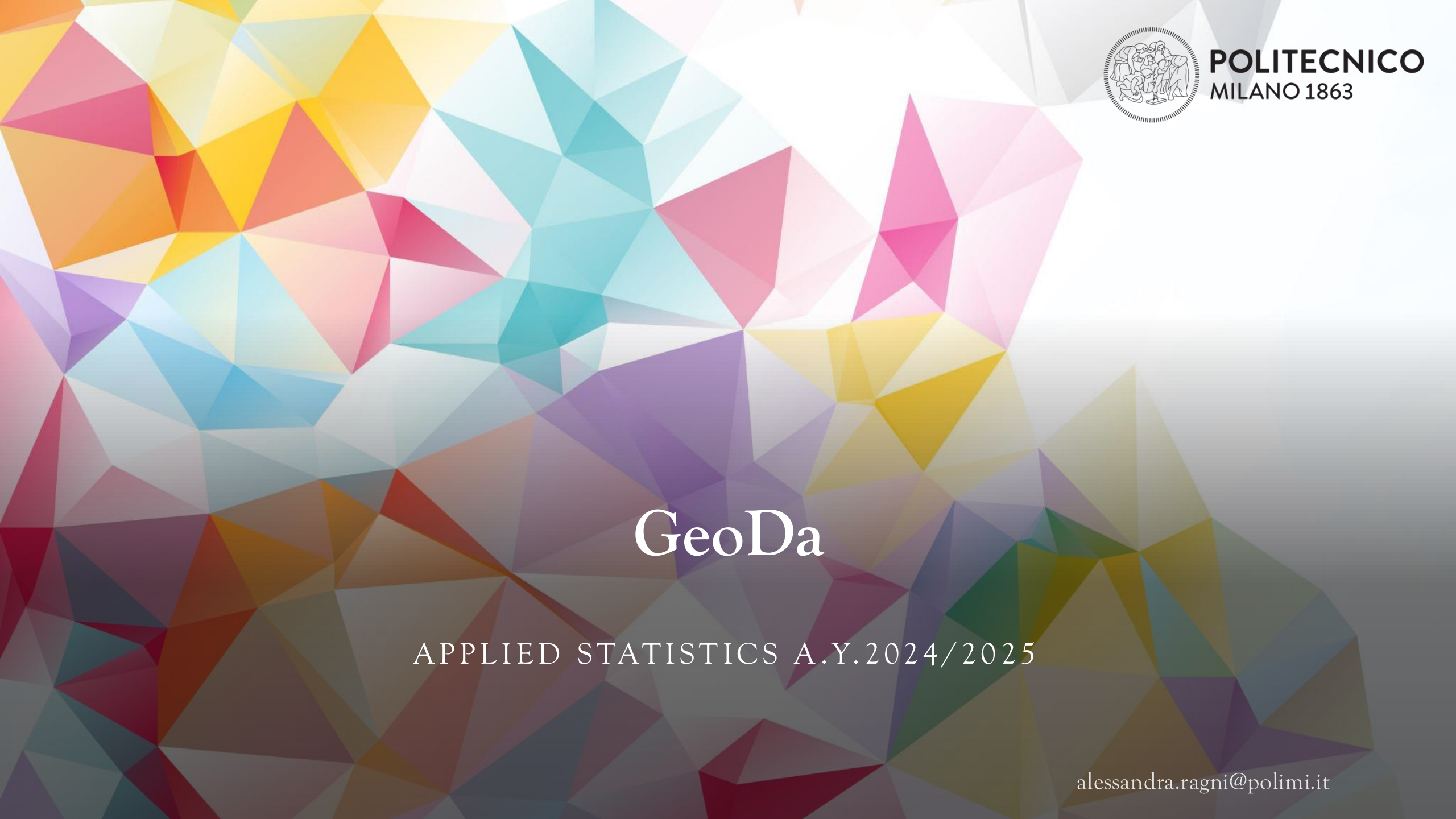


POLITECNICO
MILANO 1863

GeoDa

APPLIED STATISTICS A.Y.2024/2025

alessandra.ragni@polimi.it

POLITECNICO MILANO 1863

Download GeoDa!

https://geodacenter.github.io/download.html

Data sets we'll use today:

- **Nyc**: https://geodacenter.github.io/data-and-lab/nyc/
  Demographic and **housing data** for New York City's 55 sub-boroughs (2000s).
- **Natregimes**: https://geodacenter.github.io/data-and-lab/natregimes/
  Homicides and selected **socio-economic characteristics** for continental U.S. counties. Data for four decennial census years: 1960, 1970, 1980 and 1990.
- **Cleveland**: https://geodacenter.github.io/data-and-lab//clev_sls_154_core/
  Location and **sales price of home sales** in a core area of Cleveland, OH for the fourth quarter of 2015.
- **Guerry**: https://geodacenter.github.io/data-and-lab/Guerry/
  Classic social science foundational study by Andre-Michel Guerry on **crime, suicide, literacy** and other "**moral statistics**" in 1830s France.

POLITECNICO
MILANO 1863

Spatial Data Wrangling

https://geodacenter.github.io/workbook/01_datawrangling_1/lab1a.html

https://geodacenter.github.io/workbook/01_datawrangling_2/lab1b.html

alessandra.ragni@polimi.it

Spatial file formats: the *Shapefile* format

- A collection of three (or four) files, with extension:
  - .shp
  - .shx
  - .dbf
  - .prj (for the projection)
- The files should all be in the **same directory** and have the **same file name**

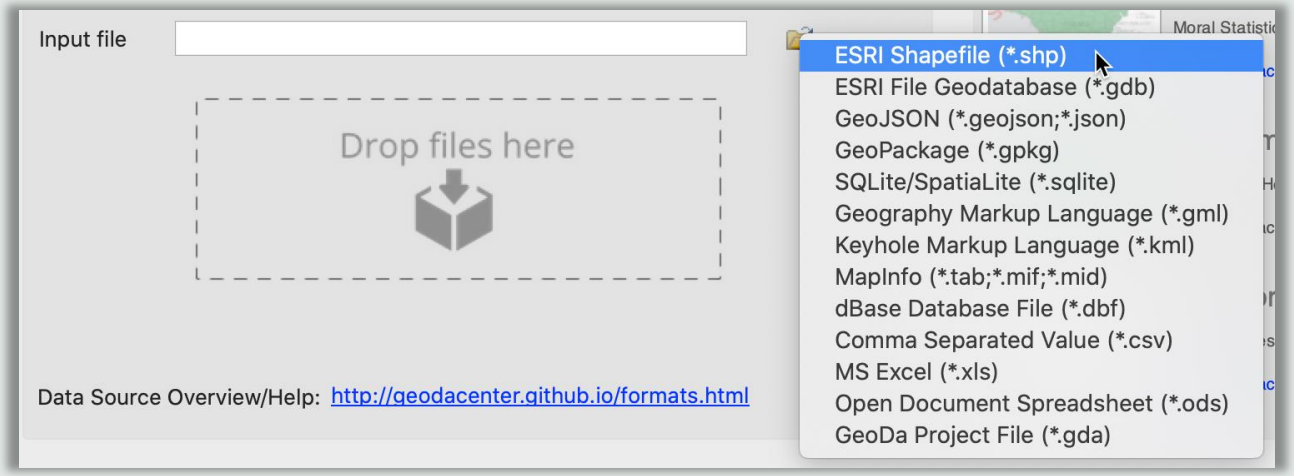
alessandra.ragni@polimi.it

How to import dataframes in GeoDa

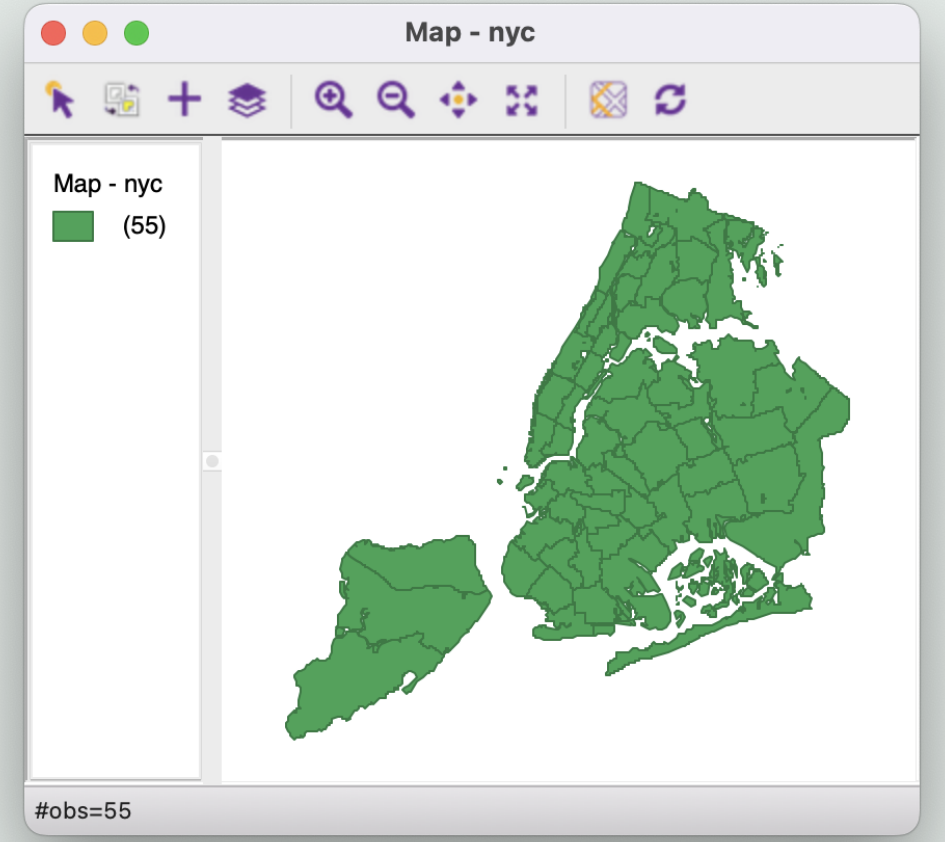
Open GeoDa, select 

And drop the .shp file you are interested in

(or select ESRI Shapefile and choose your file in your laptop).

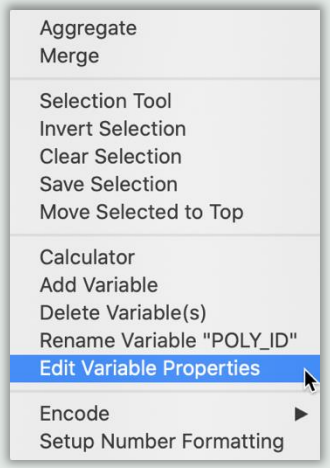


Example: import *nyc.shp*



The features are imported with the import of .shp and can be visualized through 

By right clicking on the table itself, we directly operate on the table

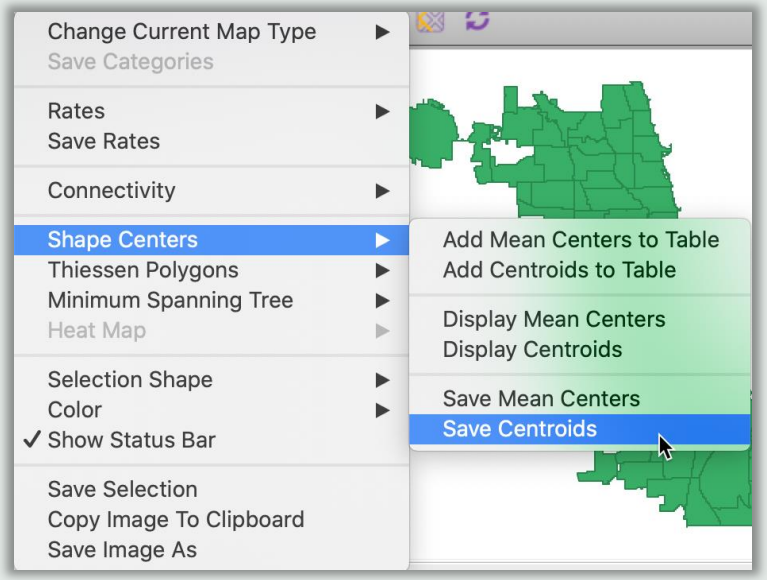


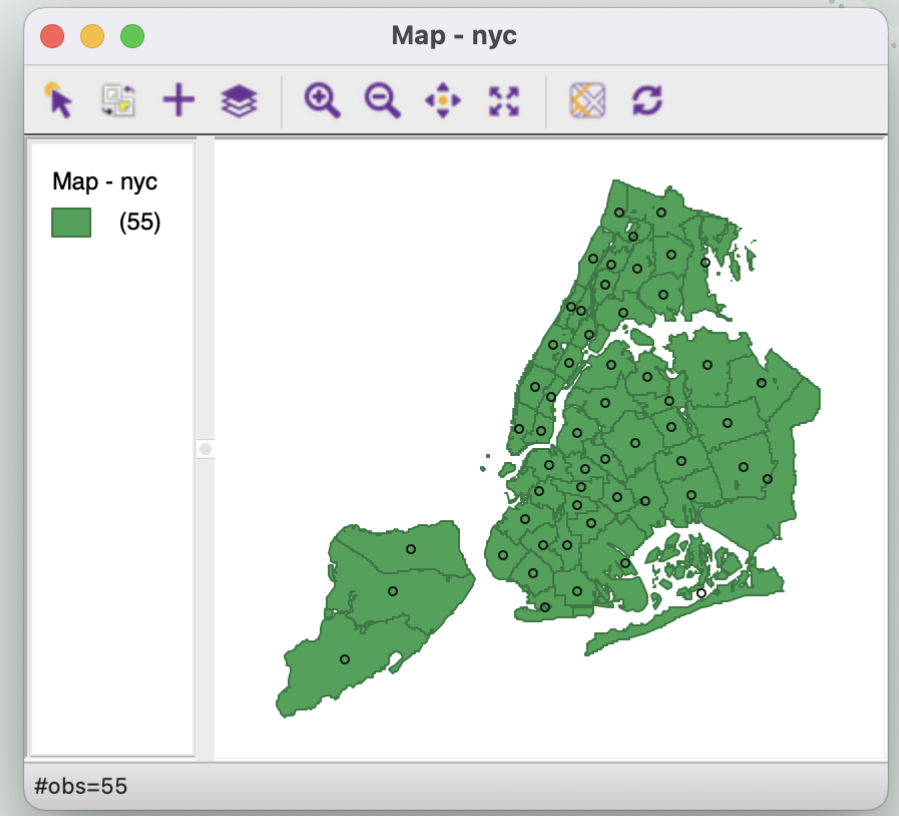
See https://geodacenter.github.io/workbook/01_datawrangling_1/lab1a.html#table-manipulations for further information

nyc data frame

Converting between Points and Polygons

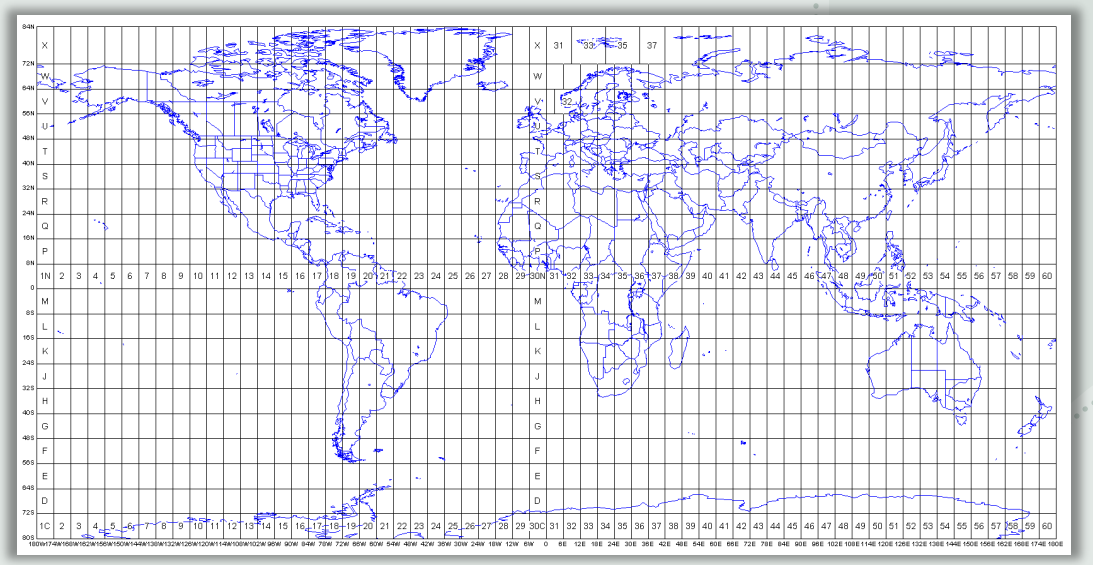
- Right click on the map





Changing the projection to UTM

- We first need to recognise whether spatial coordinates are **projected or unprojected** (an operation such as Euclidean distance is only supported for projected coordinates)
- > **File** > **Save As** > check the CRS (in proj4 format) and eventually load the CRS from a data source. Close the project and open it again



- see Reprojection in https://geodacenter.github.io/workbook/01_datawrangling_2/lab1b.html#projections

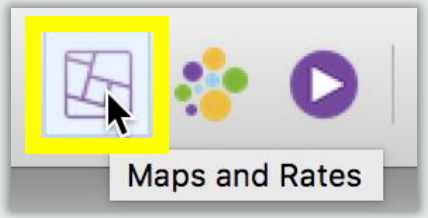
POLITECNICO MILANO 1863

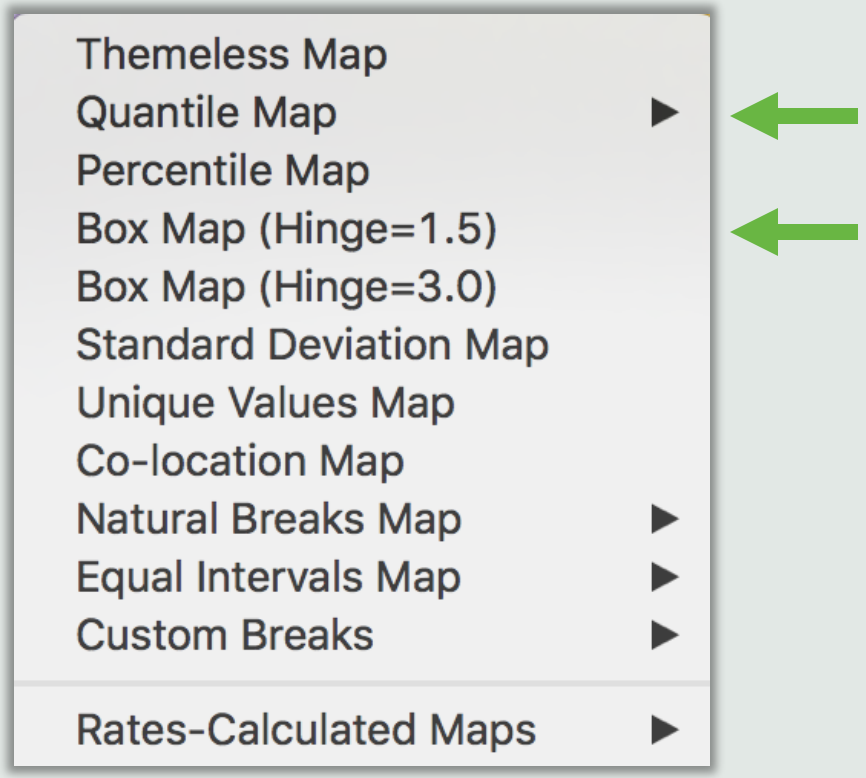
Basic Mapping

https://geodacenter.github.io/workbook/3a_mapping/lab3a.html

alessandra.ragni@polimi.it

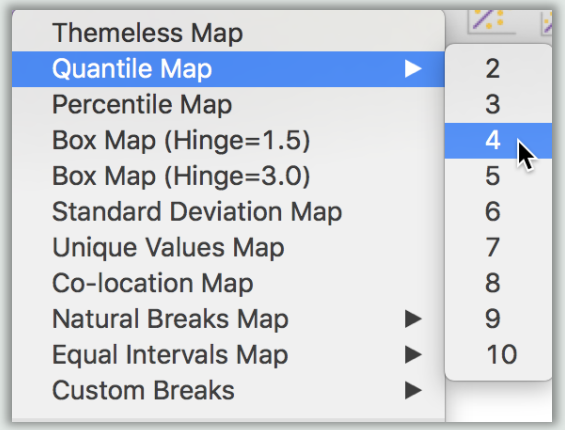
Basic mapping

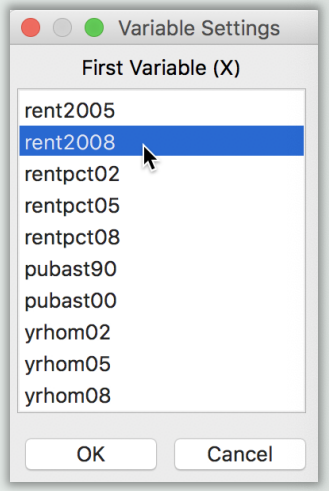


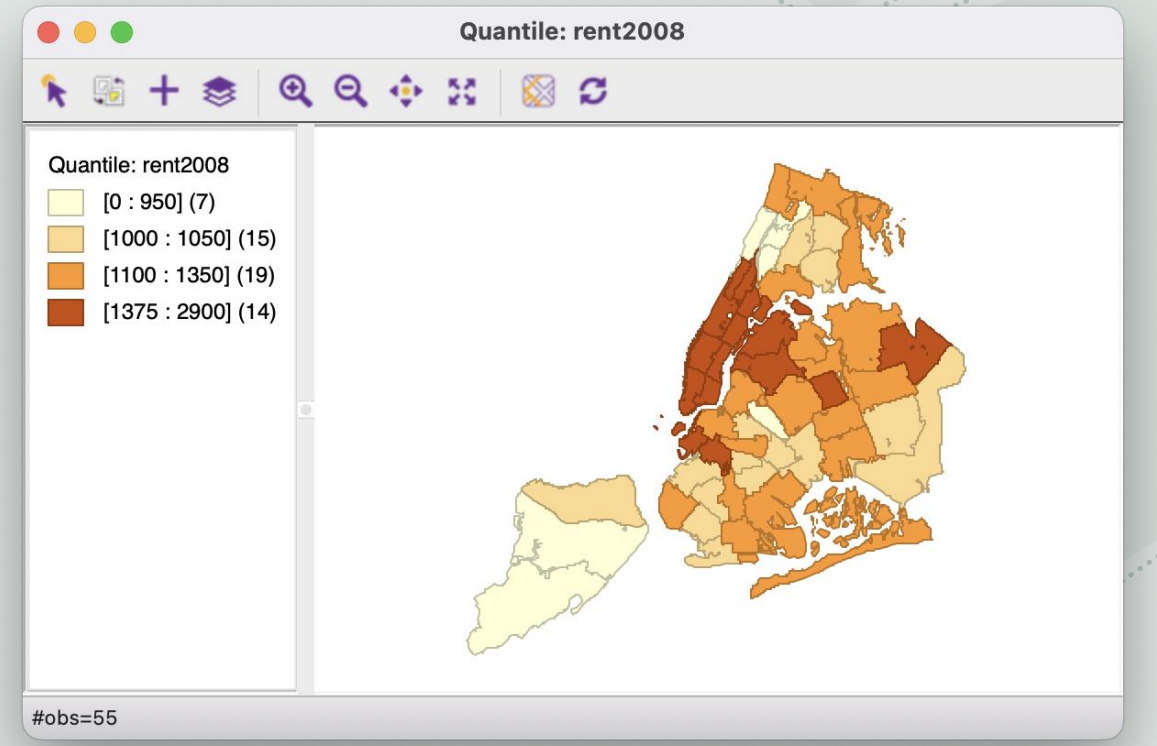


Quantile map

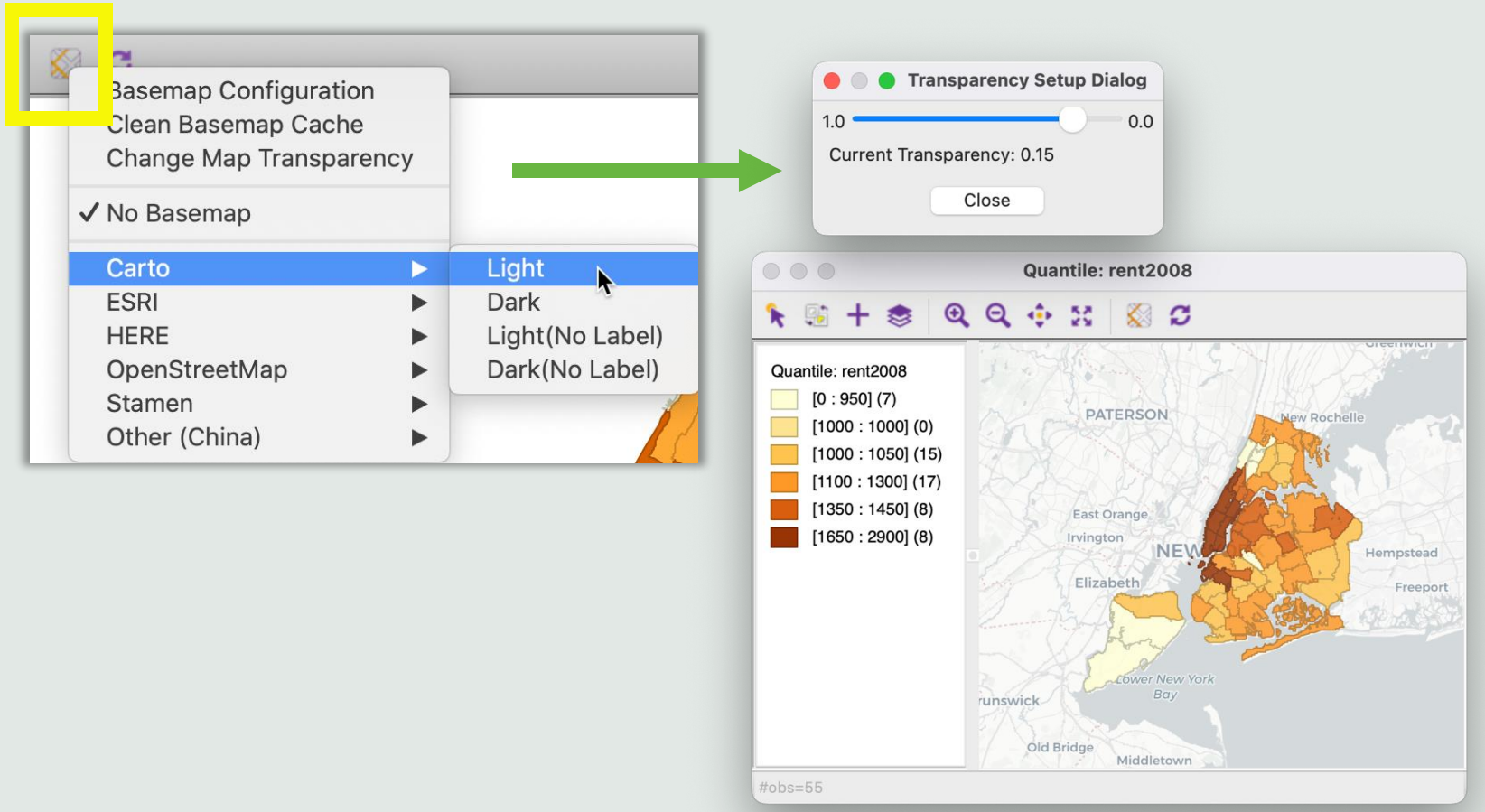
- Example: Quartile map (4 categories)



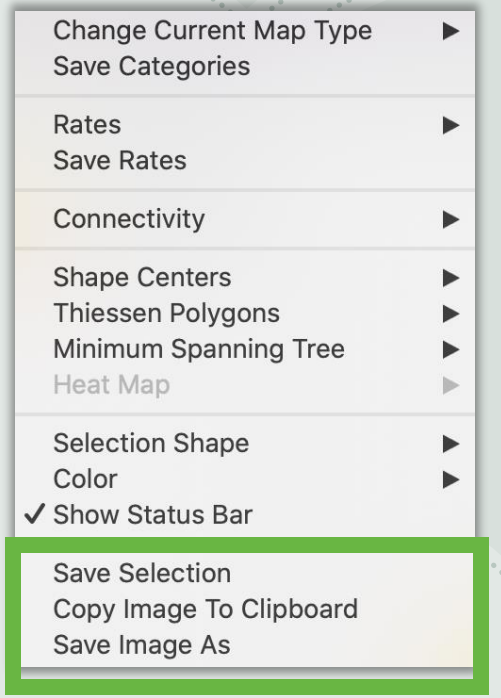




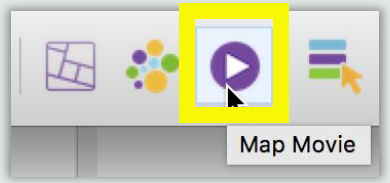
Add a Basemap

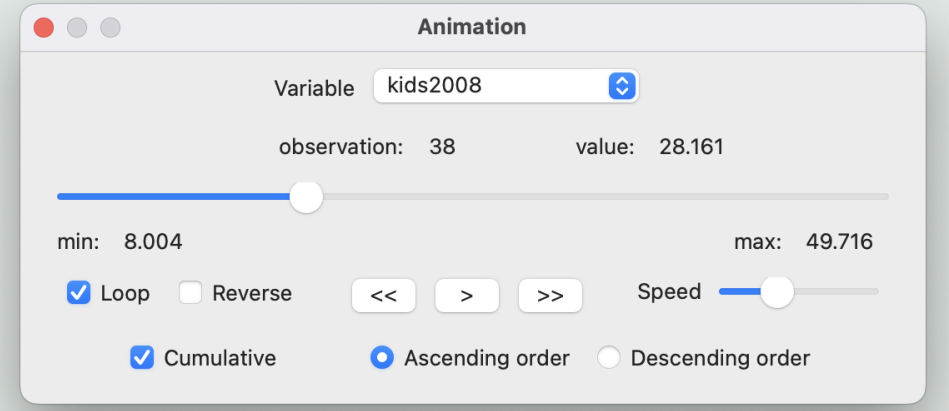


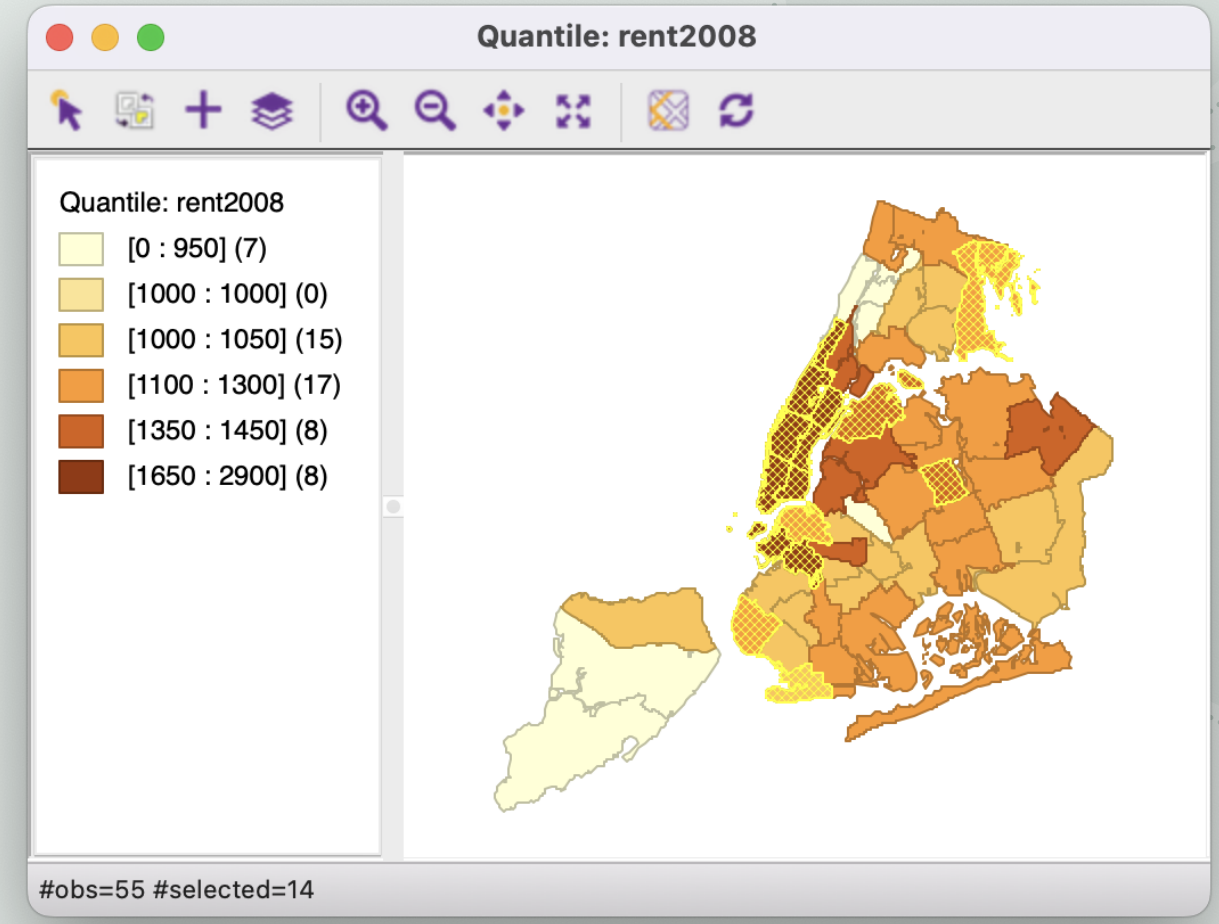
Right click on the map



Map Movie


Map Movie

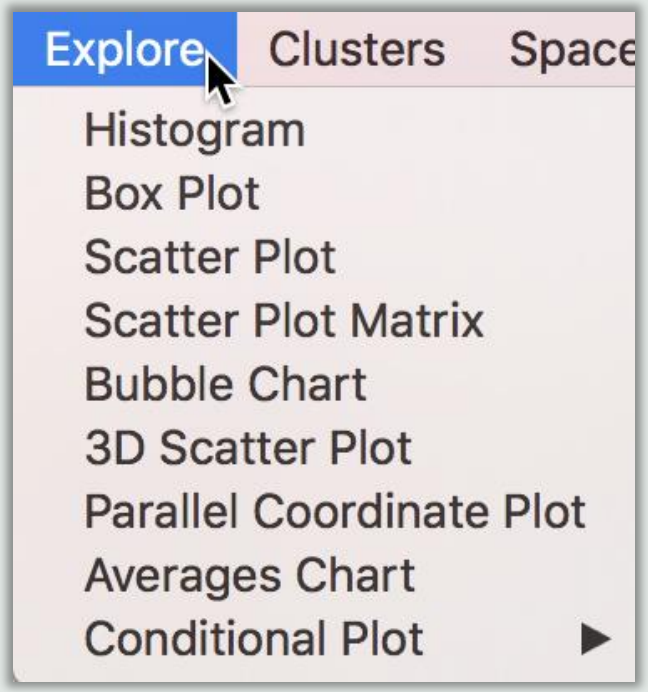

Animation
Variable
kids2008
observation: 38
value: 28.161
min: 8.004
max: 49.716
Loop
Reverse
<<
>
>>
Speed
Cumulative
Ascending order
Descending order


Quantile: rent2008
Quantile: rent2008
[0 : 950] (7)
[1000 : 1000] (0)
[1000 : 1050] (15)
[1100 : 1300] (17)
[1350 : 1450] (8)
[1650 : 2900] (8)
#obs=55 #selected=14

POLITECNICO
MILANO 1863

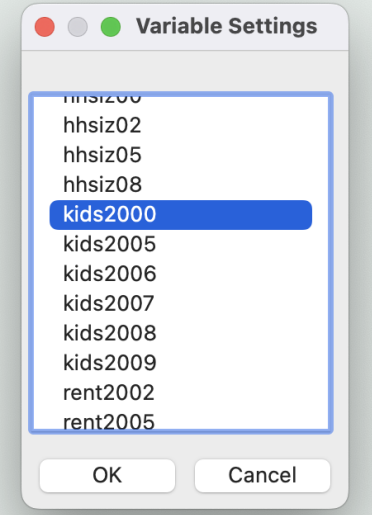
Exploratory Data Analysis

Exploratory data analysis

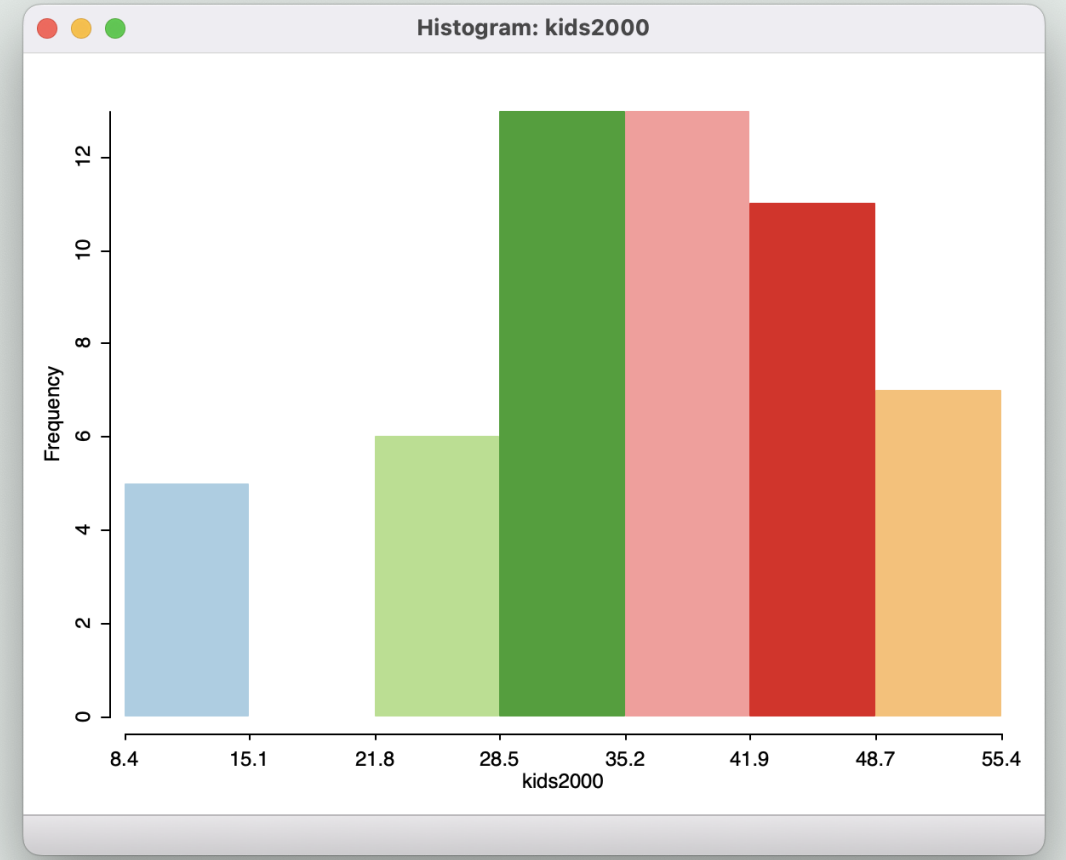


OR

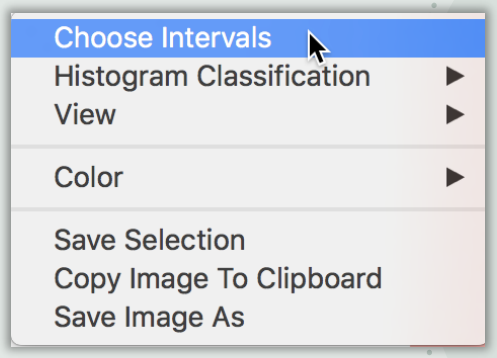
Histogram

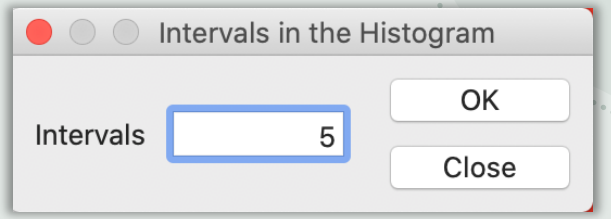


Kids2000 is the percentage of households with kids under age 18 in 2000



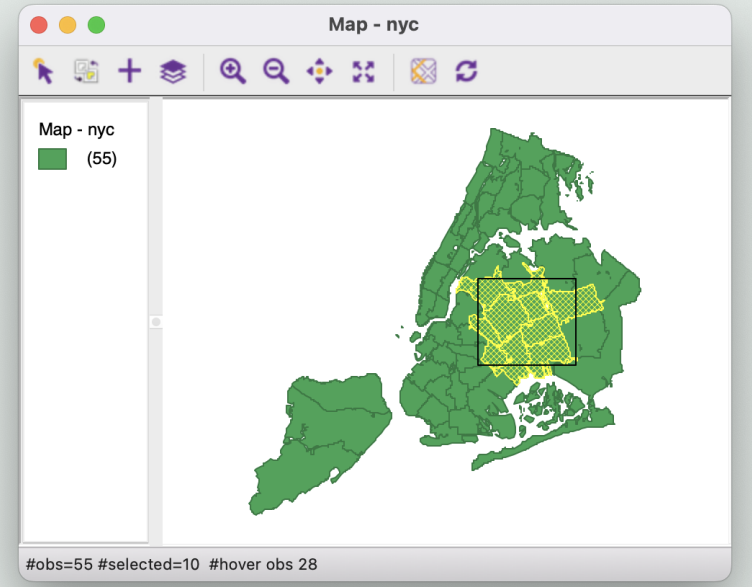
Right click

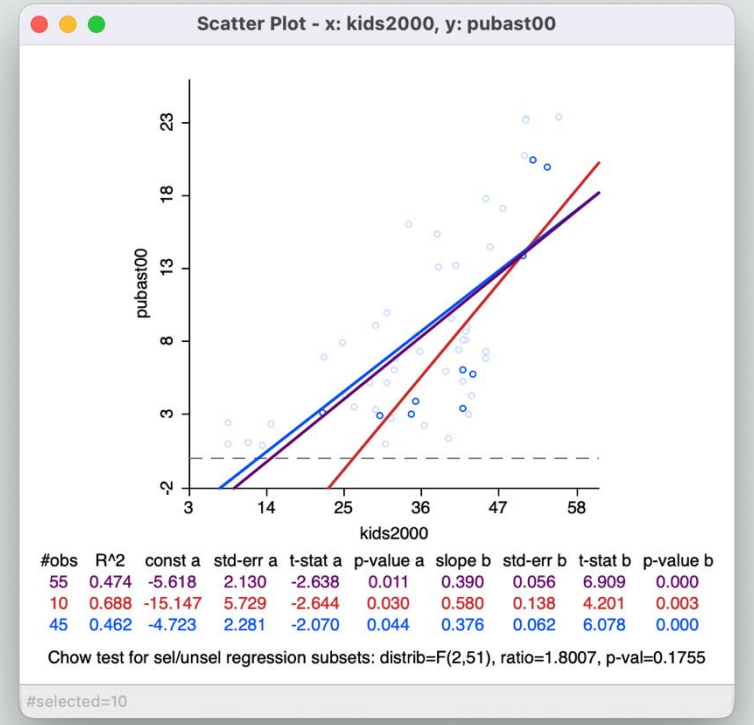




Scatterplot

- The difference between the red (selected) and blue (unselected) slopes is insufficient to lead to a rejection of the null hypothesis by means of the Chow test (p-value 0.1755)

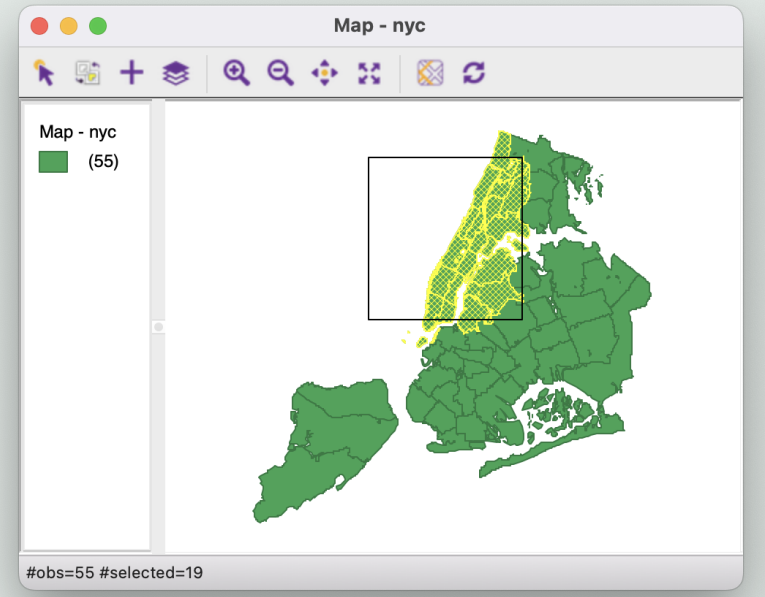


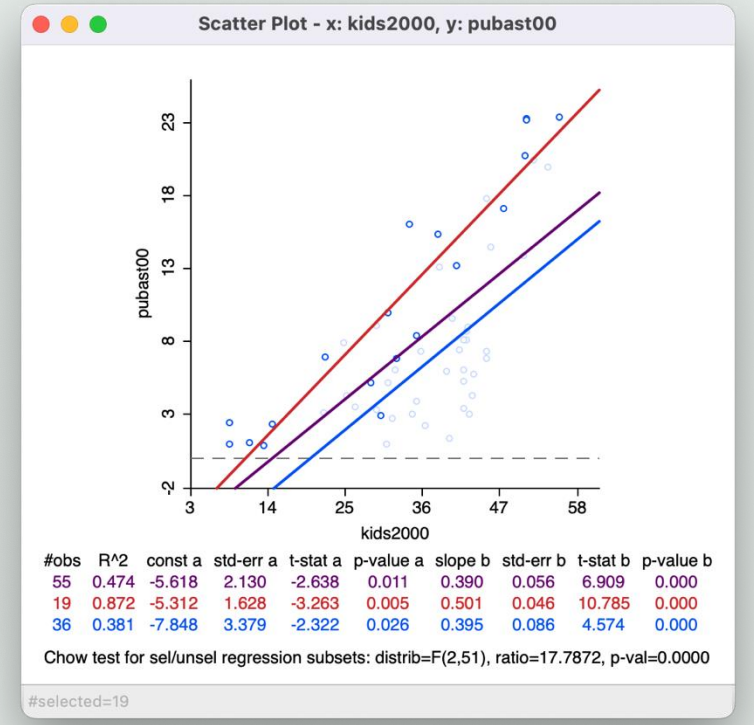


Pubast00 (percentage of households receiving public assistance)
VS
Kids2000 (percentage of households with kids under age 18 in 2000)

Scatterplot

- In this case instead, we have **evidence for a structural change much stronger**, as evidenced by the Chow test with p-value<0.005





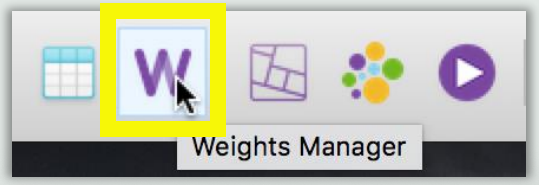
nyc data frame

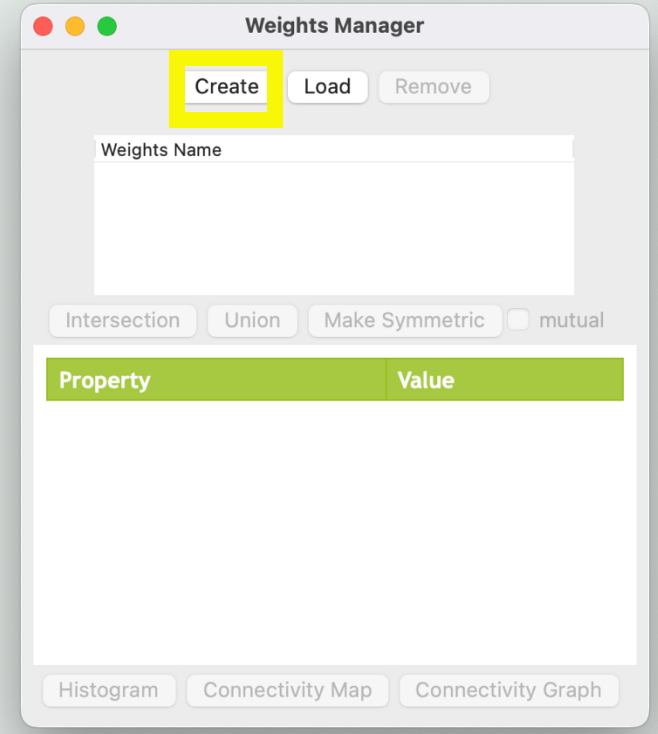
Contiguity-Based Spatial Weights

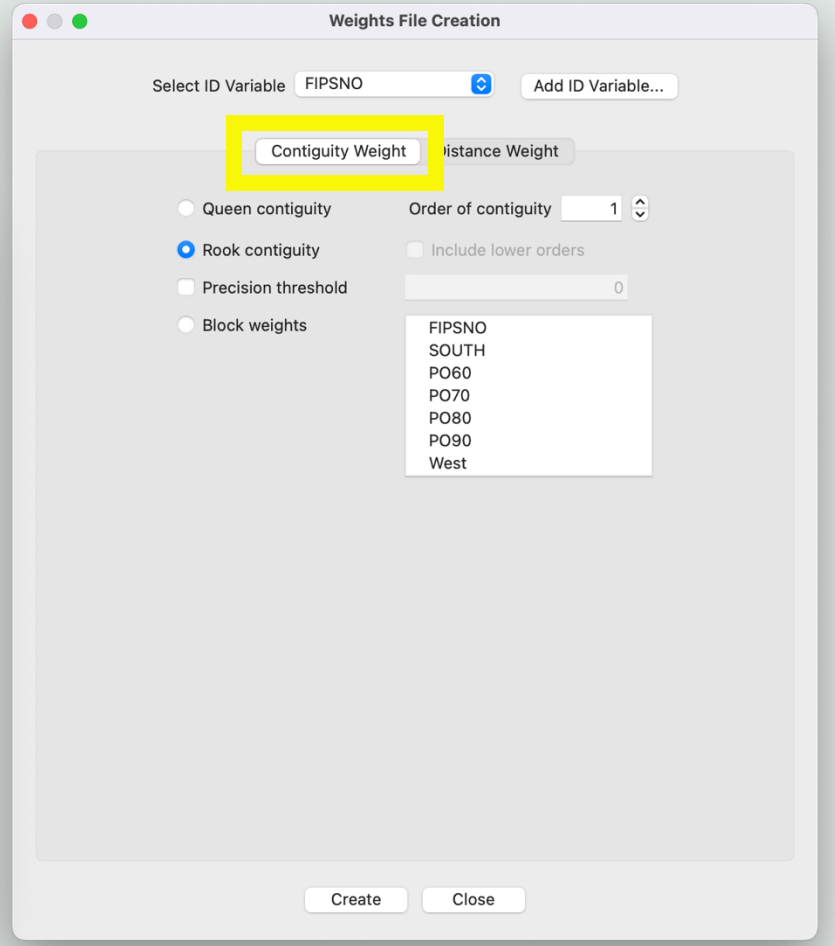
https://geodacenter.github.io/workbook/4a_contig_weights/lab4a.html

alessandra.ragni@polimi.it

The Weights Manager







- **ID Variable**: it's a key that links the data to the weights; better to have an integer variable.
  If not present , press "Add ID Variable".

- FIPSNO is a fips code as a numeric variable

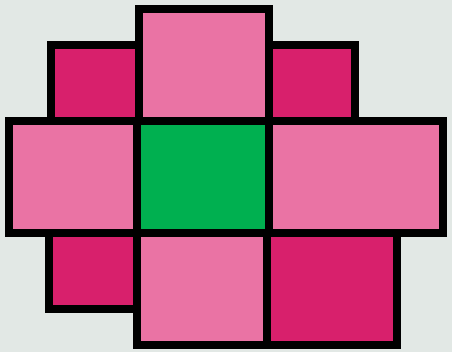
natregimes data frame

Rook VS Queen contiguity

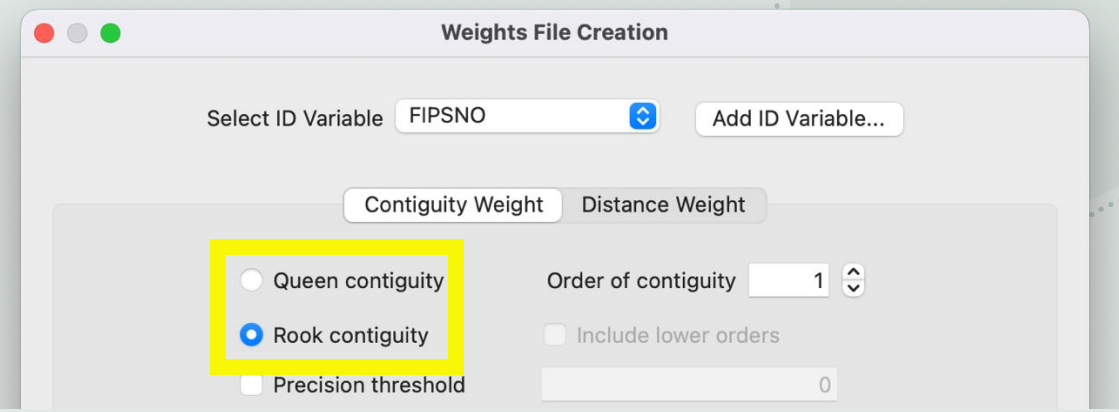
Choose the most suitable according to your specific application!

- **Rook contiguity**, i.e., when only common sides of the polygons are considered to define the neighbor relation (*common vertices are ignored*).
- The difference between the **Rook** and **Queen** criterion to determine neighbors is that the latter also *includes common vertices*.

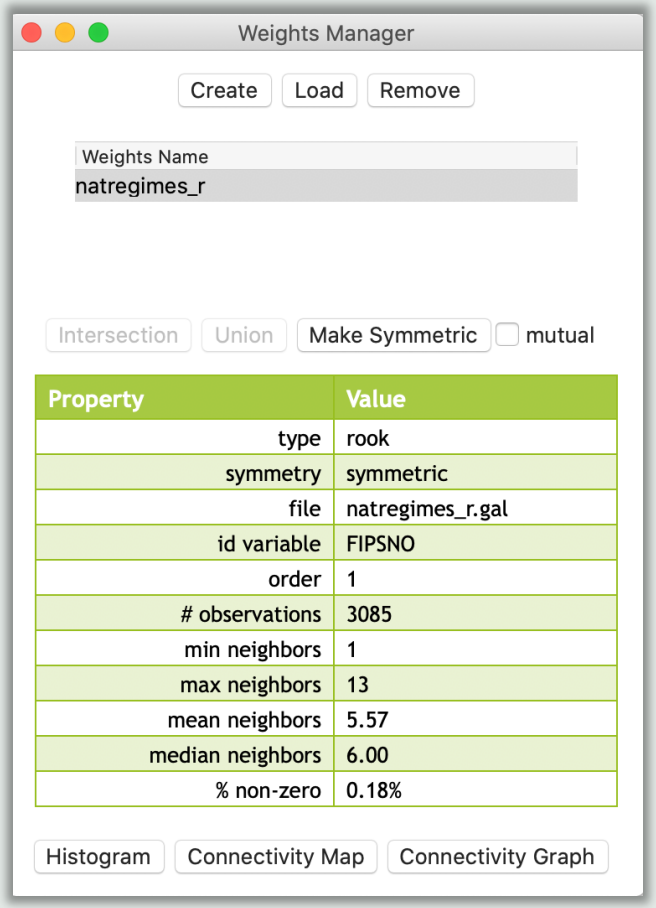
For example:



Green polygon has:
With Rook criterion → 4 neighbors
With Queen criterion → 8 neighbors



Summary Properties



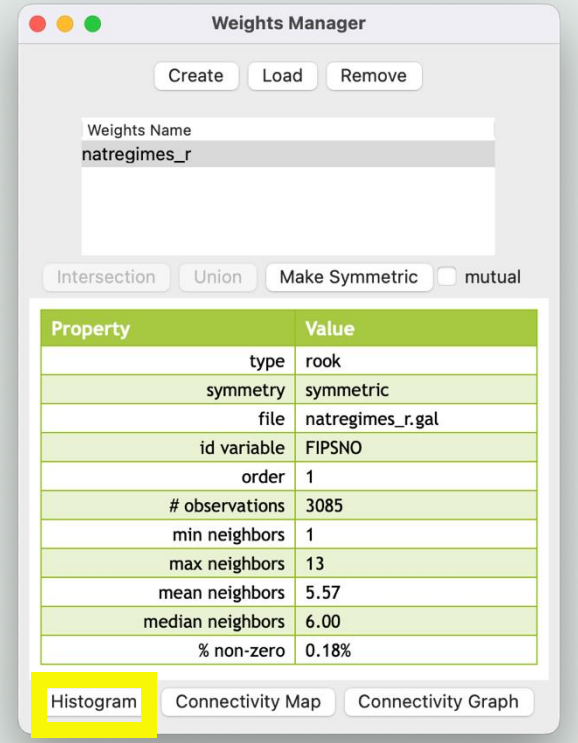
Summary statistics are listed, for example:

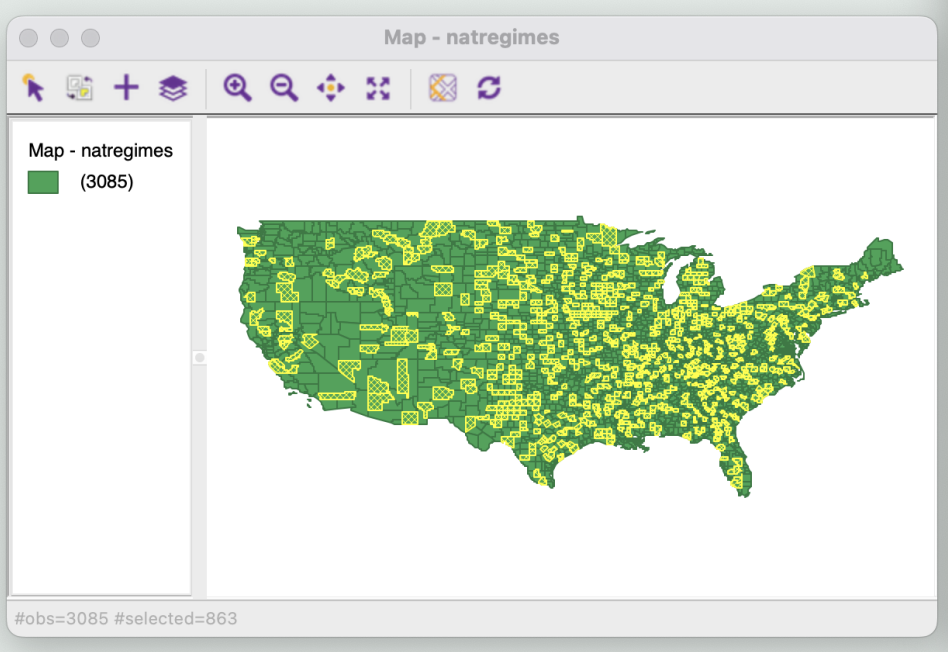
- the minimum (1) and maximum (13) number of neighbors,
- the mean (5.57) and median (6) number of neighbors,
- the percent nonzero cells in the matrix (0.18%), which is an indication of the sparsity of the weights.

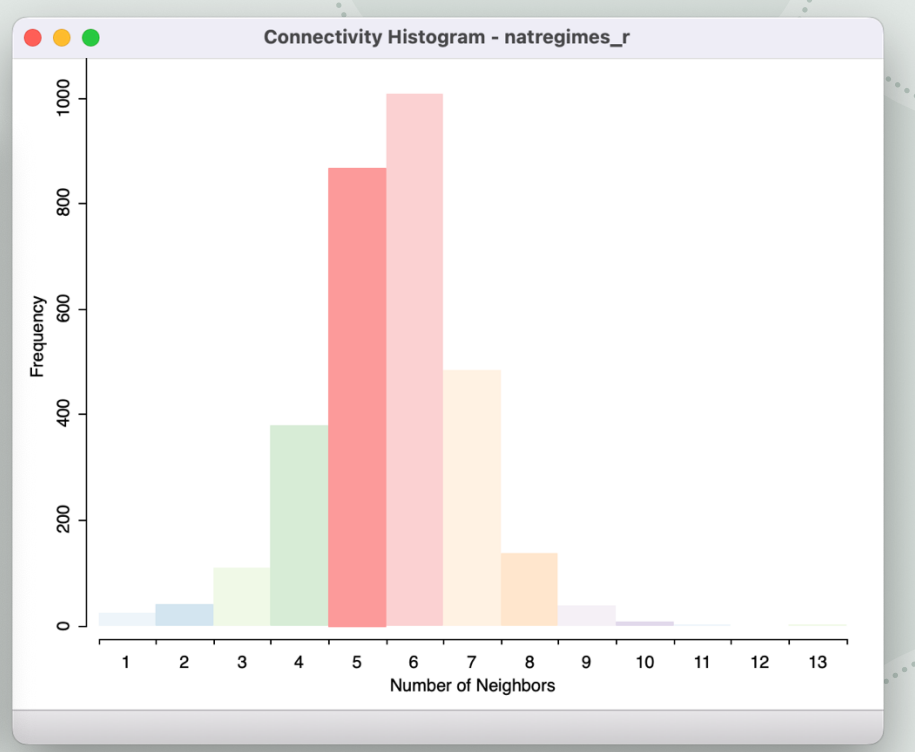
File .gal →

```
0 3085 natregimes FIPSNO
27077 3
27135 27071 27007
53019 3
53065 53047 53043
53065 4
53019 53051 53063 53043
53047 7
53019 53073 53057 53007 53017 53025 53043
53051 4
53065 16021 16017 53063
```

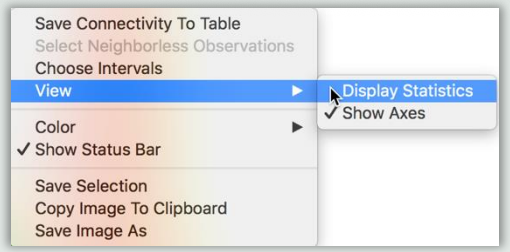
Histogram

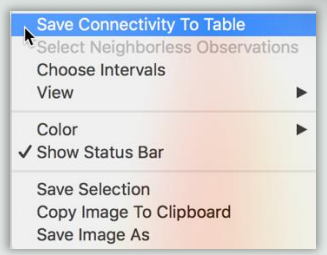






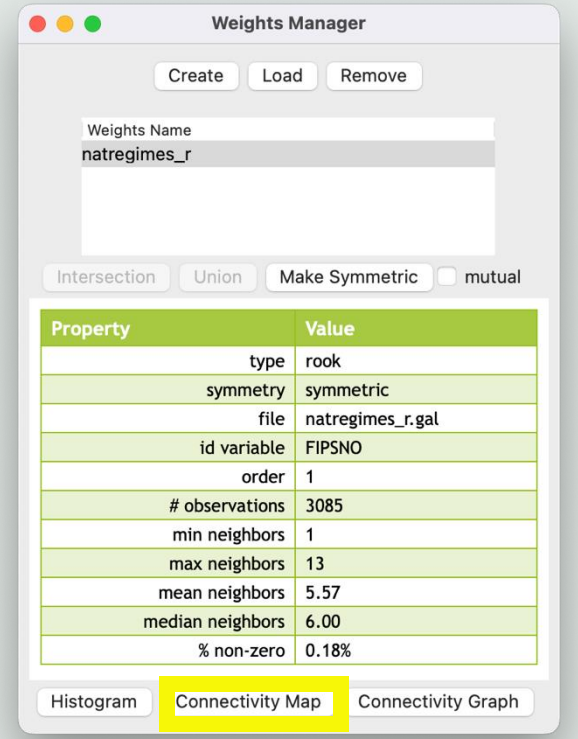
Right click on the map





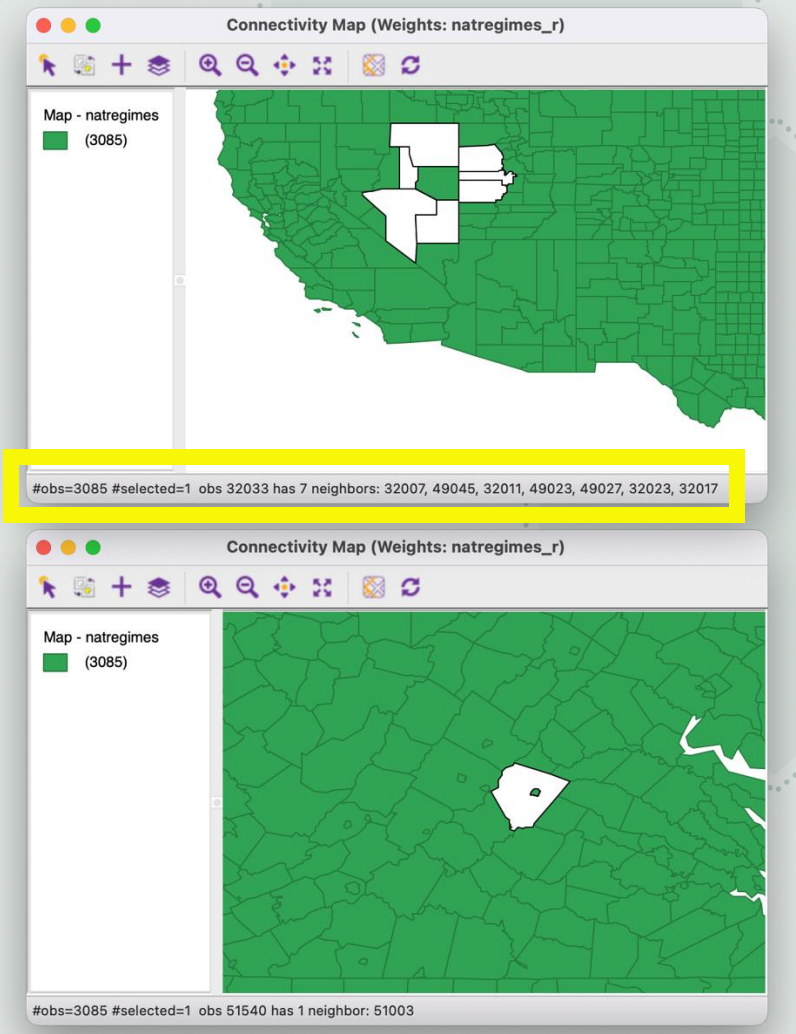
natregimes data frame

Connectivity Map

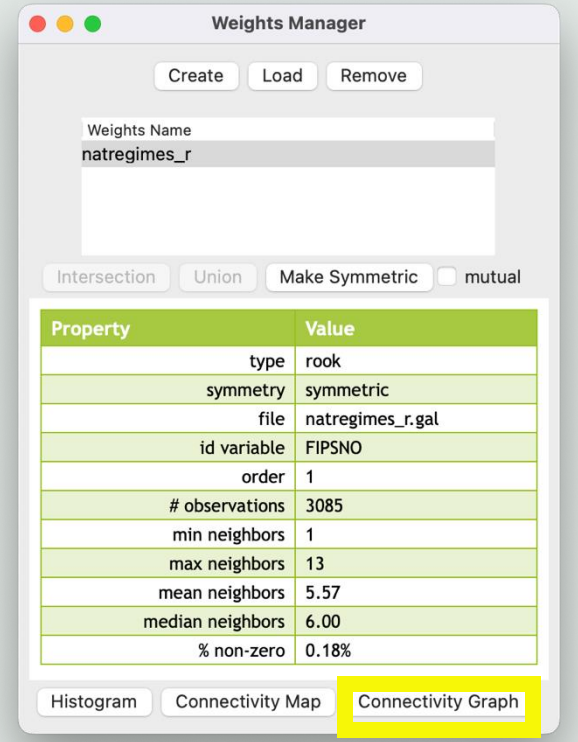


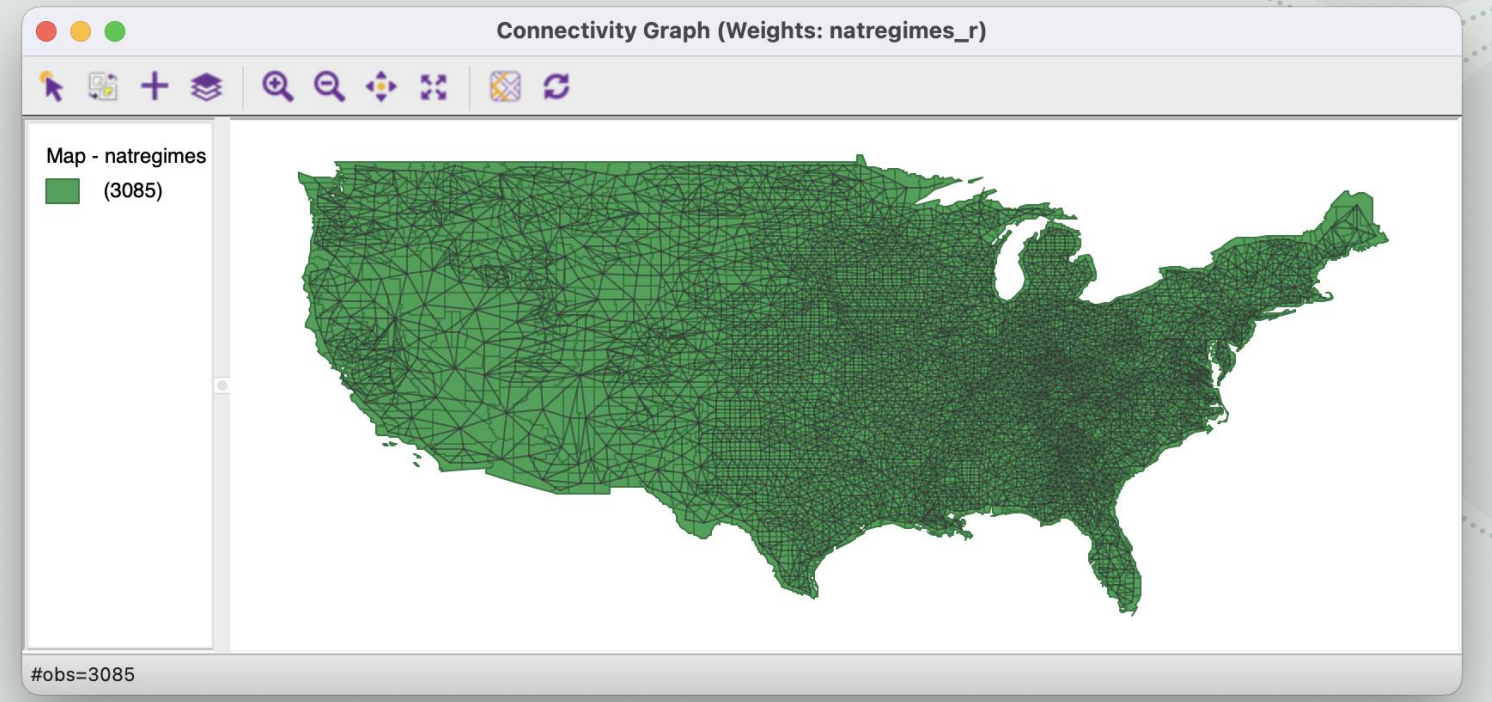
- The county is selected (the other counties become transparent) and the **outlines of the neighboring counties are shown**.
  Its ID and the IDs of the 7 neighbors are shown in the status bar.

- The **Connectivity Map** allows us to investigate *strange patterns* in the neighbor structure for some counties.
  For example, several counties in the state of Virginia are so-called **city counties**, surrounded by a larger area county.



Connectivity Graph





The graph structure that corresponds to the spatial weigths connectivity is superimposed on the map.

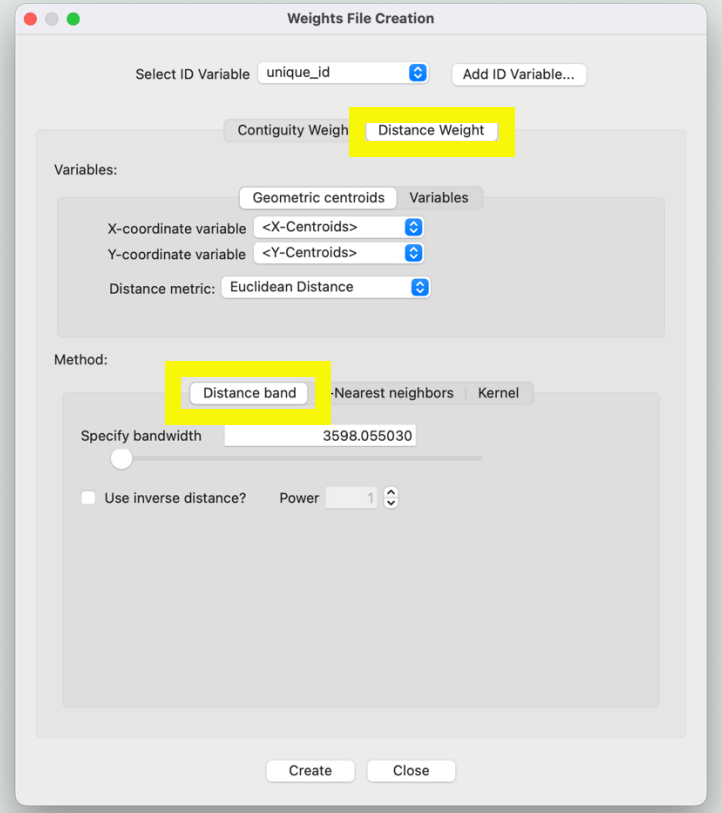
POLITECNICO
MILANO 1863

Distance-Based Spatial Weights

https://geodacenter.github.io/workbook/4b_dist_weights/lab4b.html

1) Distance-Band Weights

$w_{ij} = 1$ when $d_{ij} \leq \delta$, and $w_{ij} = 0$ otherwise, where δ is a preset critical distance cutoff.



In order to avoid isolates (islands) that would result from too stringent a critical distance, the **distance must be chosen such that <u>each location has at least one neighbor</u>.**
Such critical distance cutoff is given in the box next to "Specify bandwidth".

```
9372 4435          2409.80829
9372 3967          3277.54008
9372 9195          1355.32616
9372 4426          2415.14824
10114 11359          3598.05503
10114 20376          3174.78125
10114 20679          3389.50675
10114 17650          2696.98647
10114 17596          3340.21392
```

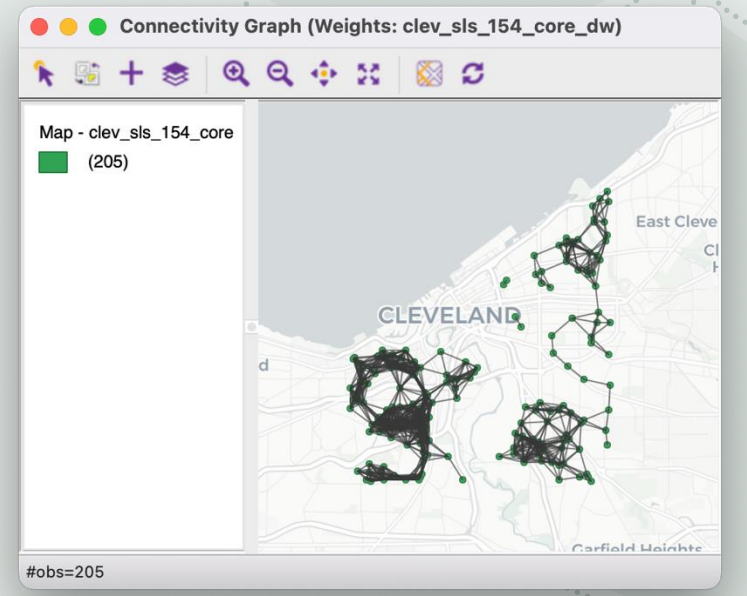
cleveland data frame

1) Distance-Band Weights

- The shape of the connectivity histogram for distance-band weights is typically **very different** from that of contiguity-based weights.

File **.gwt** →

```
0 205 clev_sls_154_core unique_id
1183 2024          1858.90398
1183 1741          1160.31203
1183 1198          385.161005
1183 2341          2525.50272
1183 2170          3367.15013
1183 1624          3013.07086
1183 7058           3369.6439
1183 6842          3253.02459
1183 6845          3390.73458
1198 2024          1758.55651
1198 1741          1170.90777
```

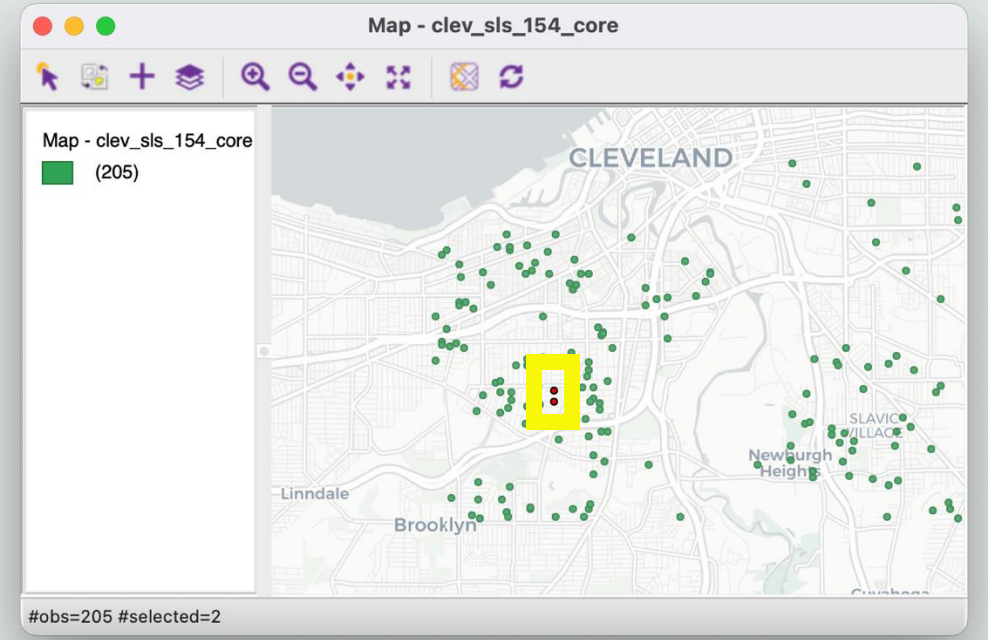


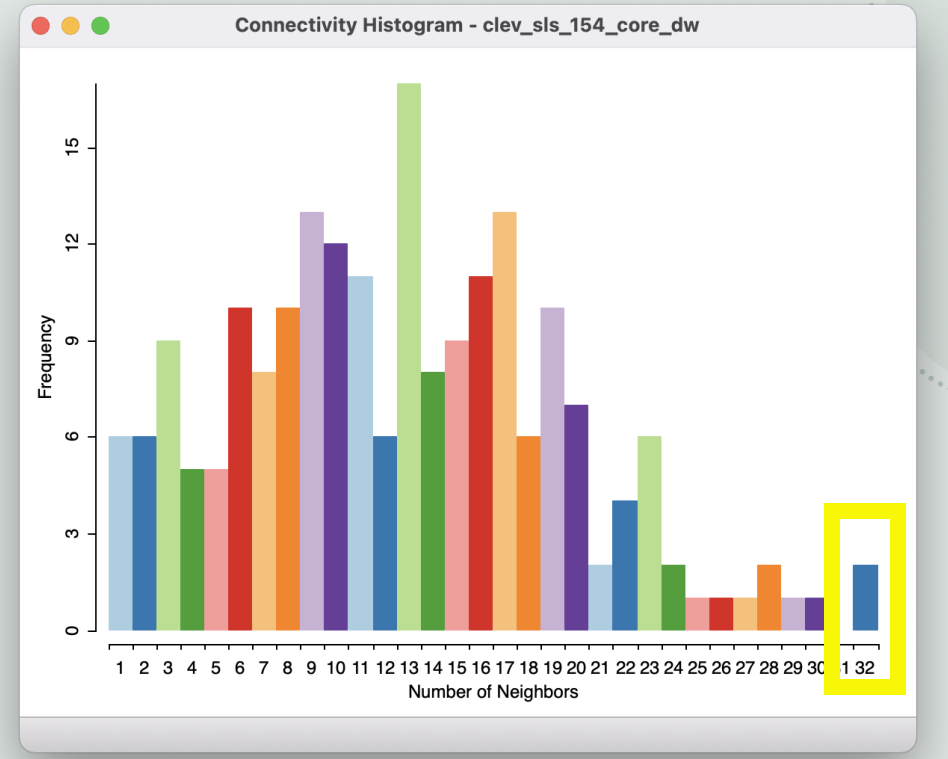
The header line is the same as for GAL files, but **each pair of neighbors is listed**, with the ID of the observation, the ID of its neighbor and the distance that separates them.

This distance is included only for informational purposes.

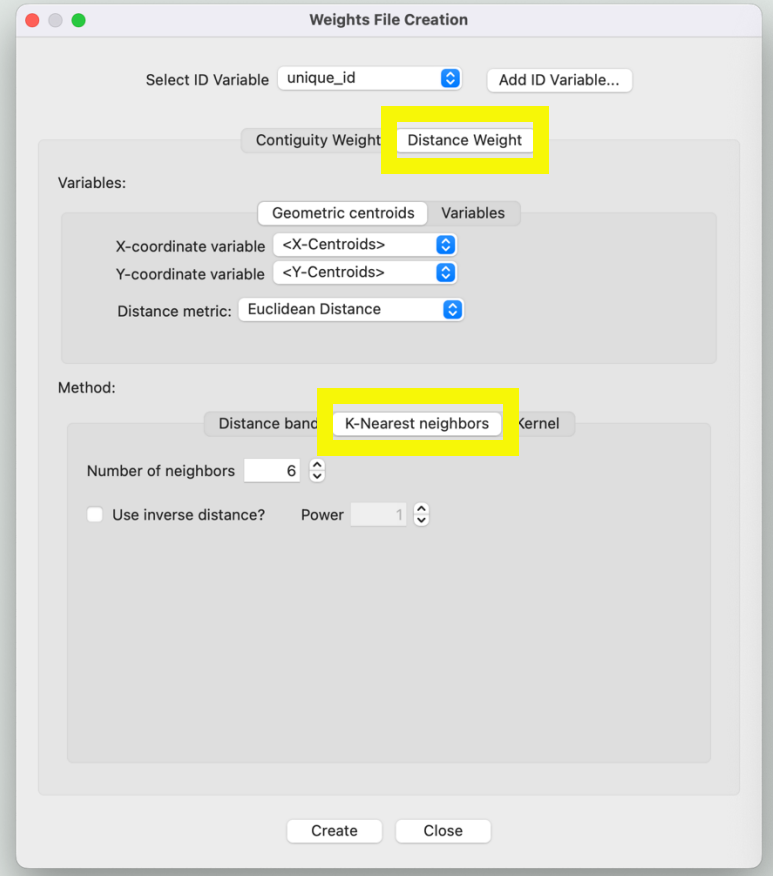
1) Distance-Band Weights

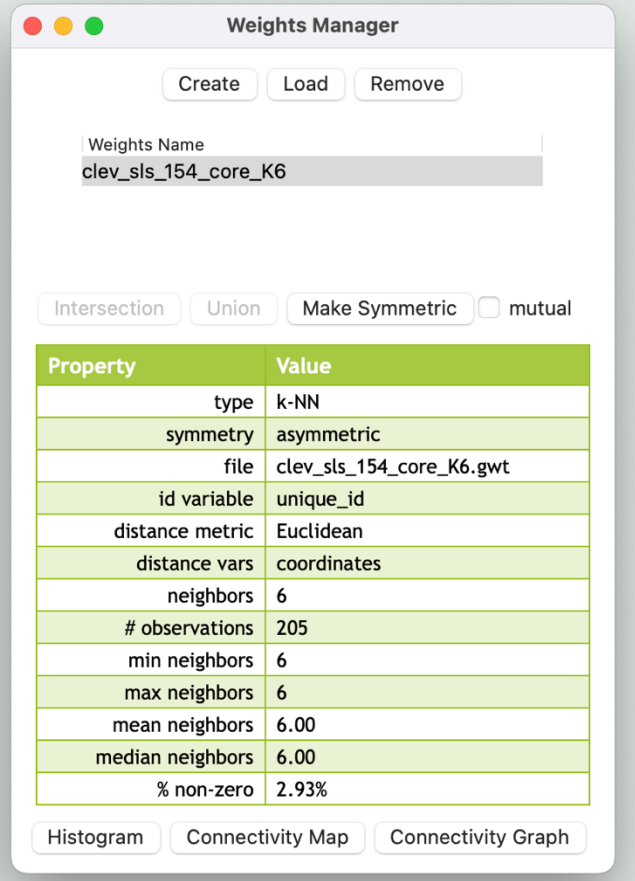
Example:





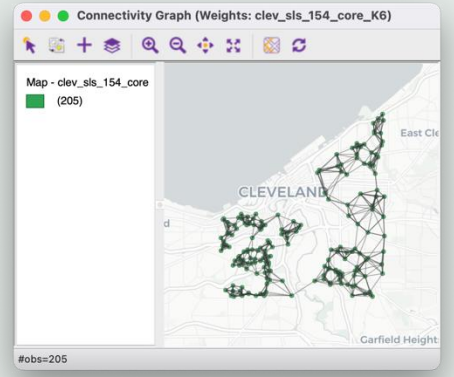
2) K-Nearest Neighbor Weights





Some comments:

- **'symmetry'** is now set to *asymmetric* (It's a fundamental difference with distance-band weights).
- The histogram doesn't make much sense, since all observations have the same number of neighbors (by construction) → just one "bar" in 6
- The Connectivity Graph is now fully connected



cleveland data frame

POLITECNICO
MILANO 1863

Spatial Weights as Distance Functions

https://geodacenter.github.io/workbook/4c_distance_functions/lab4c.html

alessandra.ragni@polimi.it

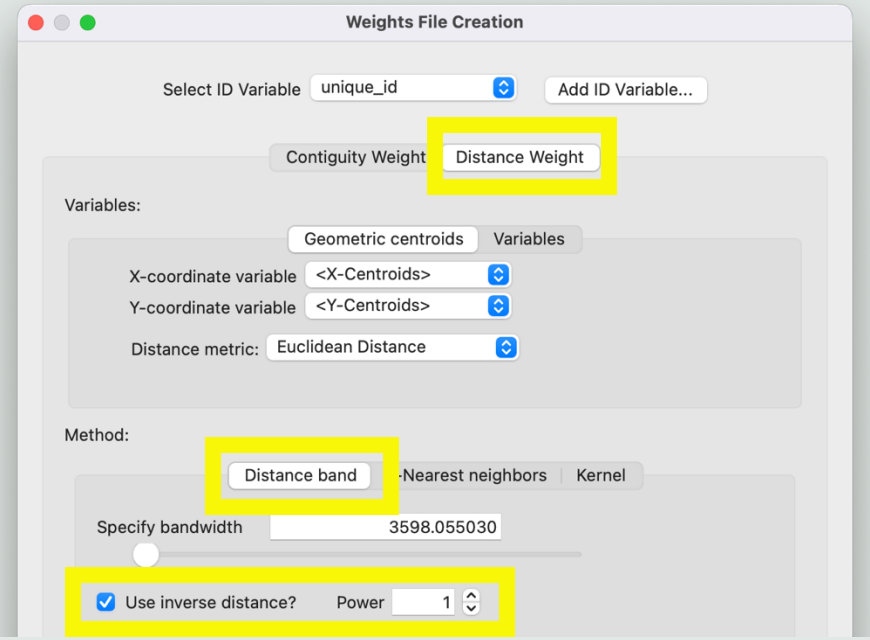
Spatial Weights as distance function

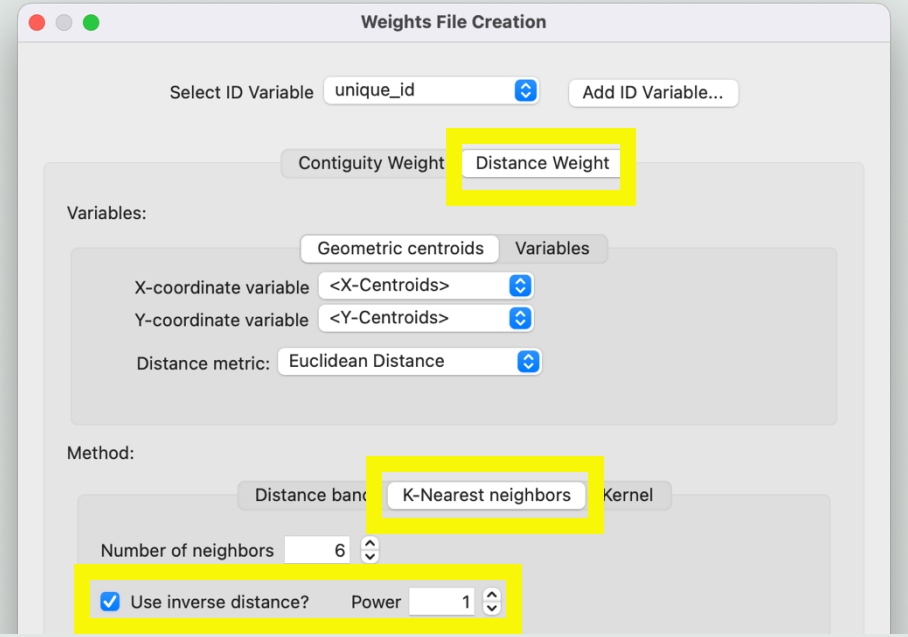
$$w_{ij} = f(d_{ij}, \theta),$$

with f as a functional form and θ a vector of parameters.

Commonly used distance functions are the inverse, with $w_{ij} = 1/d_{ij}^{\alpha}$ (and α as a parameter), and the negative exponential, with $w_{ij} = e^{-\beta d_{ij}}$ (and β as a parameter). The functions are often combined with a distance cut-off criterion, such that $w_{ij} = 0$ for $d_{ij} > \delta$.

In practice, the parameters are seldom estimated, but typically set to a fixed value, such as $\alpha = 1$ for inverse distance weights ($1/d_{ij}$), and $\alpha = 2$ for gravity weights ($1/d_{ij}^2$). By convention, the diagonal elements of the spatial weights are set to zero and not computed. Plugging in a value of $d_{ii} = 0$ would yield division by zero for inverse distance weights.





Inverse distance weights in the GWT file

A **comparison** of the entries in the GWT file for respectively the **distance-band weights** and **the inverse distance weights**.

```
0 205 clev_sls_154_core unique_id
1183 2024          1858.90398
1183 1741          1160.31203
1183 1198          385.161005
1183 2341          2525.50272
1183 2170          3367.15013
1183 1624          3013.07086
1183 7058           3369.6439
1183 6842          3253.02459
1183 6845          3390.73458
1198 2024          1758.55651
1198 1741          1170.90777
1198 1183          385.161005
```

```
1  0 205 clev_sls_154_core unique_id
2  1183 2024      0.000537951401
3  1183 1741      0.000861837141
4  1183 1198       0.00259631683
5  1183 2341      0.000395960769
6  1183 2170      0.000296987055
7  1183 1624      0.000331887316
8  1183 7058      0.000296767264
9  1183 6842      0.000307406222
10 1183 6845      0.000294921344
11 1198 2024      0.000568648203
12 1198 1741      0.000854038232
13 1198 1183       0.00259631683
```

We notice that the **pairs of neighbors are identical**, as expected.
Also, the value for the **inverse distance weight is exactly the inverse of the distance**.

alessandra.ragni@polimi.it

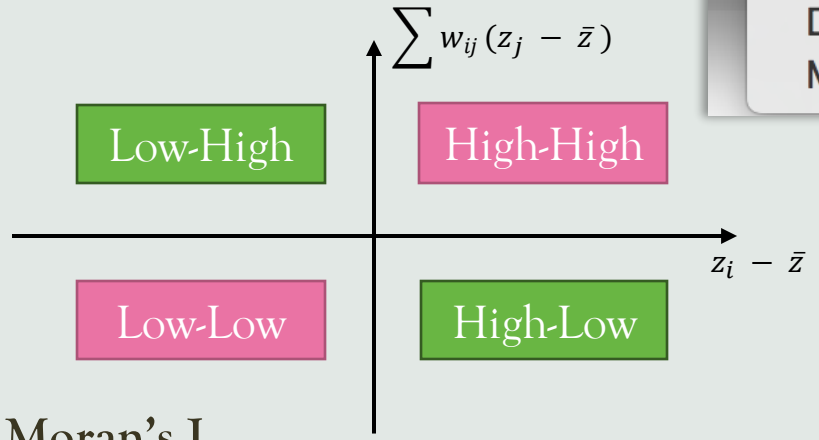
Global Spatial Autocorrelation

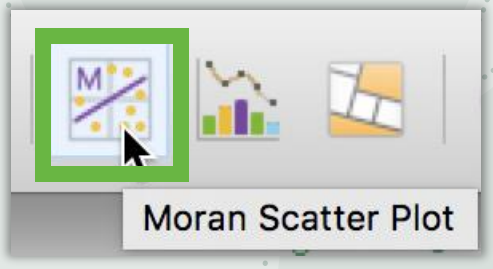
https://geodacenter.github.io/workbook/5a_global_auto/lab5a.html

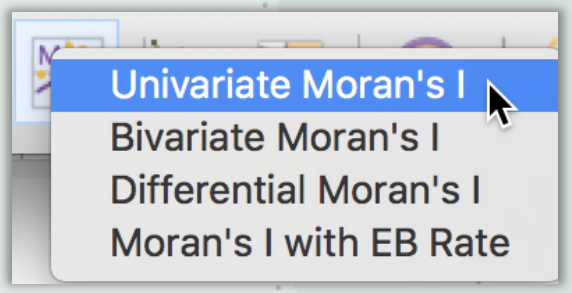
alessandra.ragni@polimi.it

Moran's I & Moran scatter plot

- **Moran's I** statistic is (one of) the most commonly used indicator of **global spatial autocorrelation;**
- **Moran scatterplot** consists of a plot with spatially lagged variable on the y-axis and the original variable on the x-axis.
  It allows the classification of the nature of spatial autocorrelation in four categories:
  - High-High & Low-Low: positive spatial autocorrelation (similar values at neighboring locations)
  - High-Low & Low-High: negative spatial autocorrelation (dissimilar values at neighboring locations)

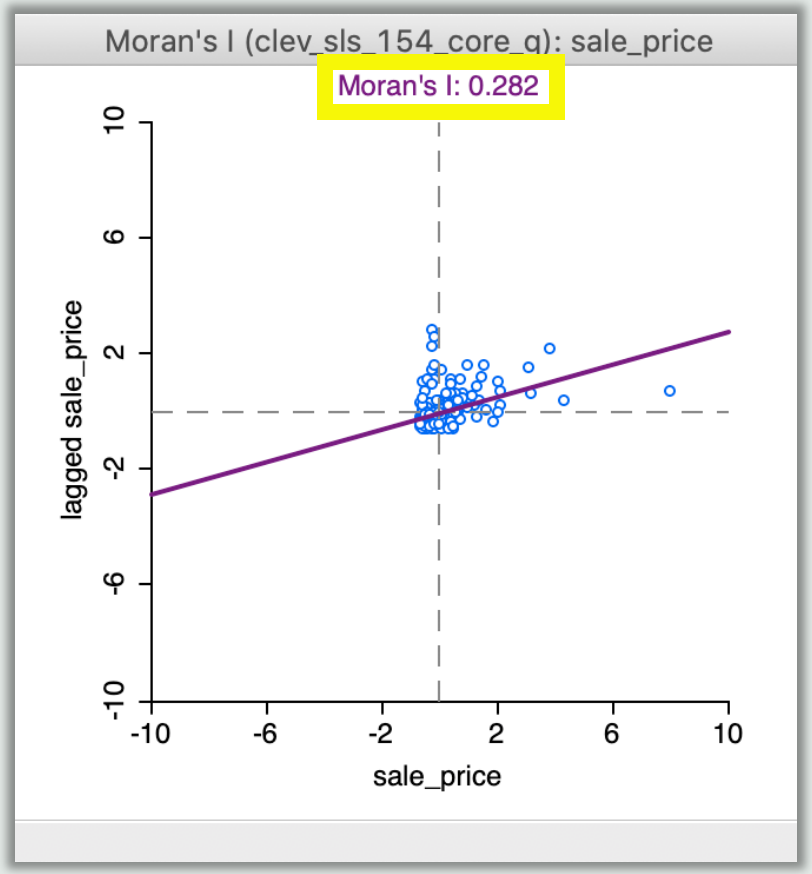






- The **slope** of the linear fit to the scatter plot equals **Moran's I.**

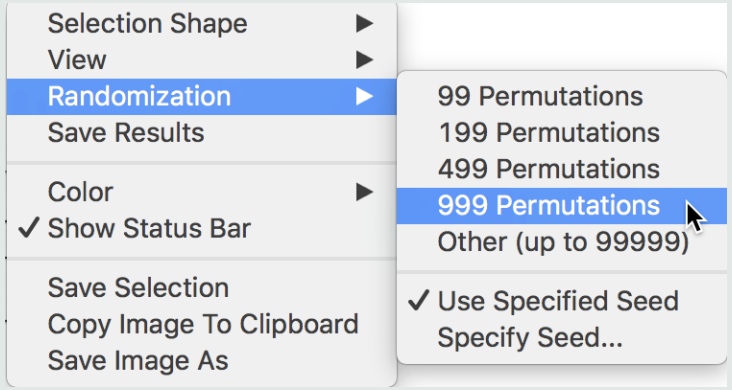
Moran's I & Moran scatter plot

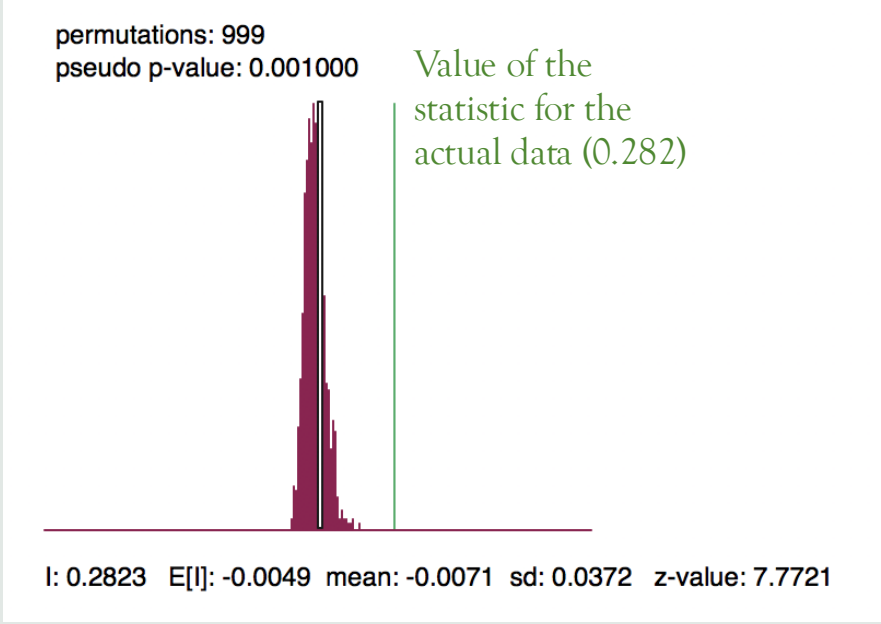


- Standardized values on both axes (sd = 1, mean = 0).
- Shape is influenced by the presence of **several outliers on the high end** (e.g. larger than 3 std deviational units from the mean).
- By **eliminating some of the outliers**, we might be able to see more detail for the remaining observation.
- We could save the scatter plot variables and create a Moran scatter plot as a standard scatterplot; see https://geodacenter.github.io/workbook/5a_global_auto/lab5a.html#moran-scatter-plot-options for more information.

Permutation inference: Randomization option

- Compute a **reference distribution for the statistic under the null hypothesis of spatial randomness by randomly permuting the observed values over the locations.**
- Right click on Moran scatter plot





→ **Strong rejection of the null hypothesis H0** (H0: «no spatial autocorrelation»)

→ Pseudo p-val = (0+1) / (999+1)

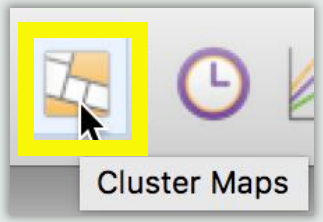
POLITECNICO MILANO 1863

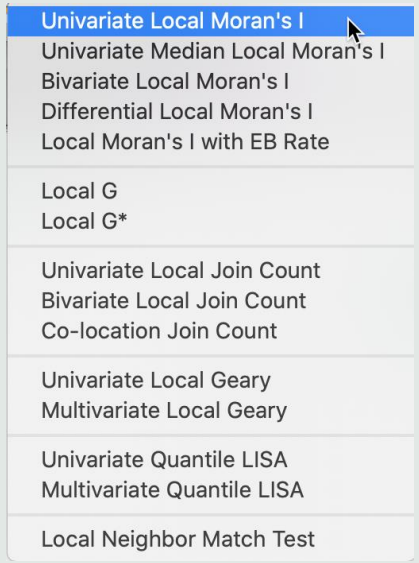
Local Spatial Autocorrelation

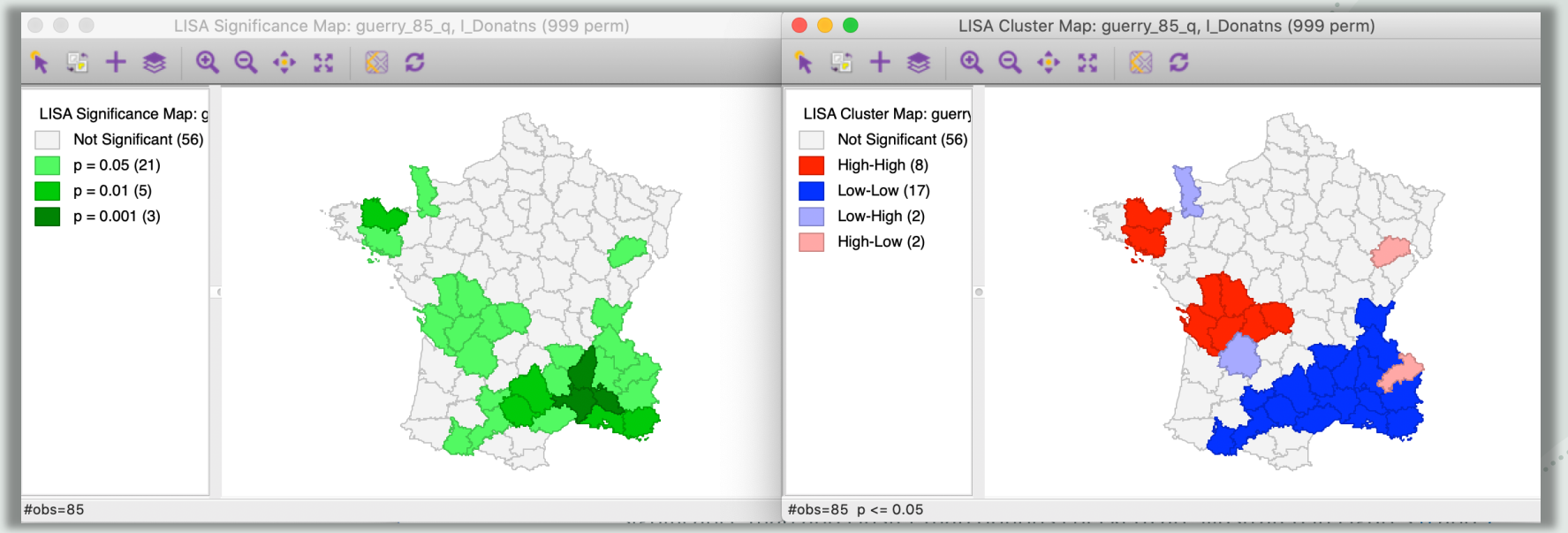
https://geodacenter.github.io/workbook/6a_local_auto/lab6a.html

Local Moran: LISA Significance & Cluster maps

→ for identifying **local clusters** and **local spatial outliers**







Significance map shows the locations with a significant local statistic, with the degree of significance reflected in increasingly darker shades of green.

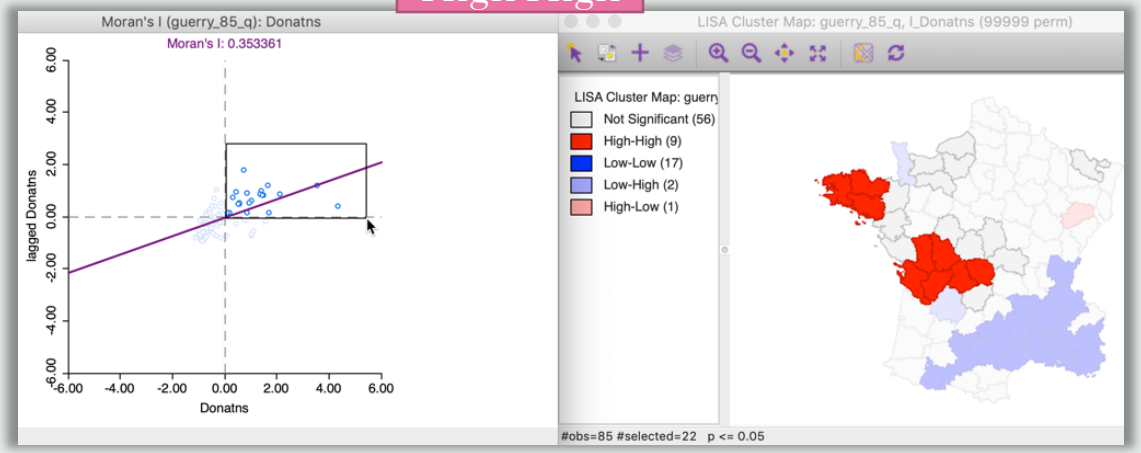
Cluster map augments the significant locations with an indication of the type of spatial association, based on the Moran scatter plot results.

guerry data frame

LISA Clusters

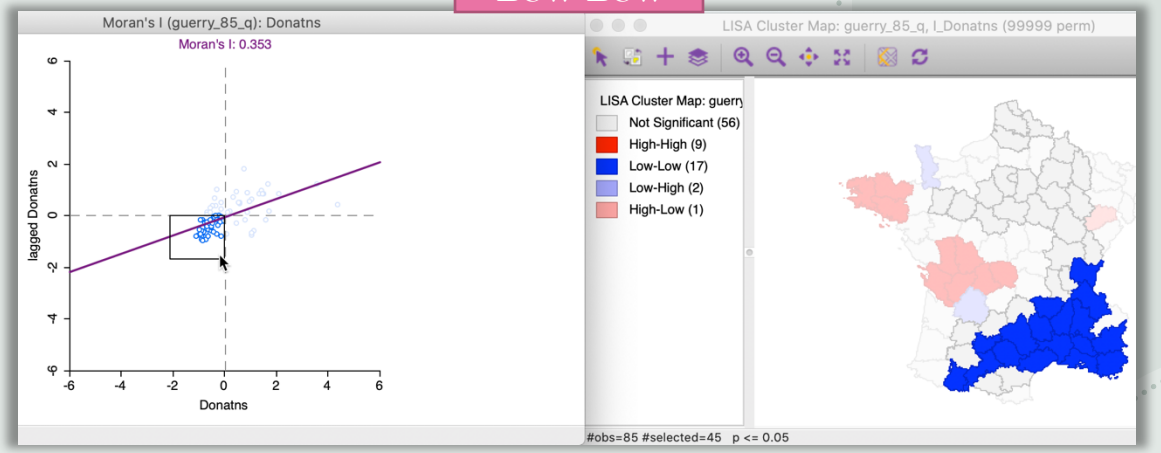
- Connection between the **Moran scatter plot** and the **cluster map**

High-High



Interpretation: Departments with a number of donants above the mean, and in the neighbourhood they have departments with a number of donants above the mean

Low-Low



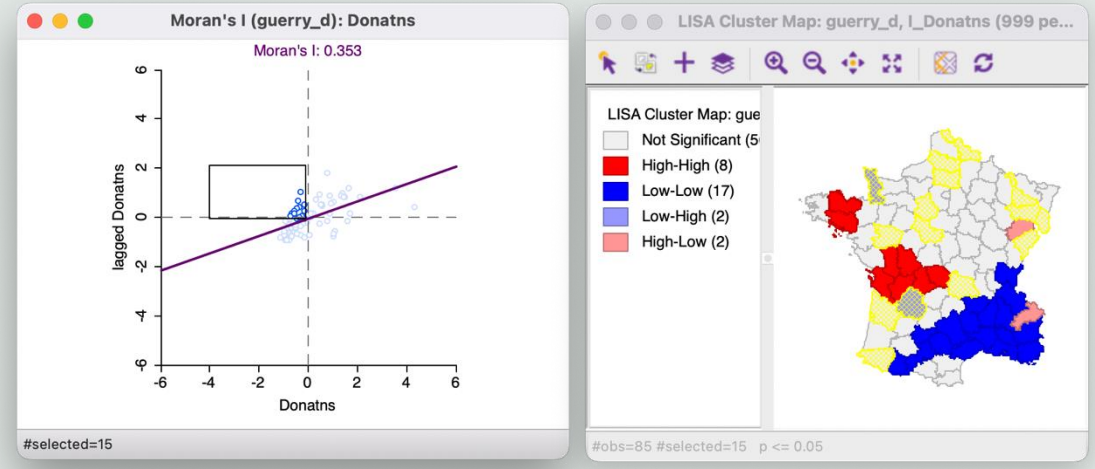
Interpretation: Departments with a number of donants below the mean, and in the neighbour hood they have departments with donants below the mean

alessandra.ragni@polimi.it

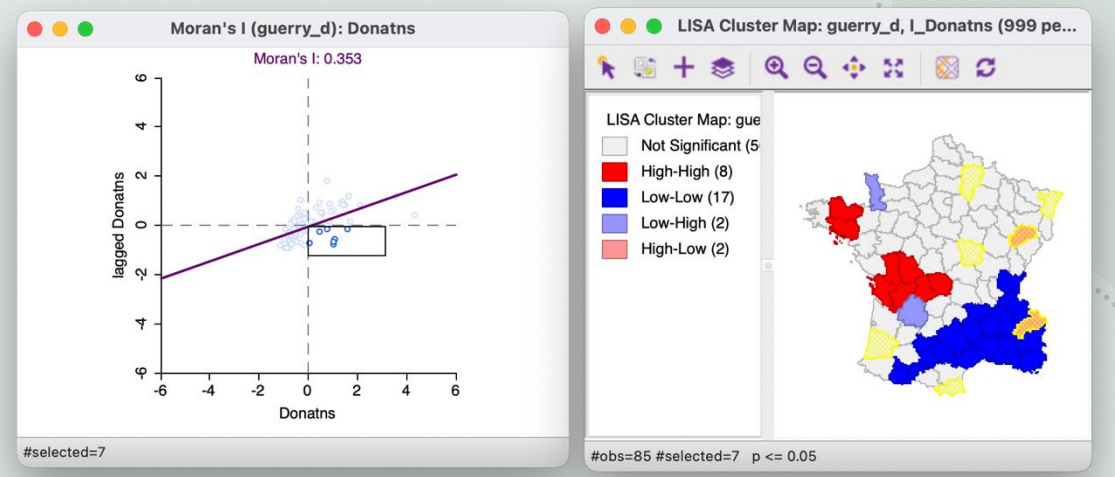
LISA Clusters: Spatial Outliers

- Connection between the **Moran scatter plot** and the **cluster map**

Low-High



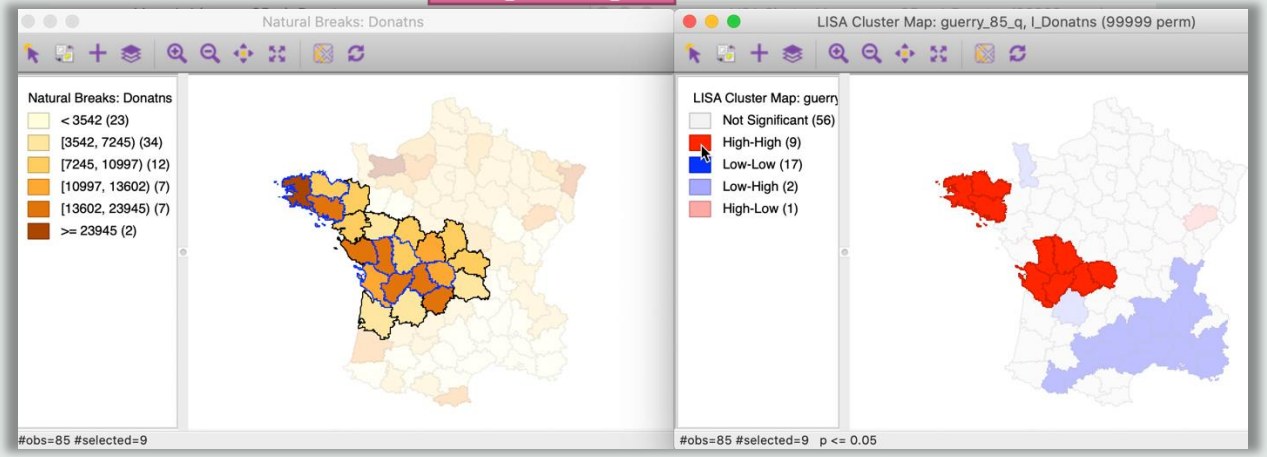
High-Low



LISA Clusters

- Connection between the **natural breaks map** (6 categories) and the **cluster map**

High-High



Low-Low