Spatial Data Analysis

An introduction to spatial autocorrelation and spatial regression analysis

january 2015

www.johanblomme.net

Many research questions require analysis of complex patterns of interrelated social, behavioral, economic and environmental phenomena. In addressing these questions, it is increasingly argued that both spatial thinking and spatial analytical perspectives have an important role to play. Indeed, research on social stratification and inequality, health, mortality and fertility and many other issues depends on the collection and analysis of individual and context-level data.

The geospatial and methodological development environment has changed. The volume, sources and forms of available geospatial data are growing rapidly. The flow of information from a host of sensors has grown exponentially in recent years to the point that many observations can be geo-referenced. Data storage and handling (e.g. cloud computing) change what, how and when we collect data on individuals and their environments.

In a world where information is increasingly seen through geographic filters, the importance of spatial thinking is addressed. More and more instances show that space and place are important elements and stress the leverage of place-based politics. For example, conventional approaches in health research underestimate the contribution of place to disease risk. Several studies reinforce the view how neighborhood context is an important condition of human wellbeing. Place emerges as an important contextual framework for considering a number of critical societal issues. Place as a social context is deeply connected to larger patterns of social advantage and disadvantage.

Since the mid 1990s, there is a renewed interest in the much earlier tradition of spatial demography that focuses on areal aggregates as units of analysis. Trends in technology during the 1980s and 1990s brought sophistication to the world of spacial demography. Factors contributing were :

- U.S. Census Bureau's TIGER files ;
- extensive natural resource, crime and epidemiological databases ;
- powerful GIS software for integrating and mapping spatial data ;
- computing hardware platforms.

These factors altered the way in which spatial demography research was carried out. Other trends that emerged were :

- the use of exploratory spacial data analysis (ESDA) ;
- the role of regression analysis in spatial demography ;
- the special nature of spatial data that requires modification to the standard regression model (e.g. the role of geographically weighted regrssion for exploring spatial variation);
- the need for attention both to global as well as local diagnostic tools.

When analyzing spatial data from a large number of units (e.g. counties), it is the natural inclination of researchers to move from simple descriptive analysis to begin asking questions as : How might these data be modeled ? How well can we account for variability in attribute values among geographic units ?

To answer these questions, analysts turned to multivariate regression modeling, the common methodology in the social sciences. However, the application of the standard regression approach to data tied to spatial units brings spacial complications because "spatial is special". Attention has been drawn to the fact that spatial data require special analytic approaches.

Two properties are particularly important in the analysis of spatial data. The first, spatial dependence, refers to the tendency for spatial data to exhibit spatial autocorrelation. For most social phenomena mapped in space, local proximity usually results in value similarity. High values tend to be located near other high values, while low values tend to be located near other low values, thus exhibiting positive spatial autocorrelation. Less often, high values may tend to be co-located with low values (or vice versa), as islands of dissimilarity (negative spatial autocorrelation).

In either case, the units of analysis in spacial demography likely fail a key assumption of classical statistics : independence among observations. With respect to statistical analysis that presumes such independence (e.g. standard regression analysis), positive autocorrelation means that the spatially autocorrelated observations bring less information to the model estimation process than would the same number of independent observations. The greater the extent of spatial autocorrelation, the more severe is the information loss.

A quick explanation for the presence of spatial autocorrelation can be found in the oft-cited "first law of geography" enunciated by Tobler in 1970 : "Everything is related to everything, but near things are more related than distant things" (Tobler, 1970 : 36). Tobler's first law is somewhat unsatisfying because it doesn't tell us why this phenomenon arises in practice. The answer to this question can only be approximated with models of the spatial process and the analysts's theory about the process.

The second concept refers to spacial heterogeneity, the tendency for phenomena distributed in many spaces to be statistically nonstationary (a lack of stability across space of one or more attribute values). Spacial heterogeneity confounds attempts to generalize because results of an analysis of a limited area will change when the boundaries of the area are shifted.

One of the more recent and fascinating developments in the design of local statistics is the theoretical background and associated software to explore how regression parameters and regression model performance vary across a study region.

Geographically weighted regression (GWR) is similar to a global regression model in that the familiar constant, regression coefficients and error term are all present within the regression specification. There are two ways in which GWR differs from standard (global) regression. First is the fact that a separate regression is carried out at each location (observation) using only the other observations that lie within a user-specified distance from that location. Second, the regression specification includes a statistical device which weights the attributes of nearby geographical units more highly than it does the attributes of distant geographical units. The result is a set of local regression parameters for each geographical unit. The regression is thus localized.

A GWR approach to regression analysis is a highly useful exploratory device for understanding parameter heterogeneity in one's data. The output of GWR enables the researcher to examine and map local parameter estimates and local regression diagnostics, thereby enabling assessment of the utility of the model for various positions of the larger study region.

In the first part of this guide, we provide a general introduction to perform spatial regression and spatial autocorrelation analysis. We use GeoDa, software developed by the Arizona State University's GeoDa Center for geospatial analysis and computation (http://geodacenter.asu.edu). In the second part, we model spatial data with geographically weighted regression to explain local variations in relationships.

CONTENTS

Part 1

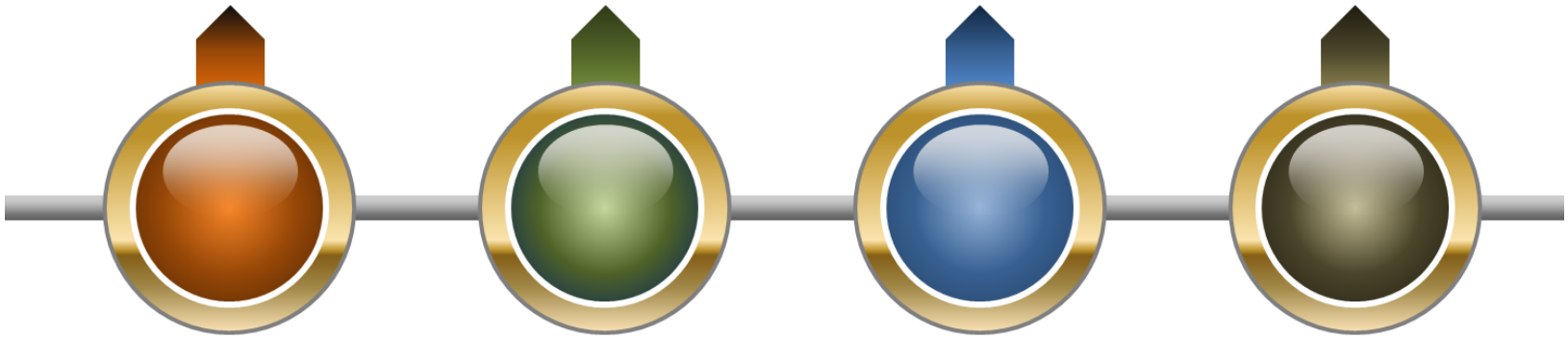
An introduction to spatial autocorrelation and spatial regression analysis with GeoDa

- The development of specialized software for spatial data analysis has seen rapid growth since the late 1980s.

- A substantial collection of spacial data analysis software is available, ranging from niche programs and commercial statistical and GIS packages to open source software environments such as R, Java and Python.

- GeoDa, for example, is the result of the effort to facilitate spatial data analysis. The main objective of the software is to provide the user with a path starting with simple **mapping and geovisualization** moving to **spatial autocorrelation** analysis and ending up with **spatial regression**.

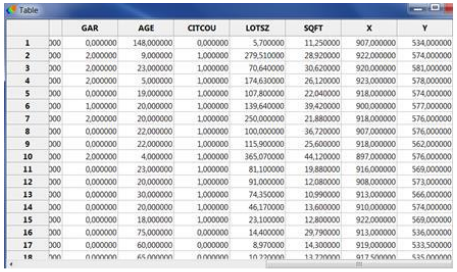
GeoDa Functionality Overview



Spatial Data (transformation)	Mapping and Exploratory Data Analysis	Spatial Autocorrelation	Spatial Regression
input from shapefile (point, polygon) and text data, centroid computation, Thiessen polygons, spatial lag variable construction, rate smoothing	chloropleth maps, percentile map, outlier map, smoothed rate map, histogram, scatterplot	spatial weights creation, Moran scatterplot	OLS and diagnostics, spatial lag model, spatial error model

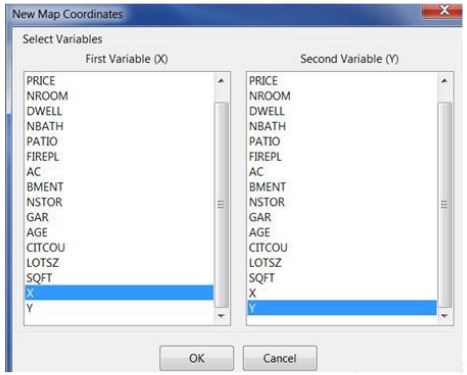
1. Manipulating Spatial Data

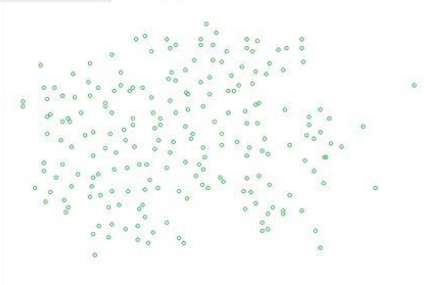
Creating point shape files from .dbf-file



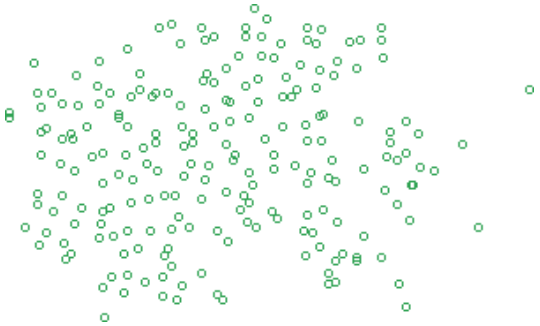
		GAR	AGE	CITCOU	LOTSZ	SQFT	X	Y
1	000	0,000000	148,000000	0,000000	5,700000	11,250000	907,000000	534,000000
2	000	2,000000	9,000000	1,000000	279,510000	28,920000	922,000000	574,000000
3	000	2,000000	23,000000	1,000000	70,640000	30,620000	920,000000	581,000000
4	000	2,000000	5,000000	1,000000	174,630000	26,120000	923,000000	578,000000
5	000	0,000000	19,000000	1,000000	107,800000	22,040000	918,000000	574,000000
6	000	1,000000	20,000000	1,000000	139,640000	39,420000	900,000000	577,000000
7	000	2,000000	20,000000	1,000000	250,000000	21,880000	918,000000	576,000000
8	000	0,000000	22,000000	1,000000	100,000000	36,720000	907,000000	576,000000
9	000	0,000000	22,000000	1,000000	115,900000	25,600000	918,000000	562,000000
10	000	2,000000	4,000000	1,000000	365,070000	44,120000	897,000000	576,000000
11	000	0,000000	23,000000	1,000000	81,100000	19,880000	916,000000	569,000000
12	000	0,000000	20,000000	1,000000	91,000000	12,080000	908,000000	573,000000
13	000	0,000000	30,000000	1,000000	74,350000	10,990000	913,000000	566,000000
14	000	0,000000	20,000000	1,000000	46,170000	13,600000	910,000000	574,000000
15	000	0,000000	18,000000	1,000000	23,100000	12,800000	922,000000	569,000000
16	000	0,000000	75,000000	0,000000	14,400000	29,790000	913,000000	536,000000
17	000	0,000000	60,000000	0,000000	8,970000	14,300000	919,000000	533,500000
18	000	0,000000	65,000000	0,000000	10,220000	13,230000	917,500000	535,000000

Tools → Shape → Points from table



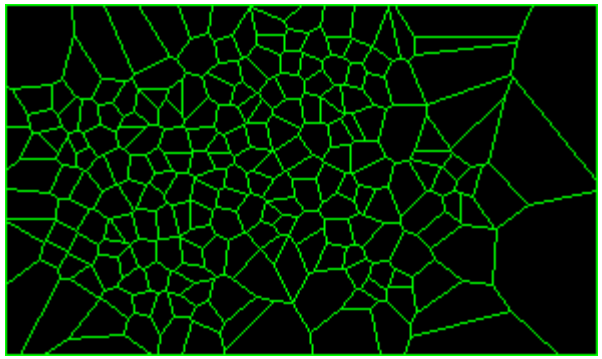


Creating Thiessen polygons as shape files



Thiessen polygons are created as a polygon shape file derived from a point shape file. Each Thiessen polygon encloses the original points in such a way that all points in a polygon are closer to the enclosed point than any other point. This correspons to the notion **of geographic market area**.

Tools → Shape → Points to polygon



Thiessen polygons allow the computation of contiguity based spatial weights for point data, using the boundaries of the polygons to establish contiguity.

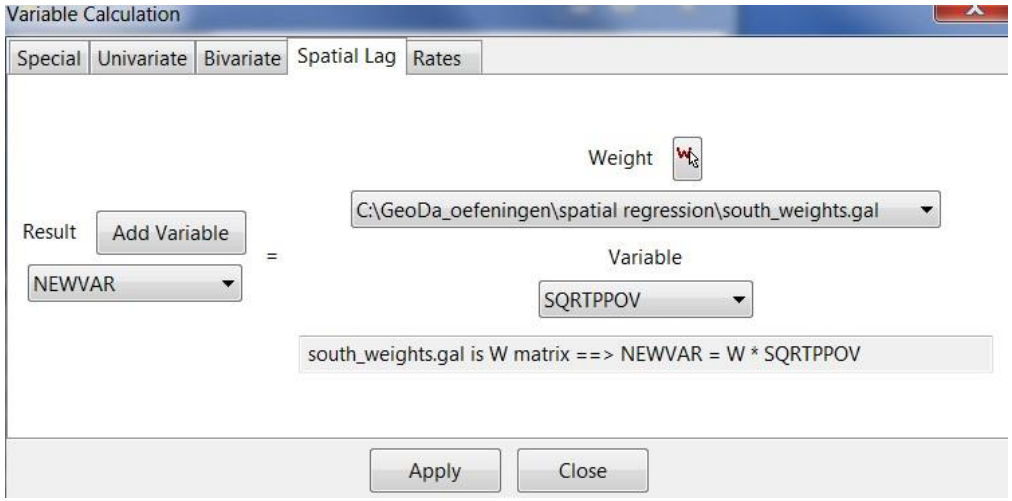
Table

	POLYID	AREA	PERIMETER	STATION	PRICE	NROOM
1	1	39,379239	23,625421	1		
2	2	19,749913	17,556413	2		
3	3	23,291446	25,572734	3		
4	4	40,577874	24,947434	4		
5	5	20,400000	18,465790	5		
6	6	41,450719	26,483263	6		
7	7	31,168254	23,806590	7		
8	8	50,452172	28,729254	8		
9	9	25,650750	19,855806	9		
10	10	76,174252	37,220197	10		
11	11	32,030299	22,232463	11		
12	12	24,584678	19,660918	12		
13	13	28,317198	21,469213	13		
14	14	36,007485	25,968686	14		
15	15	22,903019	18,652667	15		
16	16	35,494474	23,193386	16		
17	17	17.303112	16.388330	17		

Area and perimeter calculations are only supported for projected coordinates (Euclidean distance). For point shape files in unprojected latitude and longitude, the results will not be correct.

Computing spatially lagged variables

Spatially lagged variables are weighted averages of the values for neighboring locations, as specified by a spatial weights matrix.

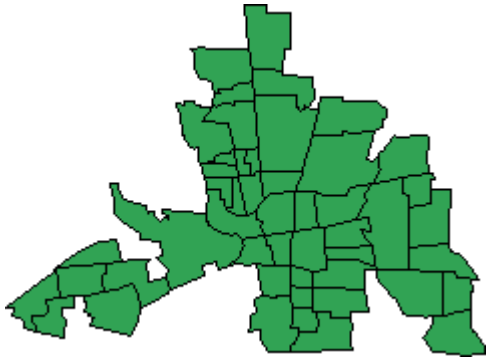


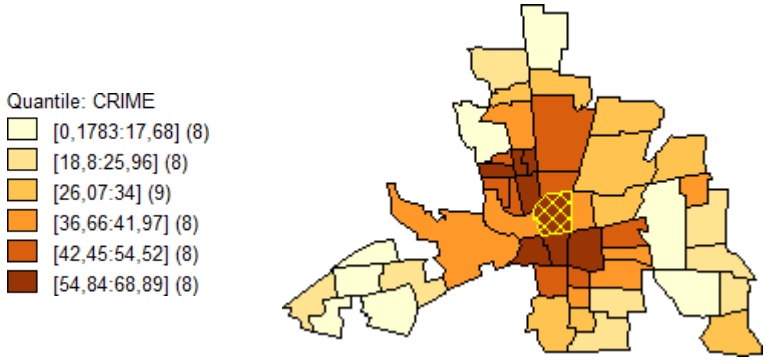
The changes and additions made to a table only reside in memory and are not permanent. In order To make them permanent, the table must be saved to a new file :

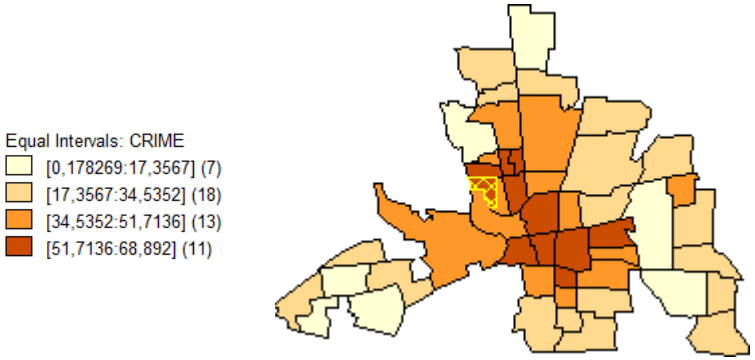
File → Save as → Shapefile name to save as

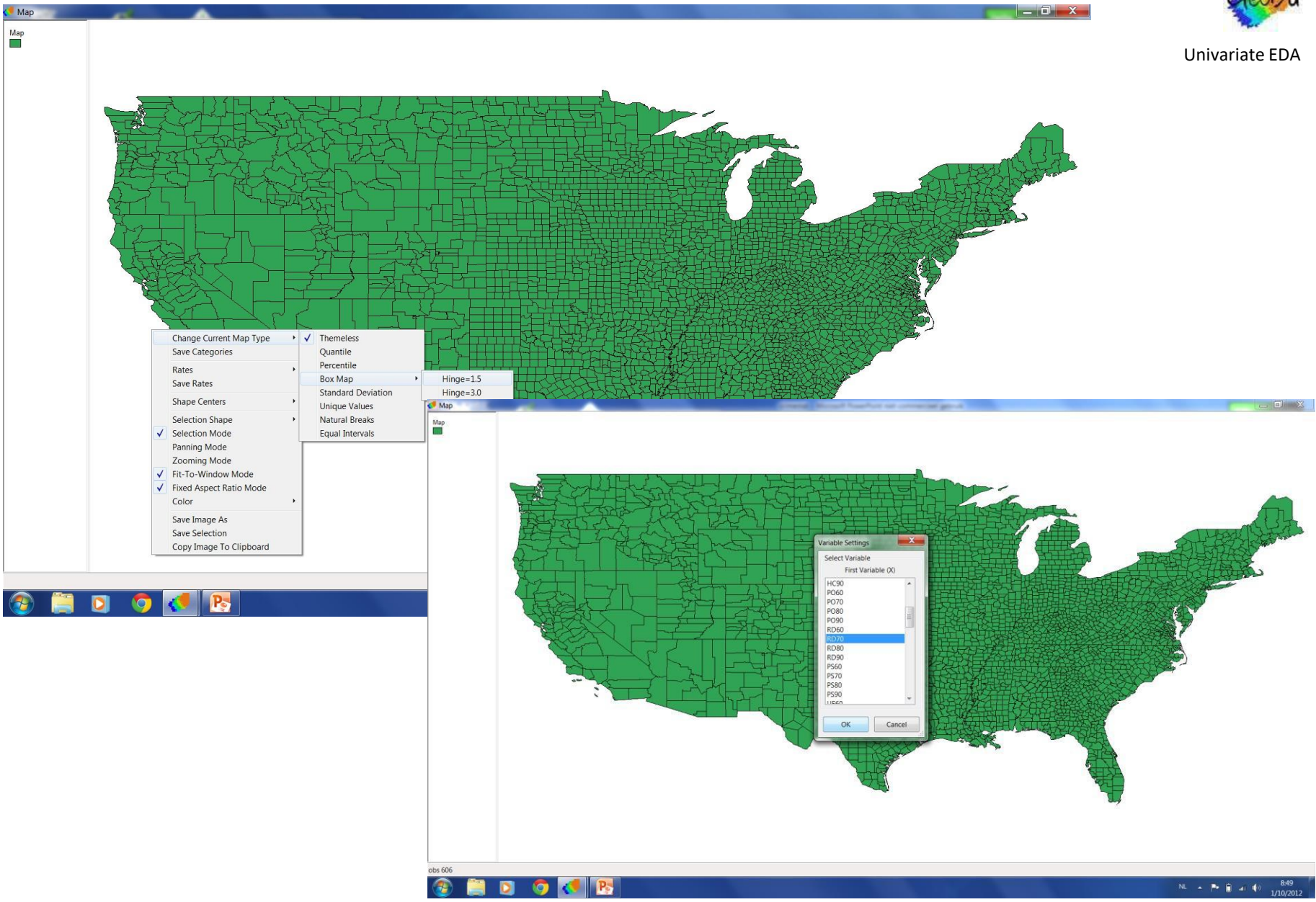
This results in three files to be saved, with file extensions .shp, .shx and .dbf.

2. Mapping and Exploratory Data Analysis



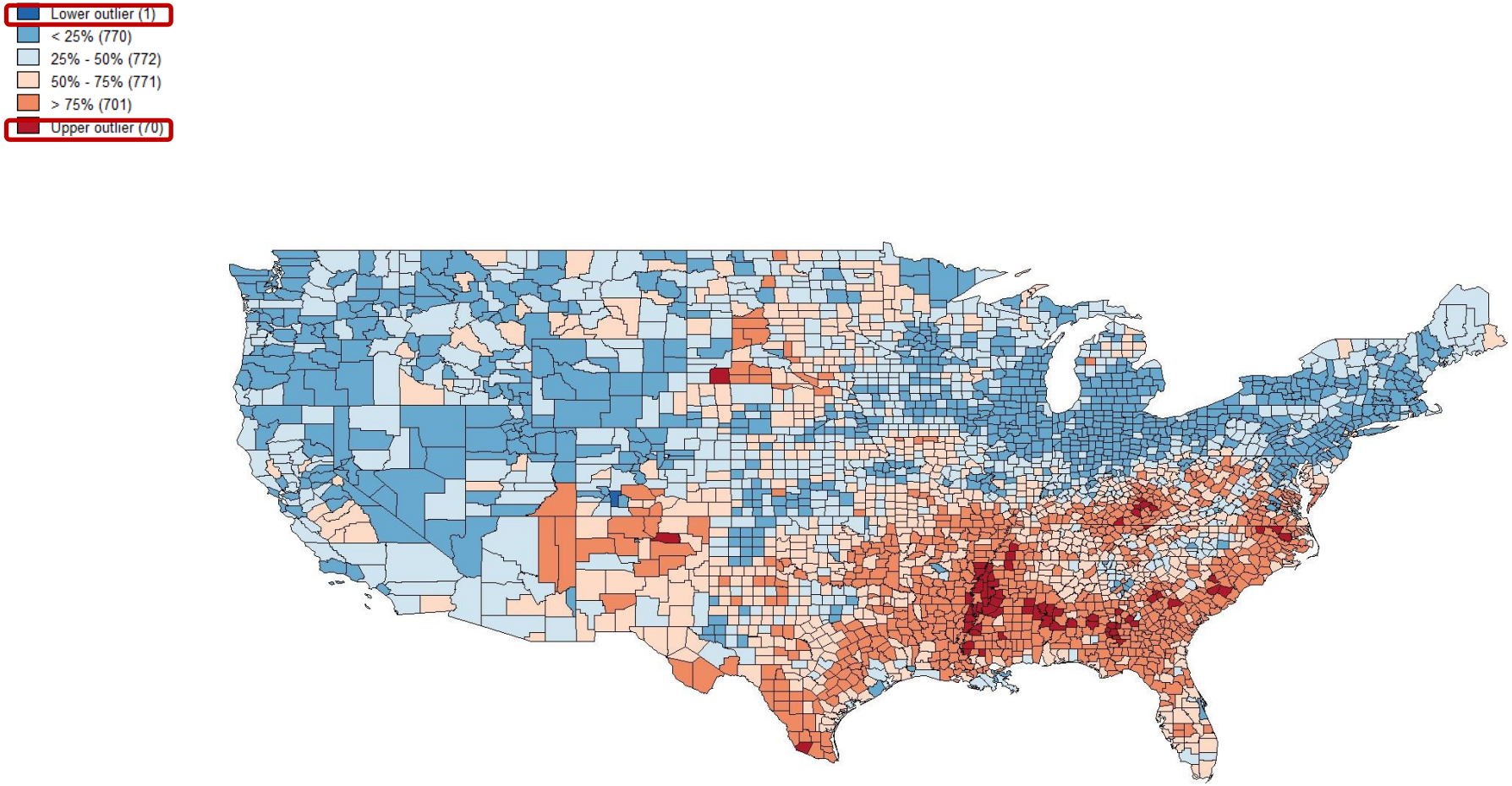

Quantile: CRIME
[0,1783:17,68] (8)
[18,8:25,96] (8)
[26,07:34] (9)
[36,66:41,97] (8)
[42,45:54,52] (8)
[54,84:68,89] (8)


Equal Intervals: CRIME
[0,178269:17,3567] (7)
[17,3567:34,5352] (18)
[34,5352:51,7136] (13)
[51,7136:68,892] (11)


Map
Map
Change Current Map Type
Save Categories
Rates
Save Rates
Shape Centers
Selection Shape
Selection Mode
Panning Mode
Zooming Mode
Fit-To-Window Mode
Fixed Aspect Ratio Mode
Color
Save Image As
Save Selection
Copy Image To Clipboard
Themeless
Quantile
Percentile
Box Map
Standard Deviation
Unique Values
Natural Breaks
Equal Intervals
Hinge=1.5
Hinge=3.0
Map
Map
Variable Settings
Select Variable
First Variable (X)
HC90
PO60
PO70
PO80
PO90
RD60
RD70
RD80
RD90
PS60
PS70
PS80
PS90
OK
Cancel
obs 606
8:49
1/10/2012

resource deprivation index (1970)

Hinge value of 1.5 = 1.5 times the interquartile range to define outliers



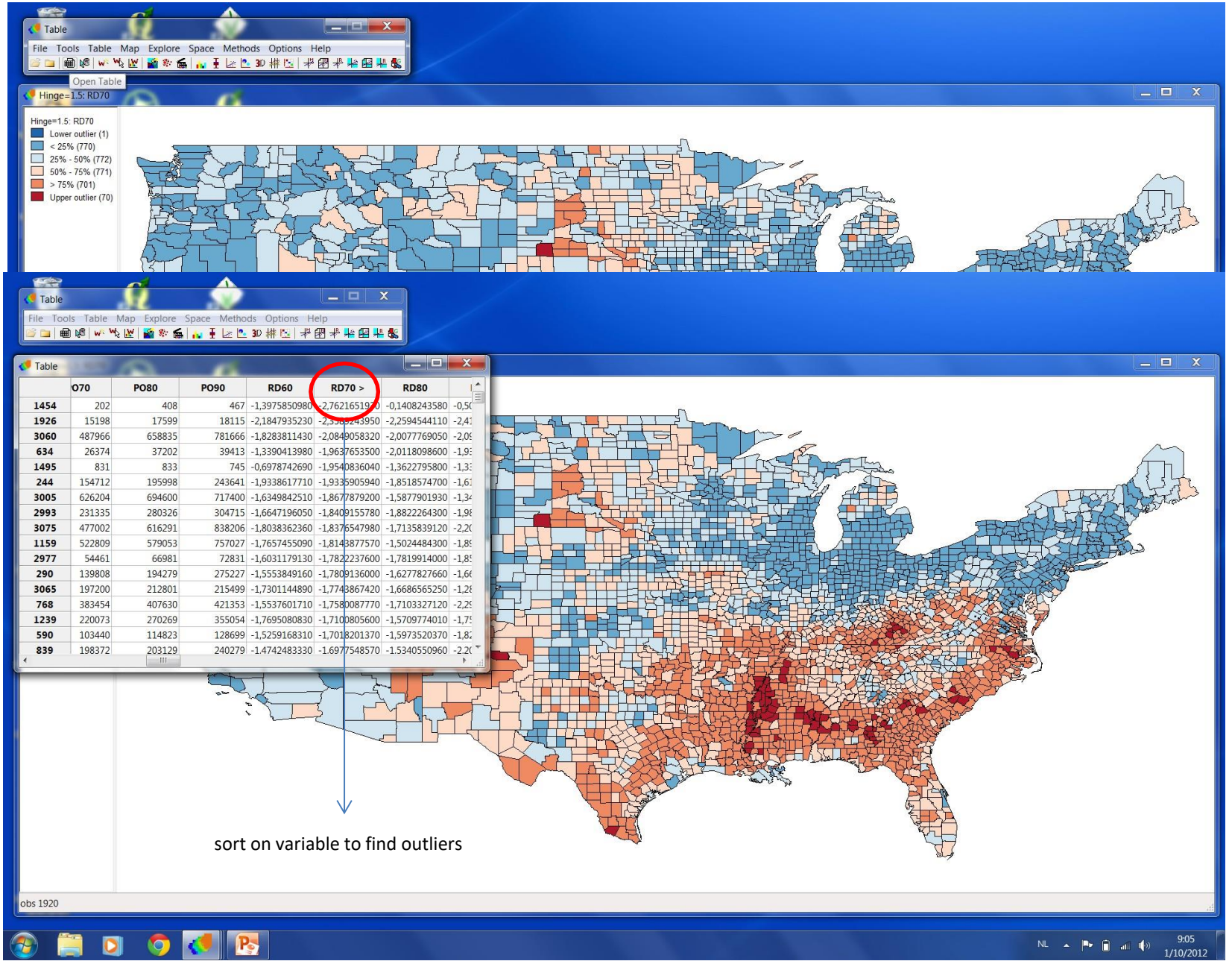


Table
File Tools Table Map Explore Space Methods Options Help
Open Table
Hinge=1.5: RD70
Hinge=1.5: RD70
Lower outlier (1)
< 25% (770)
25% - 50% (772)
50% - 75% (771)
> 75% (701)
Upper outlier (70)
Table
File Tools Table Map Explore Space Methods Options Help
Table
O70 PO80 PO90 RD60 RD70 > RD80
1454 202 408 467 -1,3975850980 -2,7621651990 -0,1408243580
1926 15198 17599 18115 -2,1847935230 -2,2594544110
3060 487966 658835 781666 -1,8283811430 -2,0849058320 -2,0077769050
634 26374 37202 39413 -1,3390413980 -1,9637653500 -2,0118098600
1495 831 833 745 -0,6978742690 -1,9540836040 -1,3622795800
244 154712 195998 243641 -1,9338617710 -1,9335905940 -1,8518574700
3005 626204 694600 717400 -1,6349842510 -1,8677879200 -1,5877901930
2993 231335 280326 304715 -1,6647196050 -1,8409155780 -1,8822264300
3075 477002 616291 838206 -1,8038362360 -1,8376547980 -1,7135839120
1159 522809 579053 757027 -1,7657455090 -1,8143877570 -1,5024484300
2977 54461 66981 72831 -1,6031179130 -1,7822237600 -1,7819914000
290 139808 194279 275227 -1,5553849160 -1,7809136000 -1,6277827660
3065 197200 212801 215499 -1,7301144890 -1,7743867420 -1,6686565250
768 383454 407630 421353 -1,5537601710 -1,7580087770 -1,7103327120
1239 220073 270269 355054 -1,7695080830 -1,7100805600 -1,5709774010
590 103440 114823 128699 -1,5259168310 -1,7018201370 -1,5973520370
839 198372 203129 240279 -1.4742483330 -1.6977548570 -1.5340550960
sort on variable to find outliers
obs 1920
NL
9:05
1/10/2012

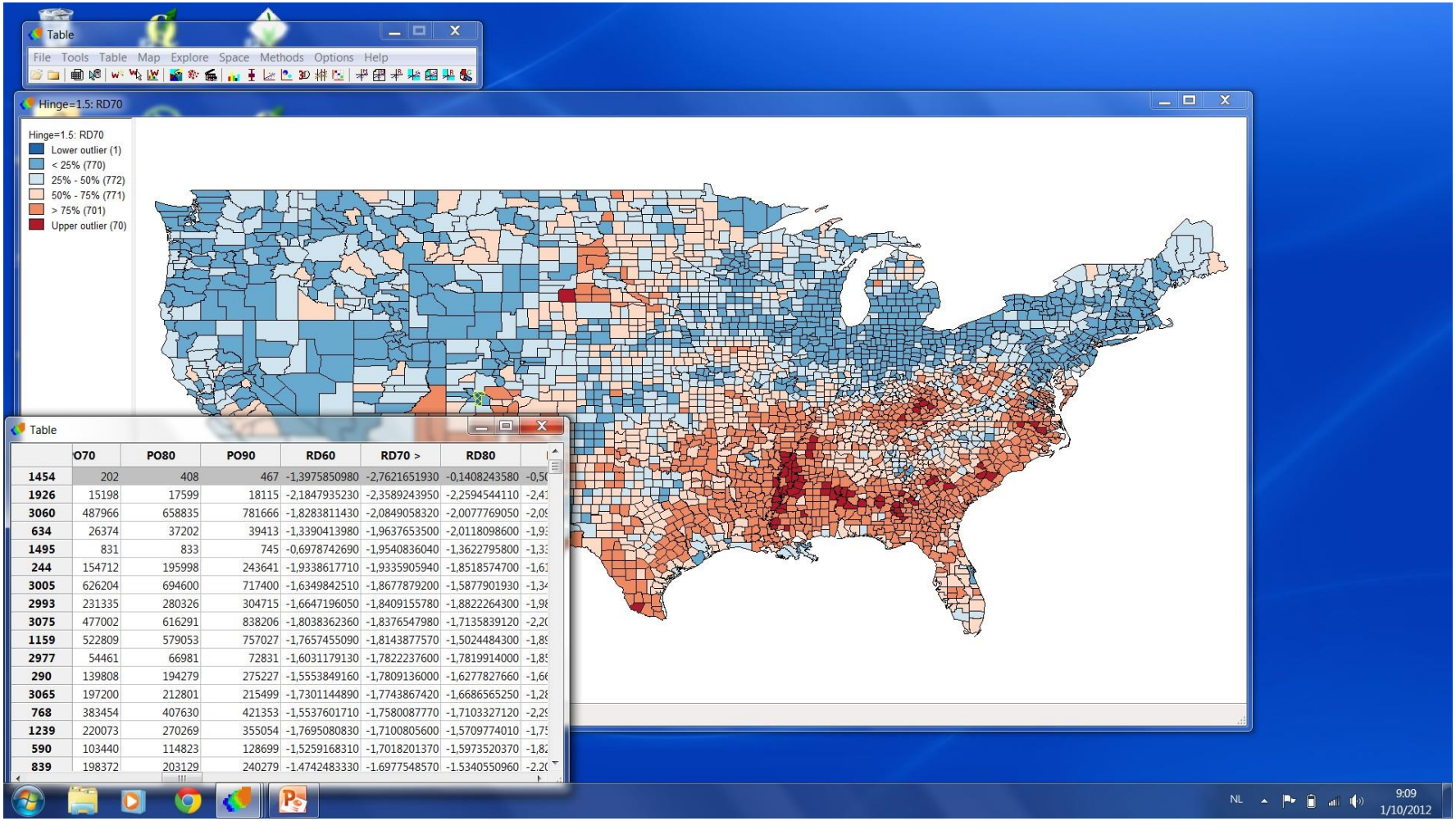
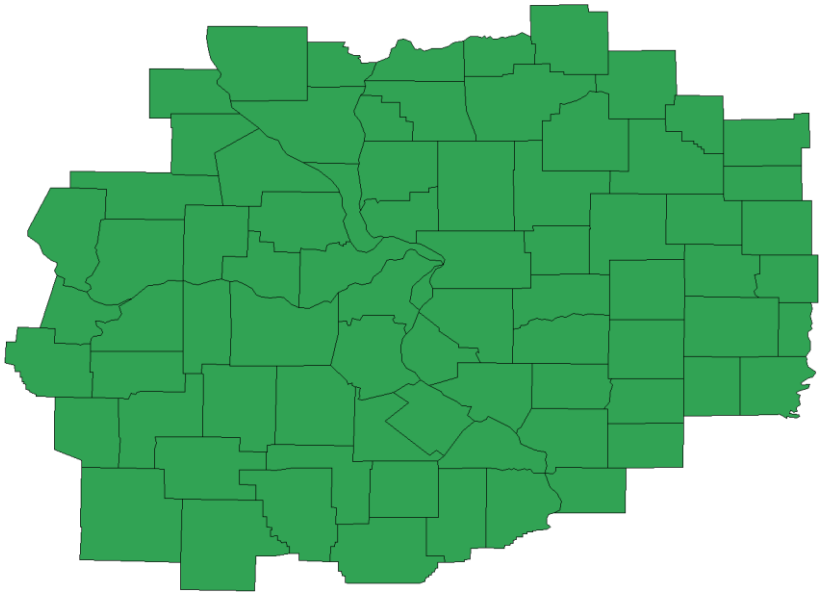
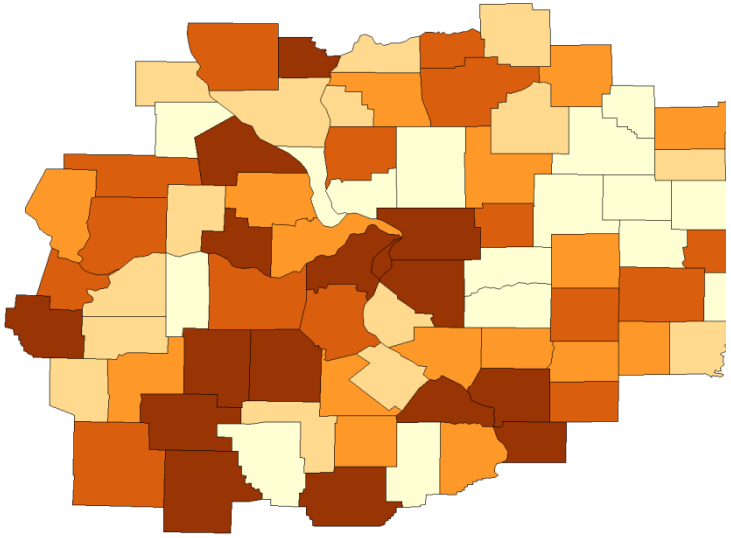


Table
File Tools Table Map Explore Space Methods Options Help
Hinge=1.5: RD70
Hinge=1.5: RD70
Lower outlier (1)
< 25% (770)
25% - 50% (772)
50% - 75% (771)
> 75% (701)
Upper outlier (70)
Table
O70 PO80 PO90 RD60 RD70 > RD80
1454 202 408 467 -1,3975850980 -2,7621651930 -0,1408243580
1926 15198 17599 18115 -2,1847935230 -2,3589243950 -2,2594544110
3060 487966 658835 781666 -1,8283811430 -2,0849058320 -2,0077769050
634 26374 37202 39413 -1,3390413980 -1,9637653500 -2,0118098600
1495 831 833 745 -0,6978742690 -1,9540836040 -1,3622795800
244 154712 195998 243641 -1,9338617710 -1,9335905940 -1,8518574700
3005 626204 694600 717400 -1,6349842510 -1,8677879200 -1,5877901930
2993 231335 280326 304715 -1,6647196050 -1,8409155780 -1,8822264300
3075 477002 616291 838206 -1,8038362360 -1,8376547980 -1,7135839120
1159 522809 579053 757027 -1,7657455090 -1,8143877570 -1,5024484300
2977 54461 66981 72831 -1,6031179130 -1,7822237600 -1,7819914000
290 139808 194279 275227 -1,5553849160 -1,7809136000 -1,6277827660
3065 197200 212801 215499 -1,7301144890 -1,7743867420 -1,6686565250
768 383454 407630 421353 -1,5537601710 -1,7580087770 -1,7103327120
1239 220073 270269 355054 -1,7695080830 -1,7100805600 -1,5709774010
590 103440 114823 128699 -1,5259168310 -1,7018201370 -1,5973520370
839 198372 203129 240279 -1.4742483330 -1.6977548570 -1.5340550960
NL
9:09
1/10/2012

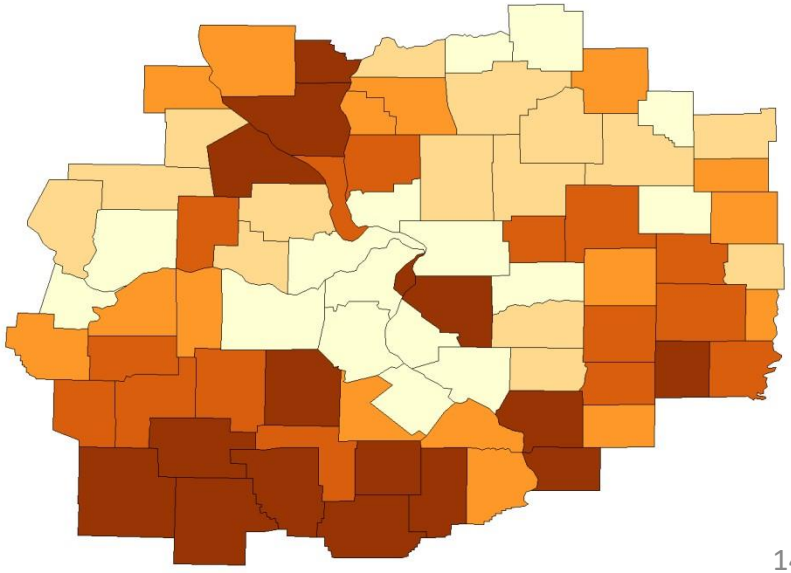
Homicide data for counties around St Louis



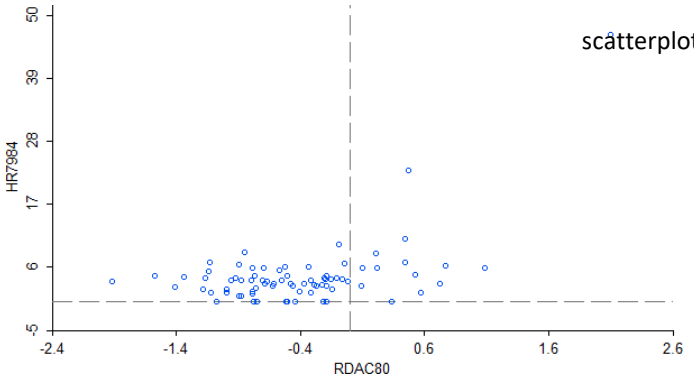
Quintile map homicide rate



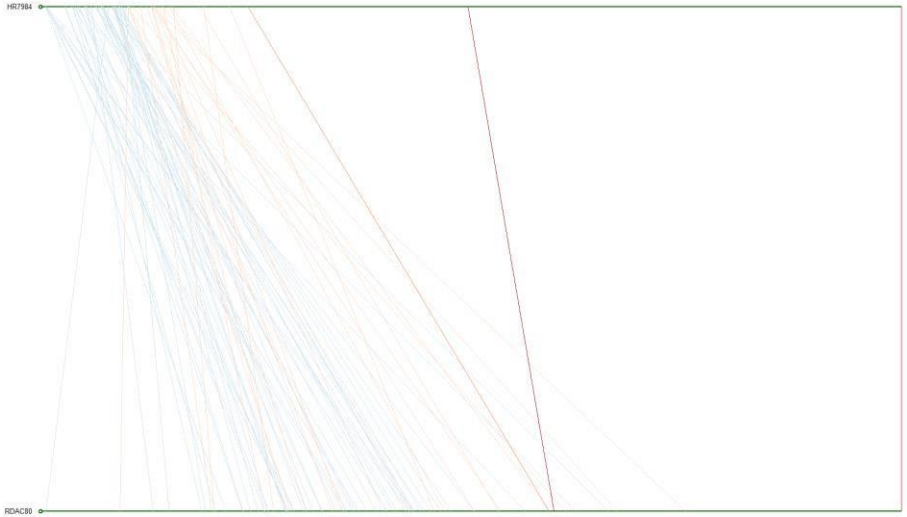
Quintile map resource deprivation



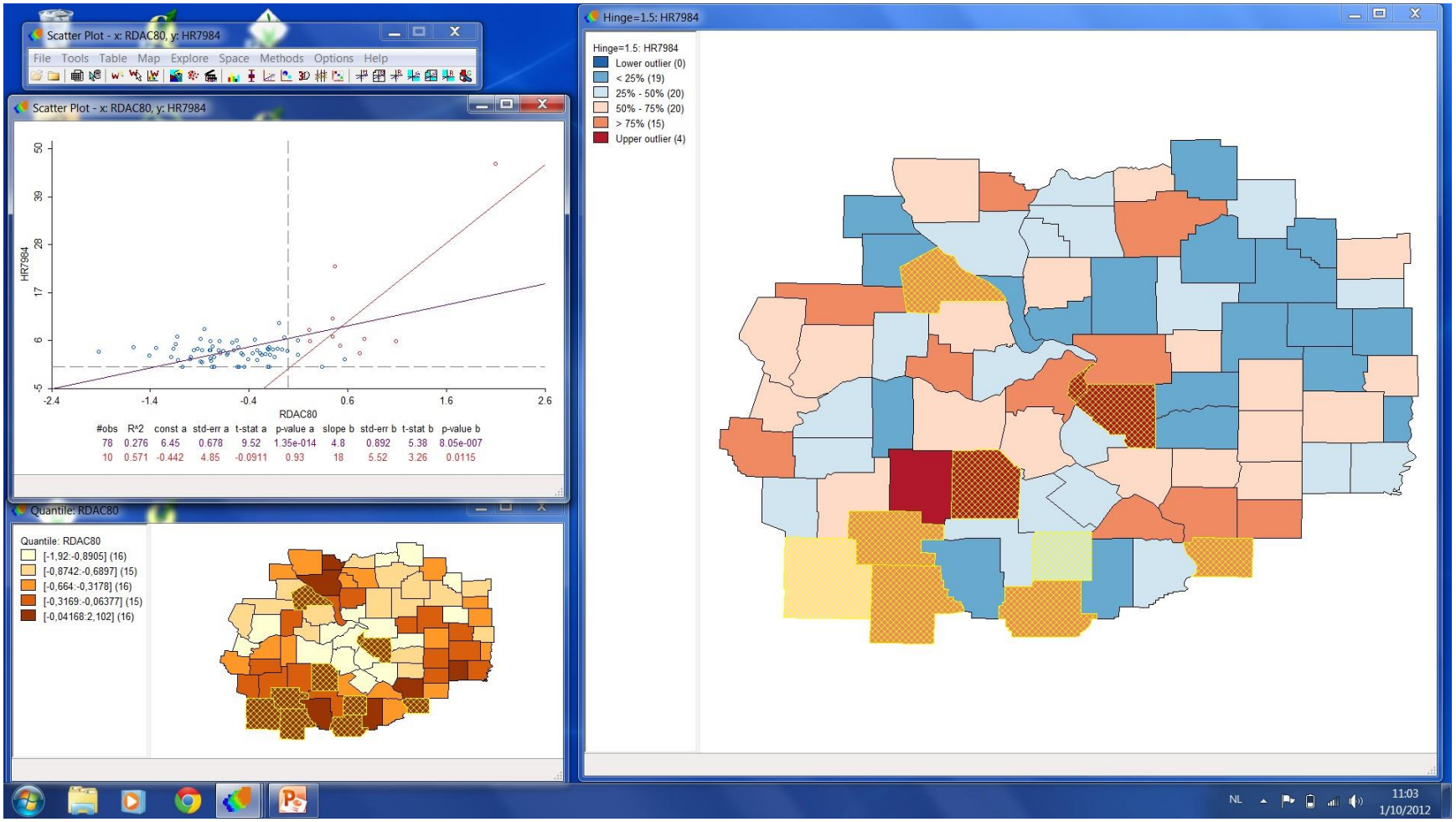
scätterplot



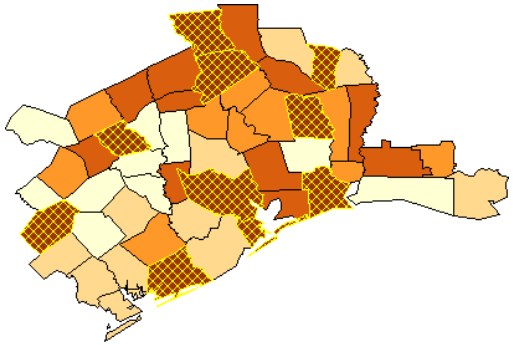
parallel coordinate plot (PCP)

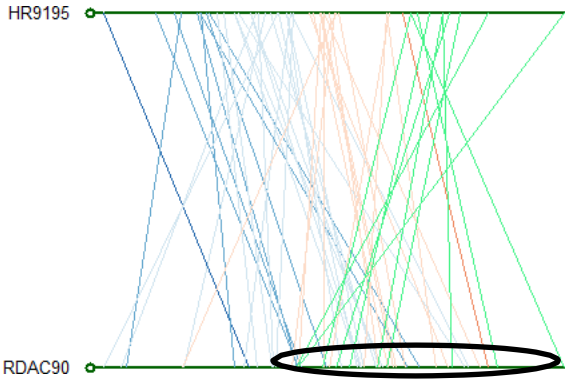


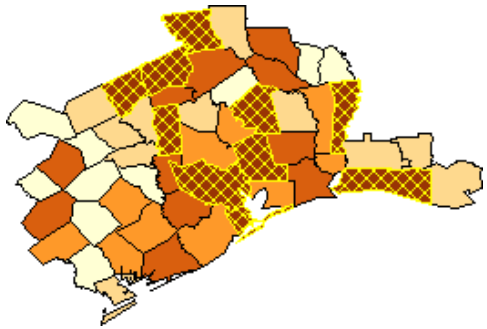
Linking and brushing

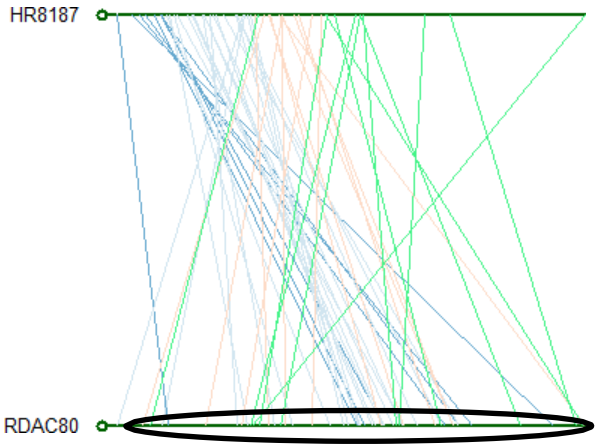


Analyzing changes over time :

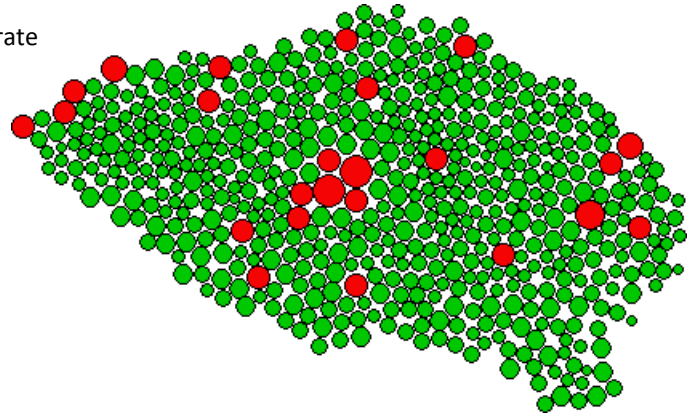


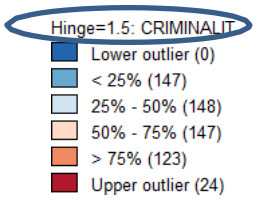


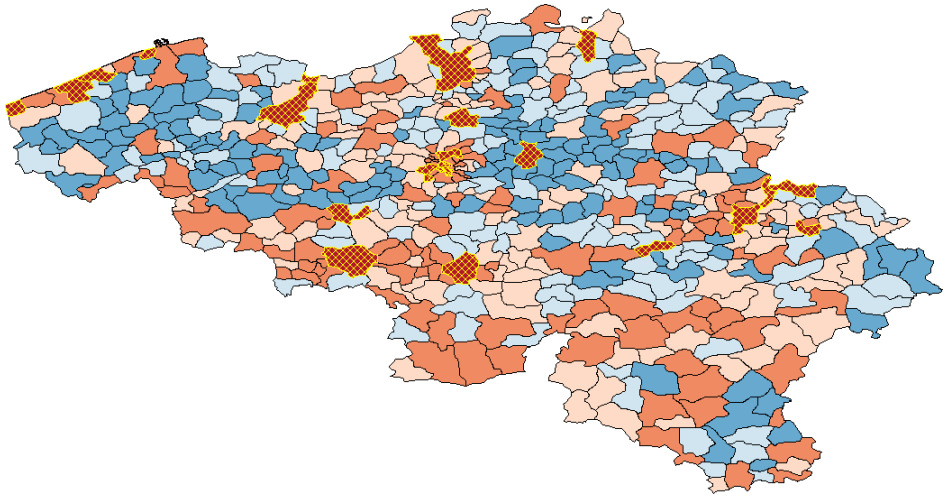


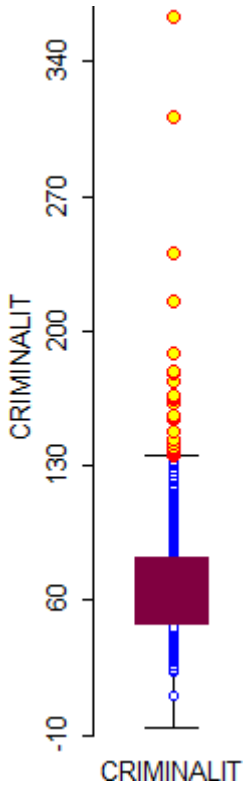


Cartogram crime rate

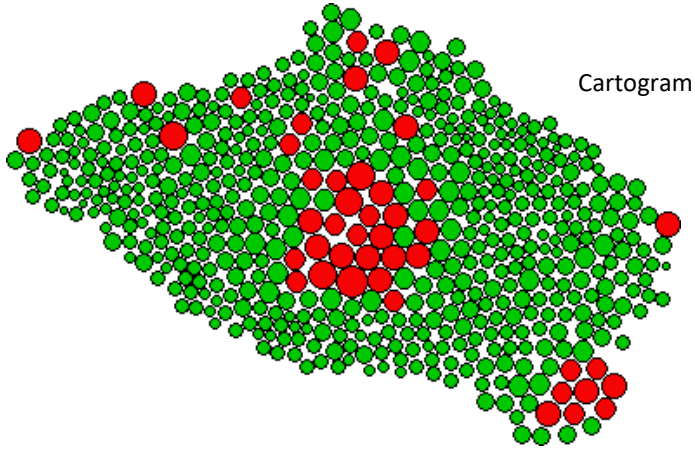


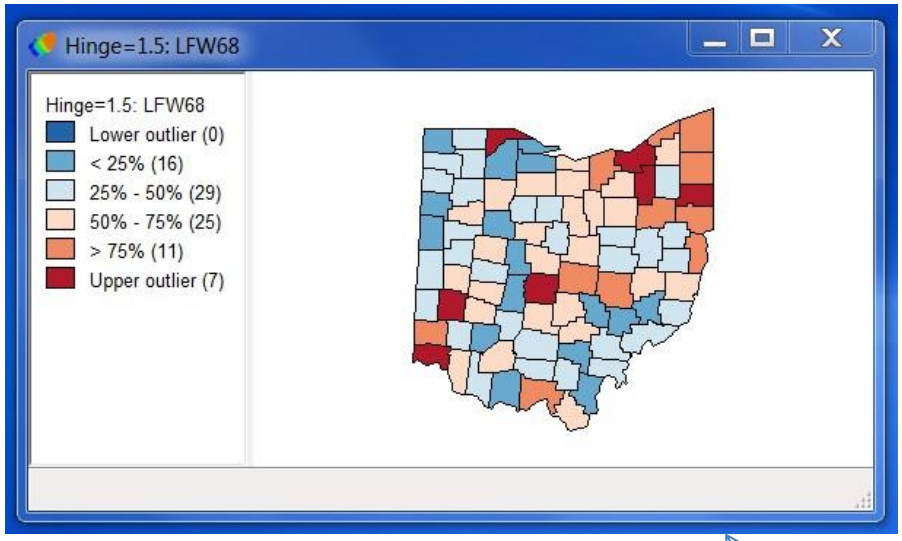






Cartogram Gini inequality

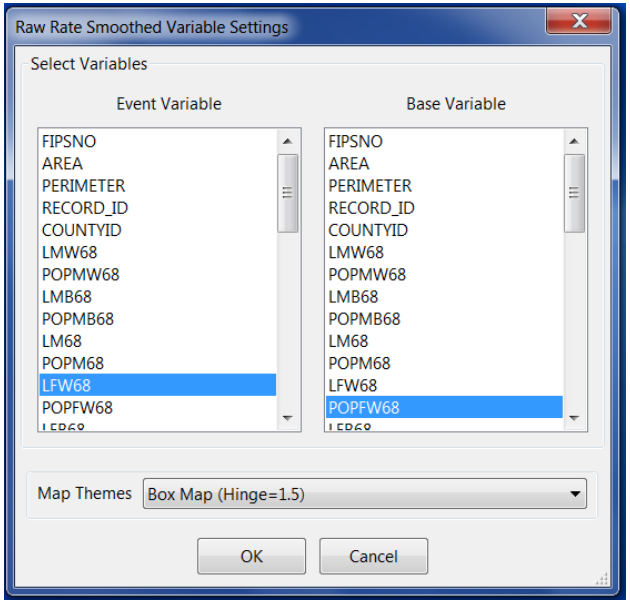


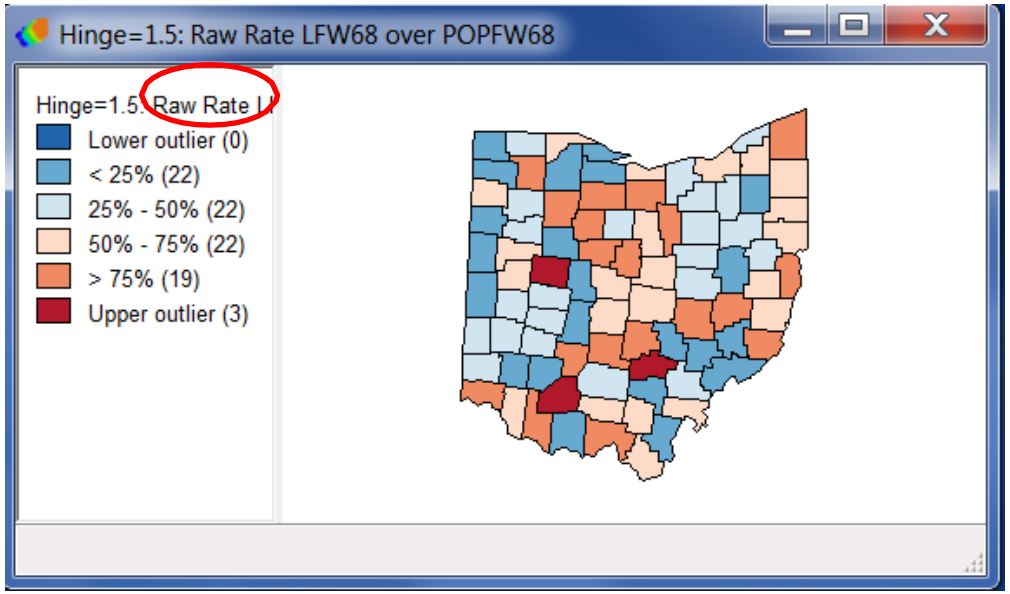


Ohio counties, total lung cancer deaths for White females, 1968

selecting a rate variable from the data set (reveals the problem of variance instability)

both the event and the population at risk are specified and the rate is calculated on the fly

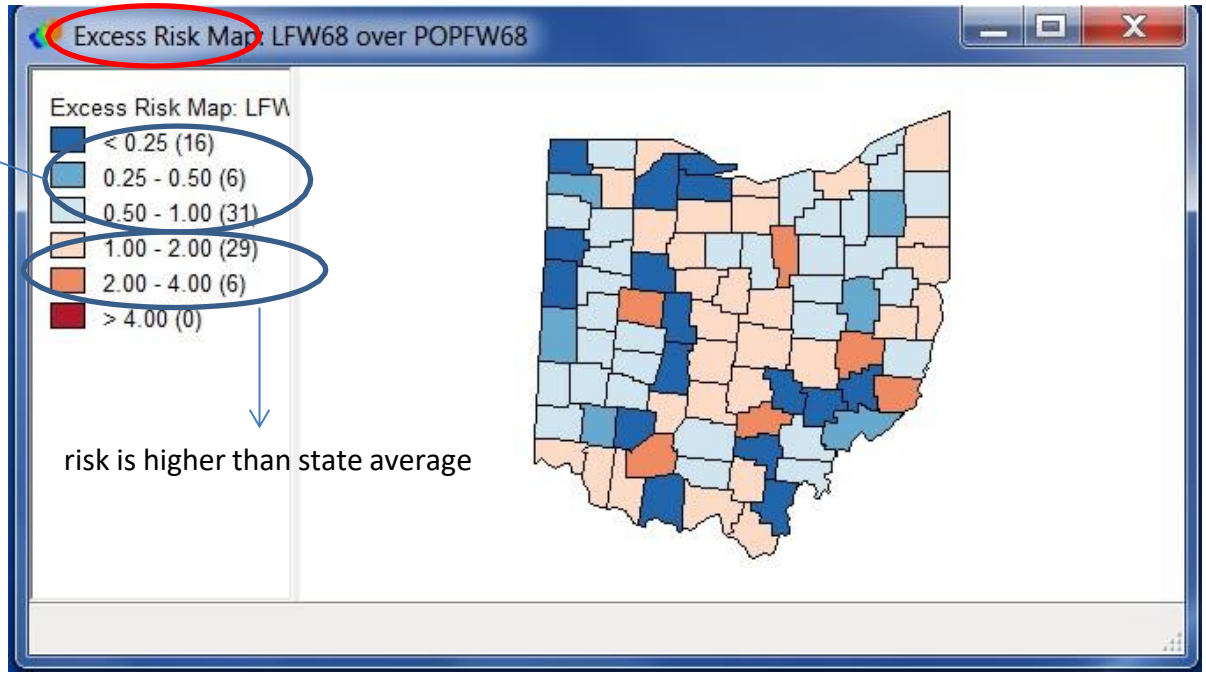




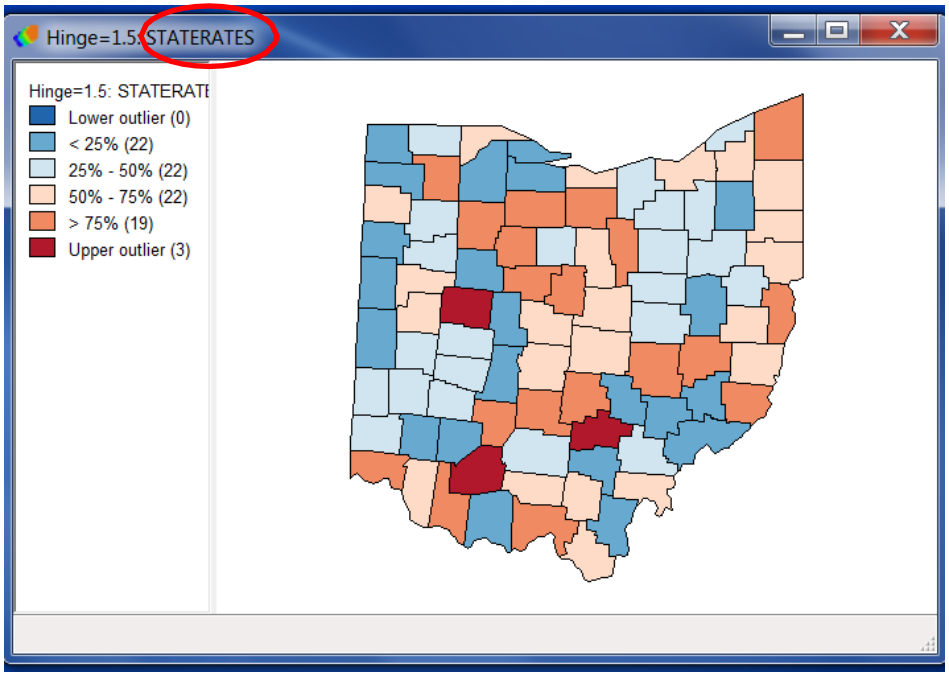
A commonly used notion in public health analysis is the concept of a standardized mortality rate (SMR), or, the ratio of the observed moratlity rate to a national (or regional) standard. GeoDa implements this in the form of an **excess risk** map.

The excess rate is the ratio of the observed rate to the average rate computed for all the data. Note that this average is not the average of the county rates (instead, it is calculated as the ratio of the total sum of all vents over the total sum of all populations at risk).

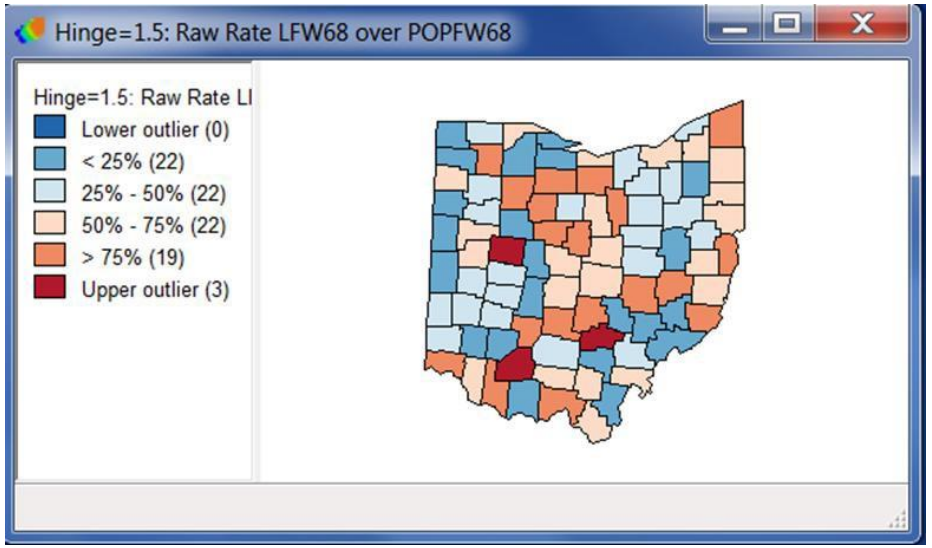
risk is lower than state average



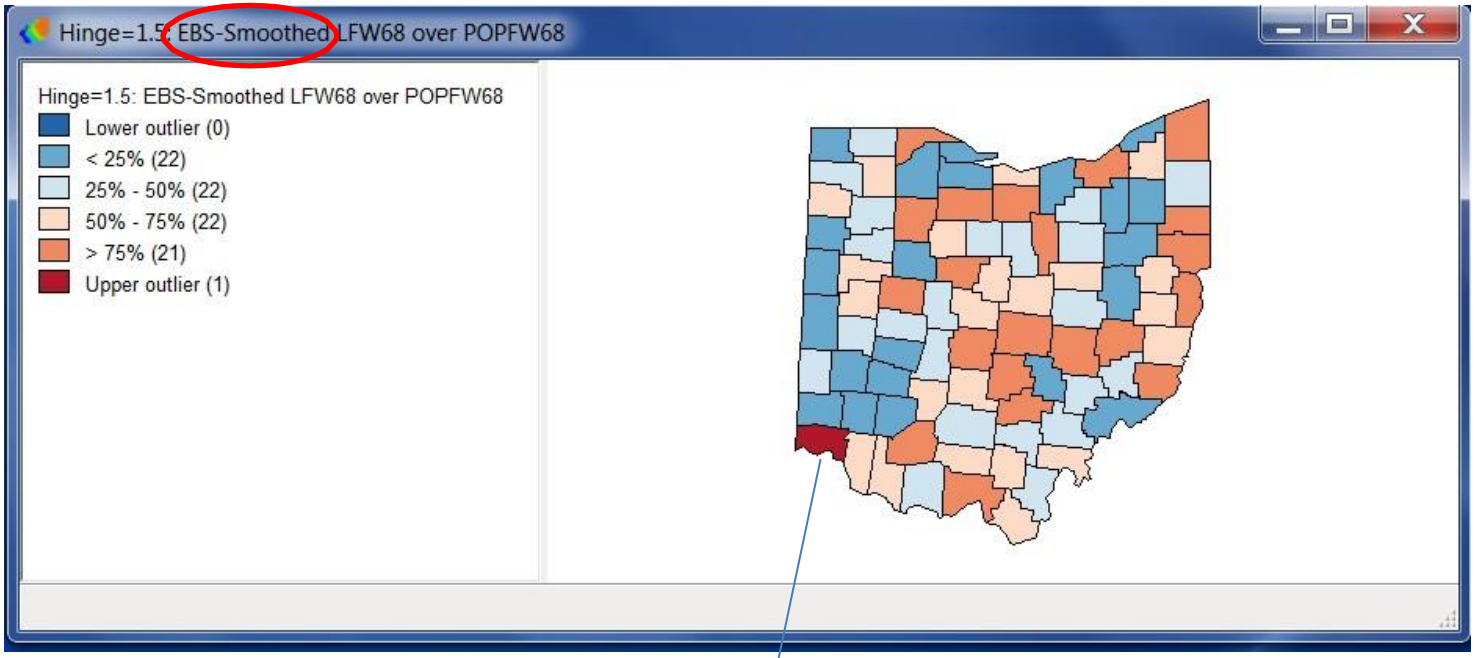
saved to the table (right click on previous map)



no difference between rescaled raw rates and raw rates



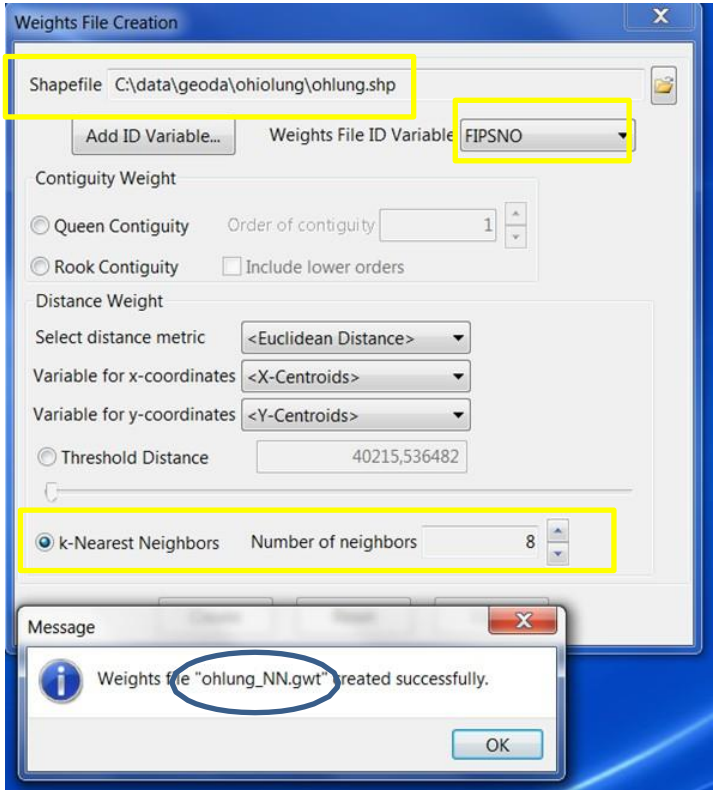
Empirical Bayes consists of computing a weighted average between the raw rate for each county and the state average, with weights proportional to the underlying population at risk. Small conties will tend to have their rates adjusted considerably, whereas for larger counties the rates will barely change.

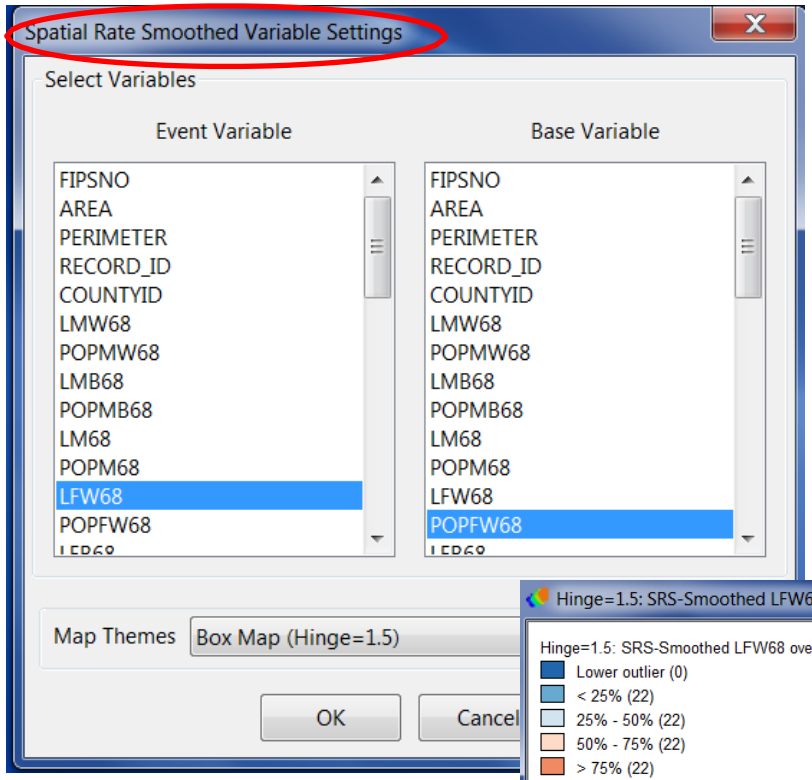


a new outlier is added

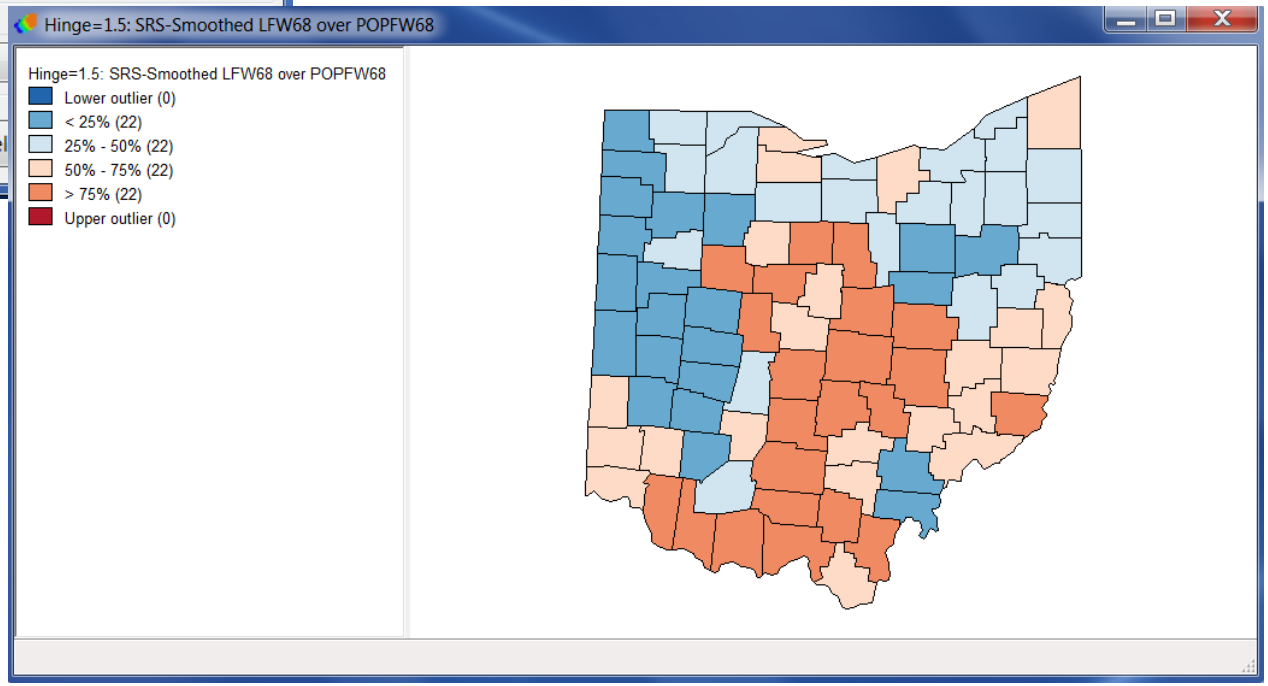
Spatial rate smoothing consists of computing the rate in a moving window that includes the county as well as its neighbors. In GeoDa neighbors are defined by means of a **spatial weights file**.

We will construct a simple spacial weights file consisting of the 8 nearest neighbors for each county in the Ohio shapefile.





A spatially smooted box map emphasizes broad regional patterns. Note how there are no more outliers.



3. Spatial autocorrelation

- **Spatial autocorrelation is a measure of spacial dependency that quantifies the degree of spatial clustering or dispersion in the values of a variable measured across a set of locations.**

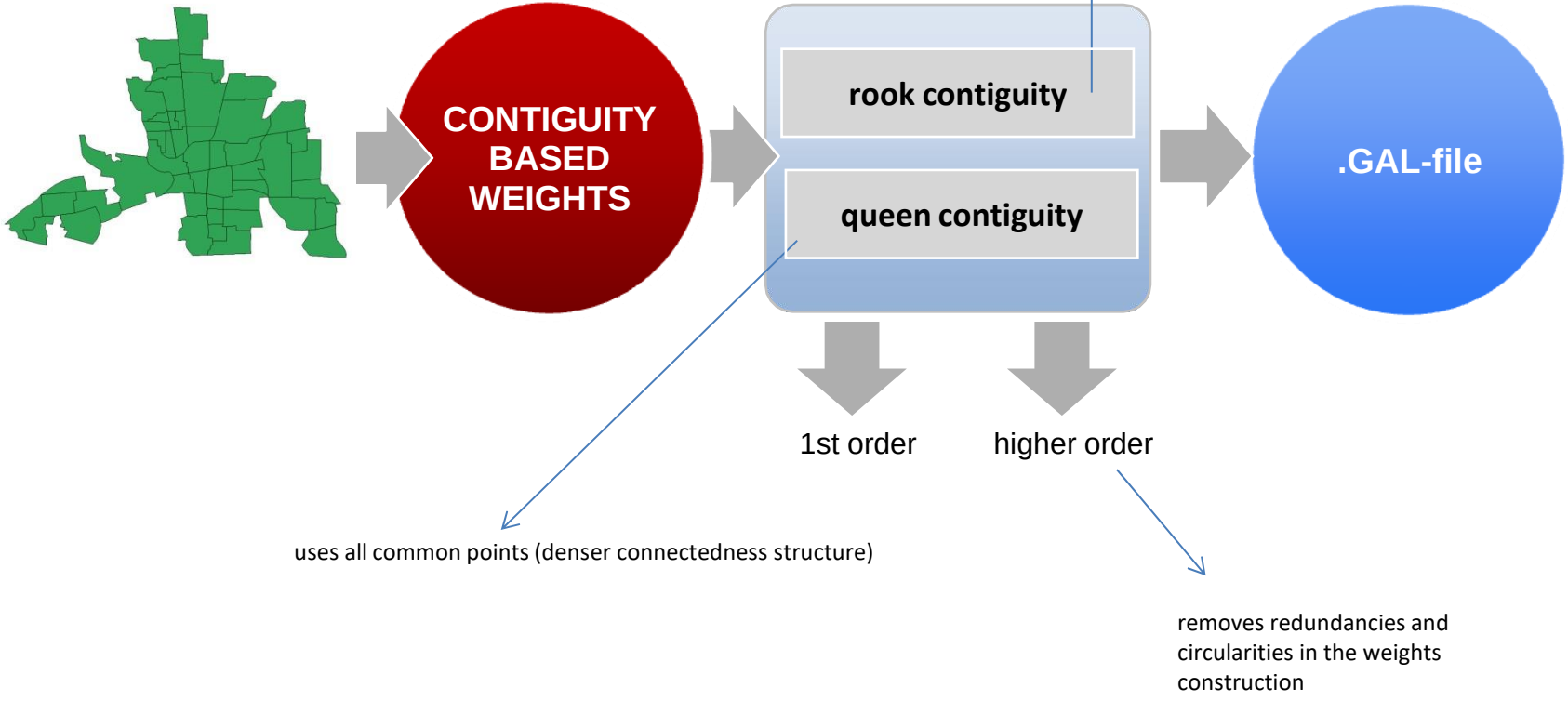
- There are two basic types of spatial autocorrelation statistics : **global measures** identify whether the values of a variable exhibit a significant overall pattern of regional clustering, whereas **local measures** identify the location of significant high and low value clusters.

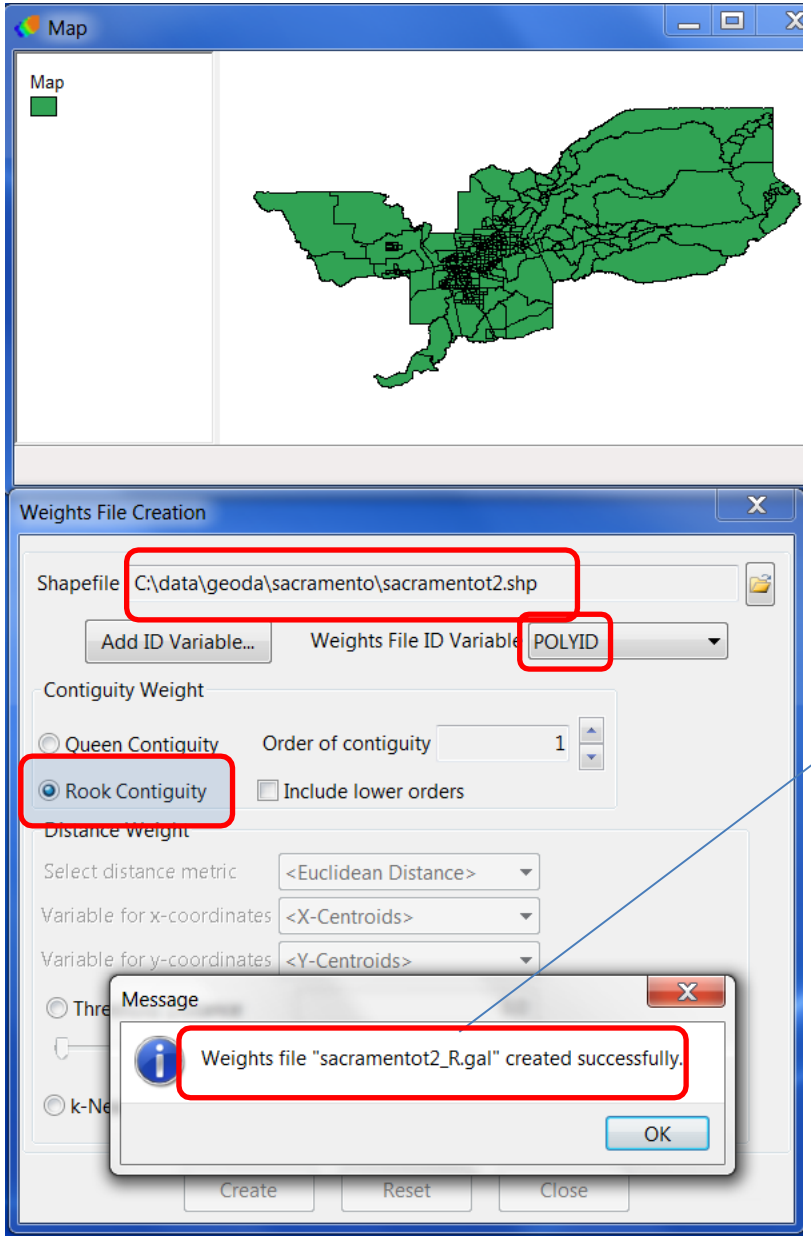
- **Basics** : ***Steps in determining the extent of spatial autocorrelation*** :
  - choose a neighborhood criterion : which areas are linked ?
  - assign weights to the areas that are linked : create a spatial weights matrix
  - run statistical tests, using weights matrix, to examine spatial autocorrelation

- Spacial autocorrelation measures the correlation of a variable with itself through space. Spacial autocorrelation can be positive or negative. Positive spatial autocorrelation occurs when similar values occur near one another. Negative spatial autocorrelation occurs when dissimilar values occur near one another.

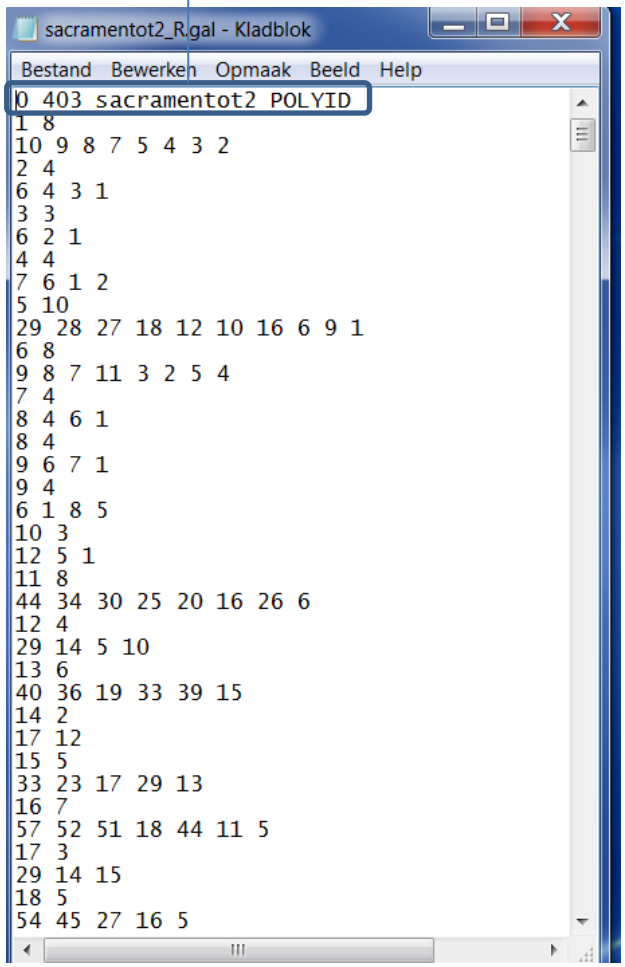
- **Spacial weights are essential for the computation of spacial autocorrelation statistics**.

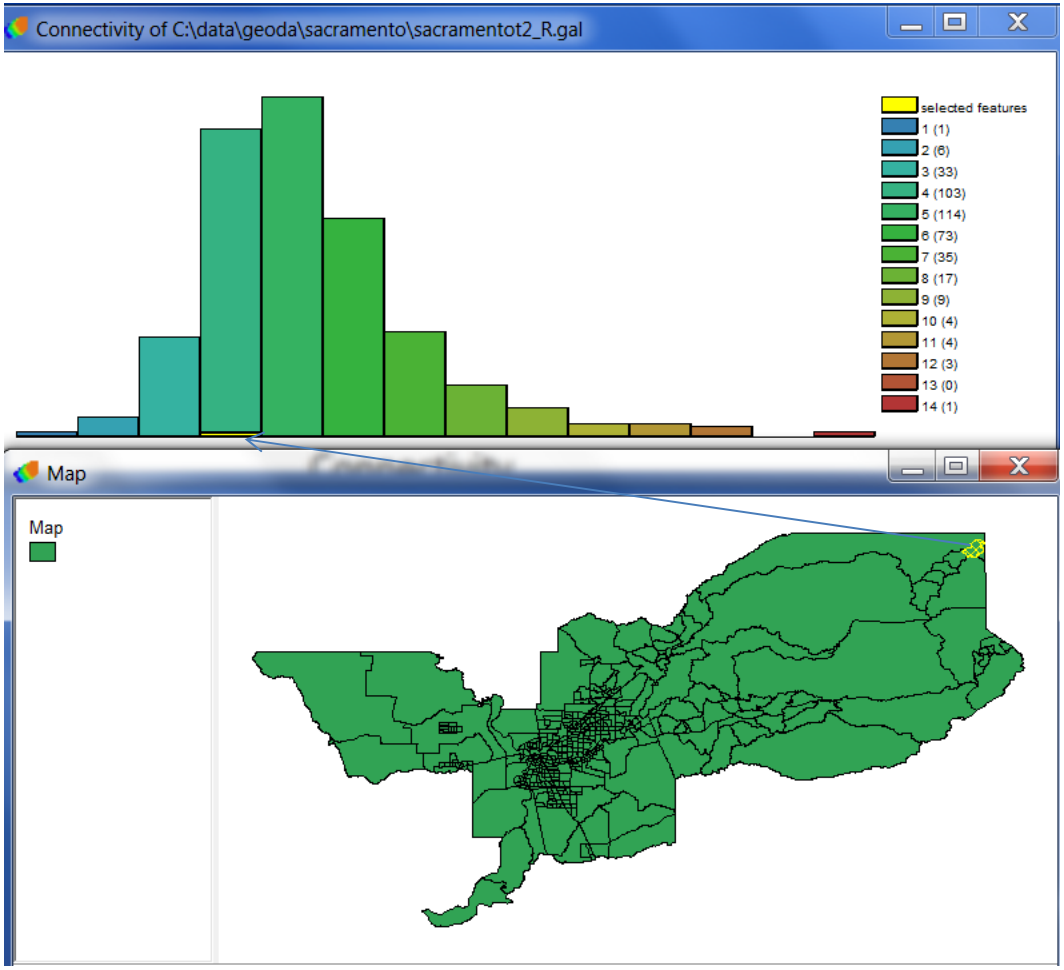
- Spacial weights can be based on contiguity from polygon boundary files or calculated from the distance between points.

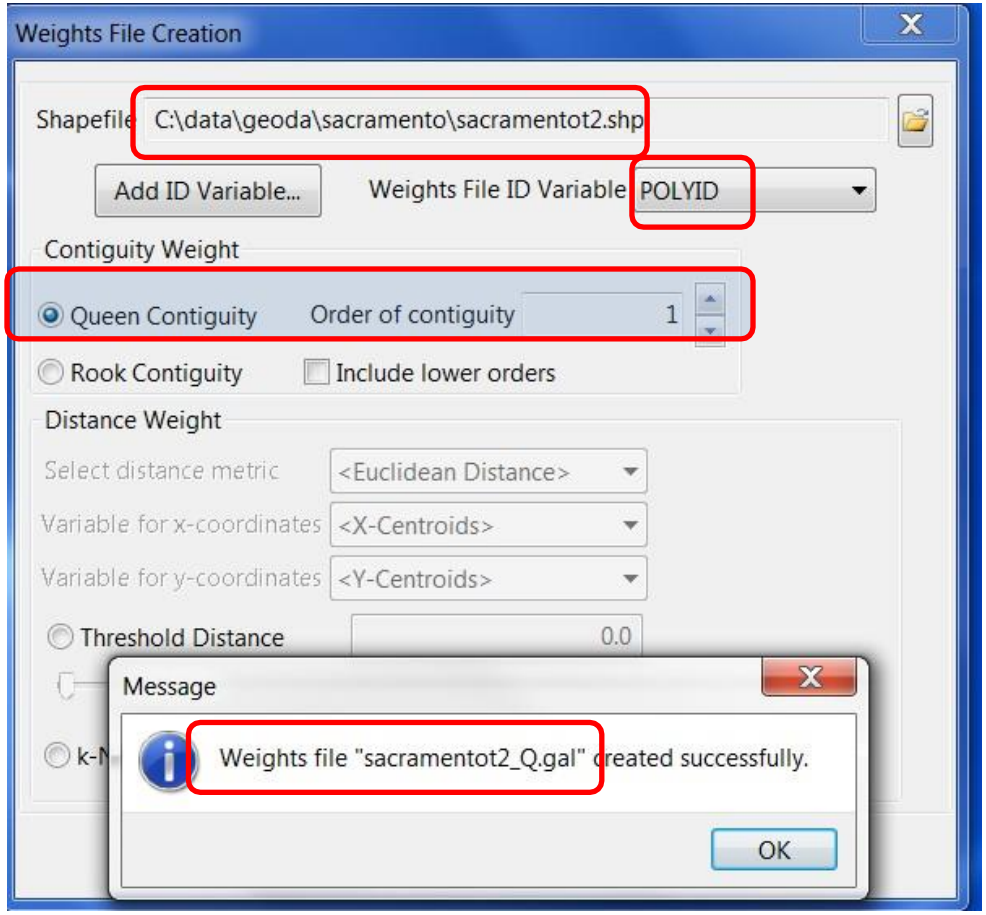

uses only common boundaries to define neighbors
polygon
shape files
CONTIGUITY
BASED
WEIGHTS
rook contiguity
queen contiguity
.GAL-file
1st order
higher order
uses all common points (denser connectedness structure)
removes redundancies and
circularities in the weights
construction

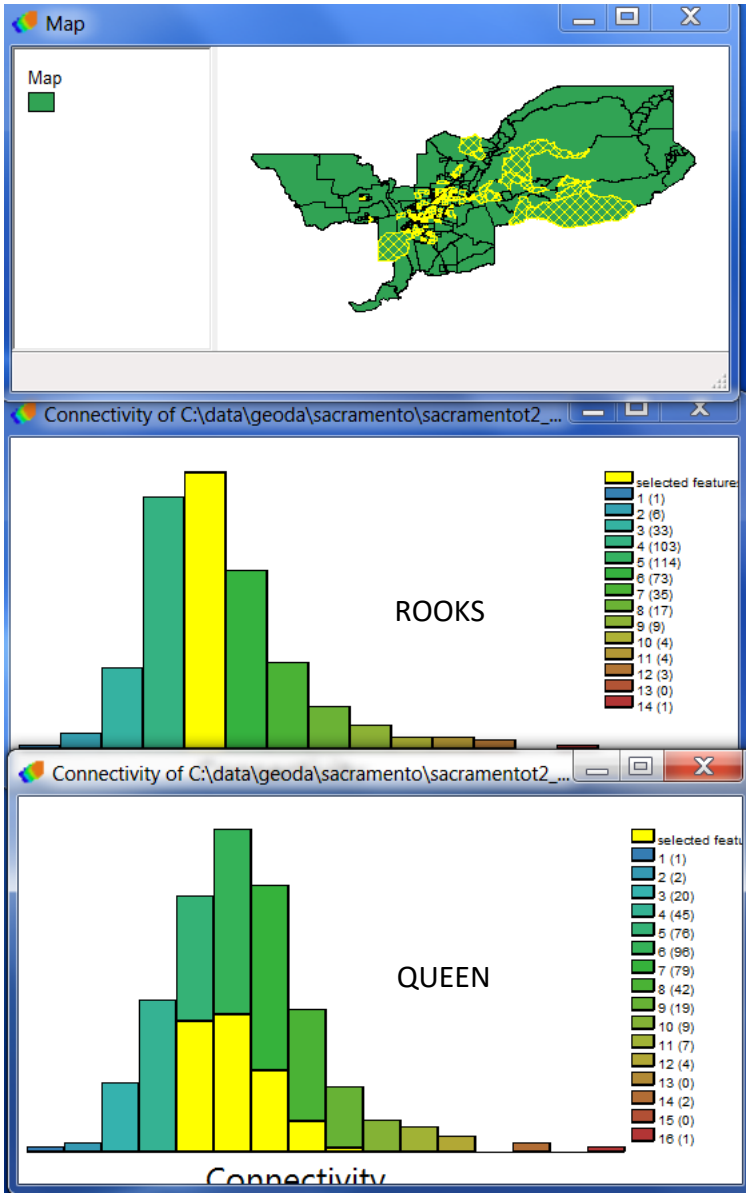


flag, number of observations, name of polygon shape file, name of the key variable



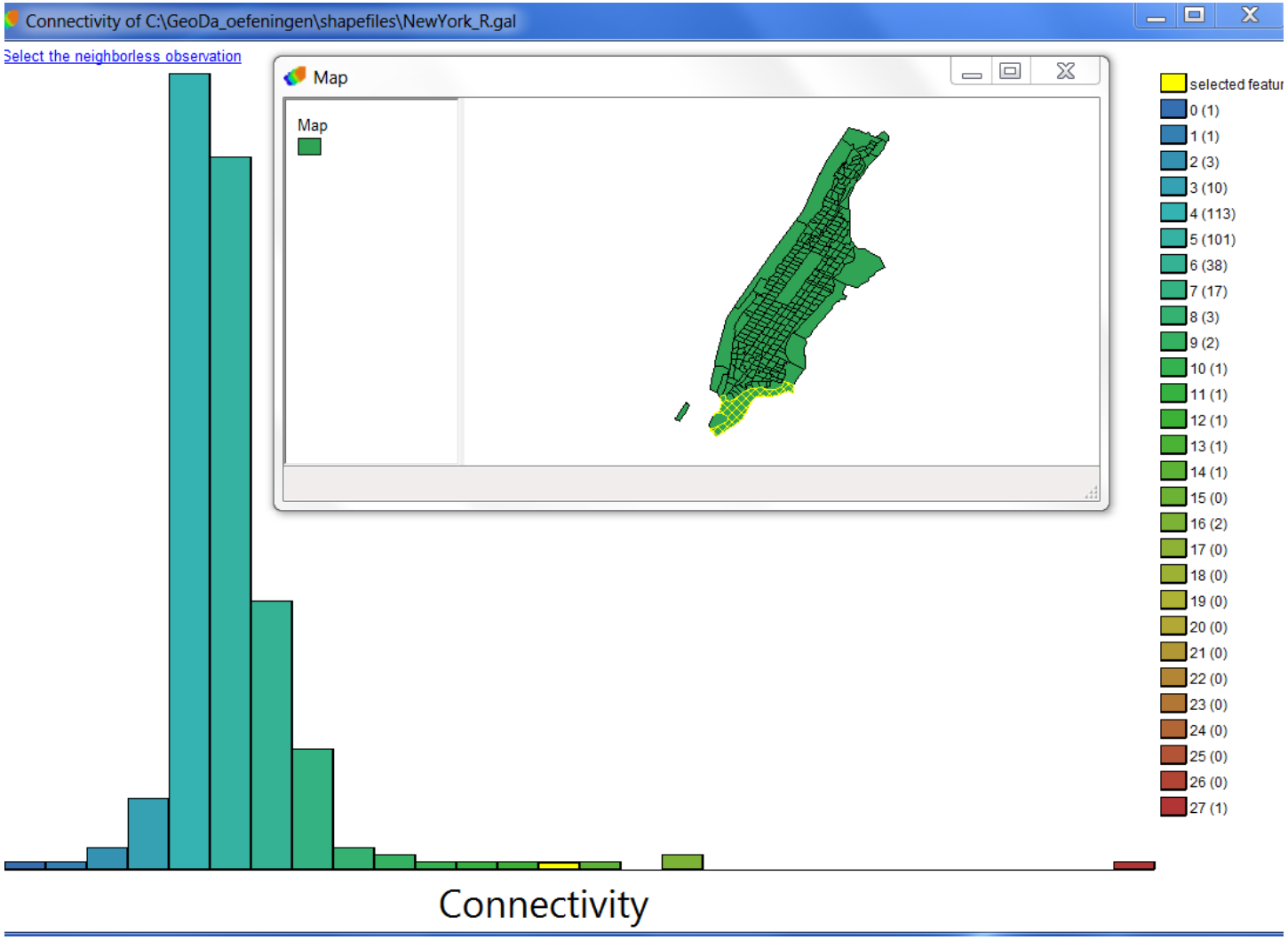

Connectivity of C:\data\geoda\sacramento\sacramentot2_R.gal
selected features
1 (1)
2 (6)
3 (33)
4 (103)
5 (114)
6 (73)
7 (35)
8 (17)
9 (9)
10 (4)
11 (4)
12 (3)
13 (0)
14 (1)
Map
Map


Weights File Creation
Shapefile
C:\data\geoda\sacramento\sacramentot2.shp
Add ID Variable...
Weights File ID Variable
POLYID
Contiguity Weight
Queen Contiguity
Order of contiguity
1
Rook Contiguity
Include lower orders
Distance Weight
Select distance metric
<Euclidean Distance>
Variable for x-coordinates
<X-Centroids>
Variable for y-coordinates
<Y-Centroids>
Threshold Distance
0.0
Message
Weights file "sacramentot2_Q.gal" created successfully.
OK

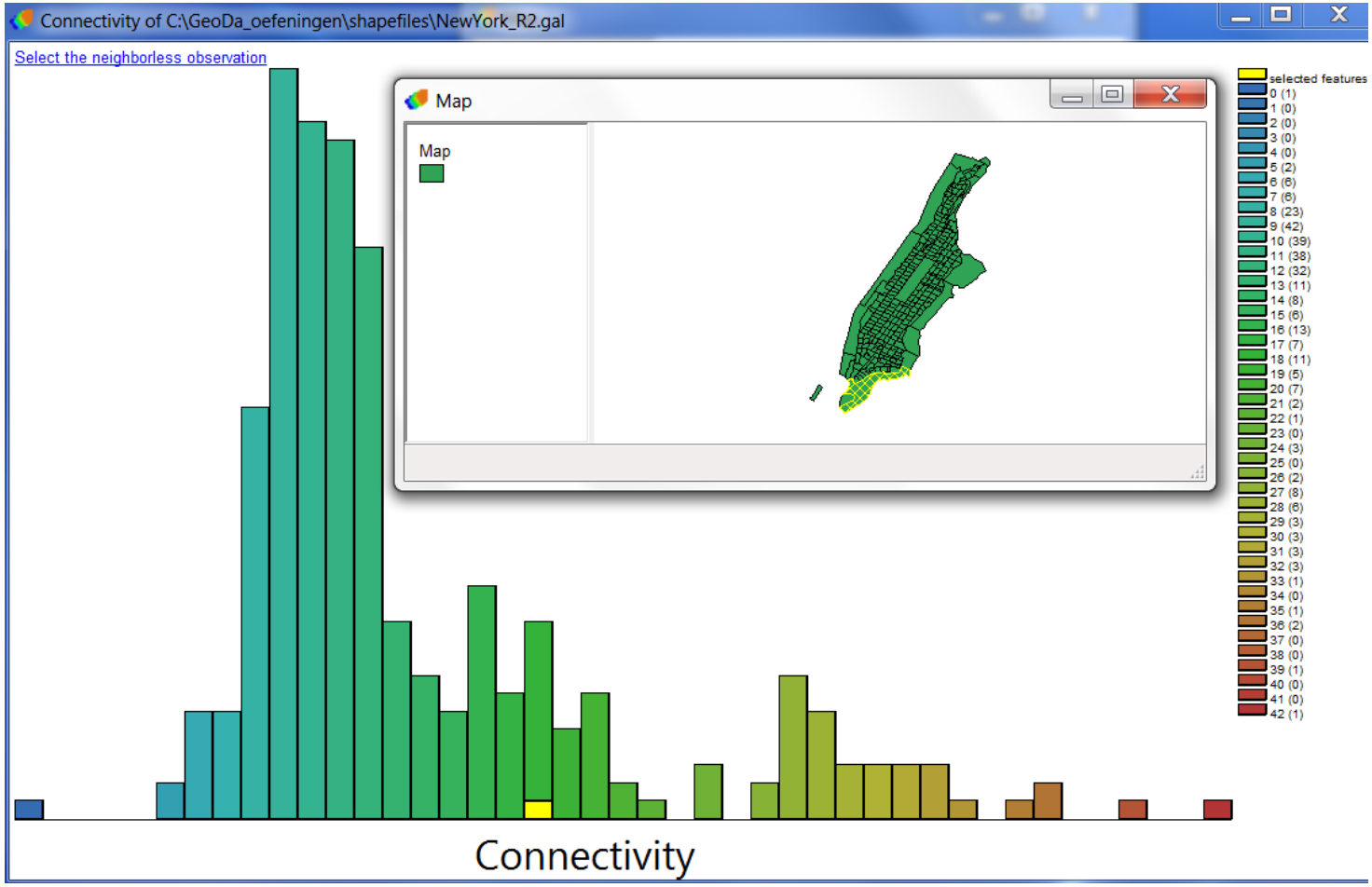


Comparison of connectedness structure for rook and queen contiguity

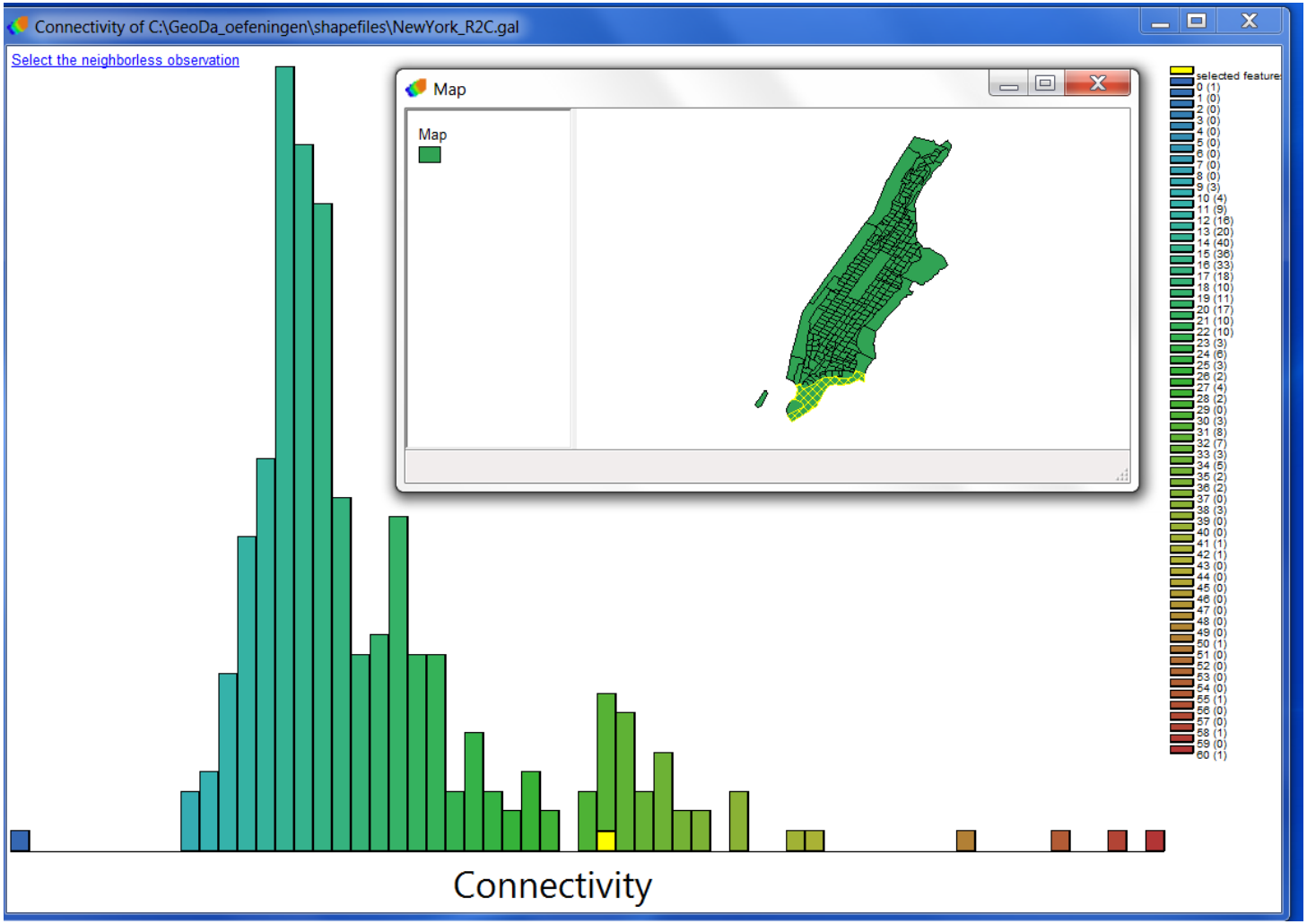
Rooks Contiguity

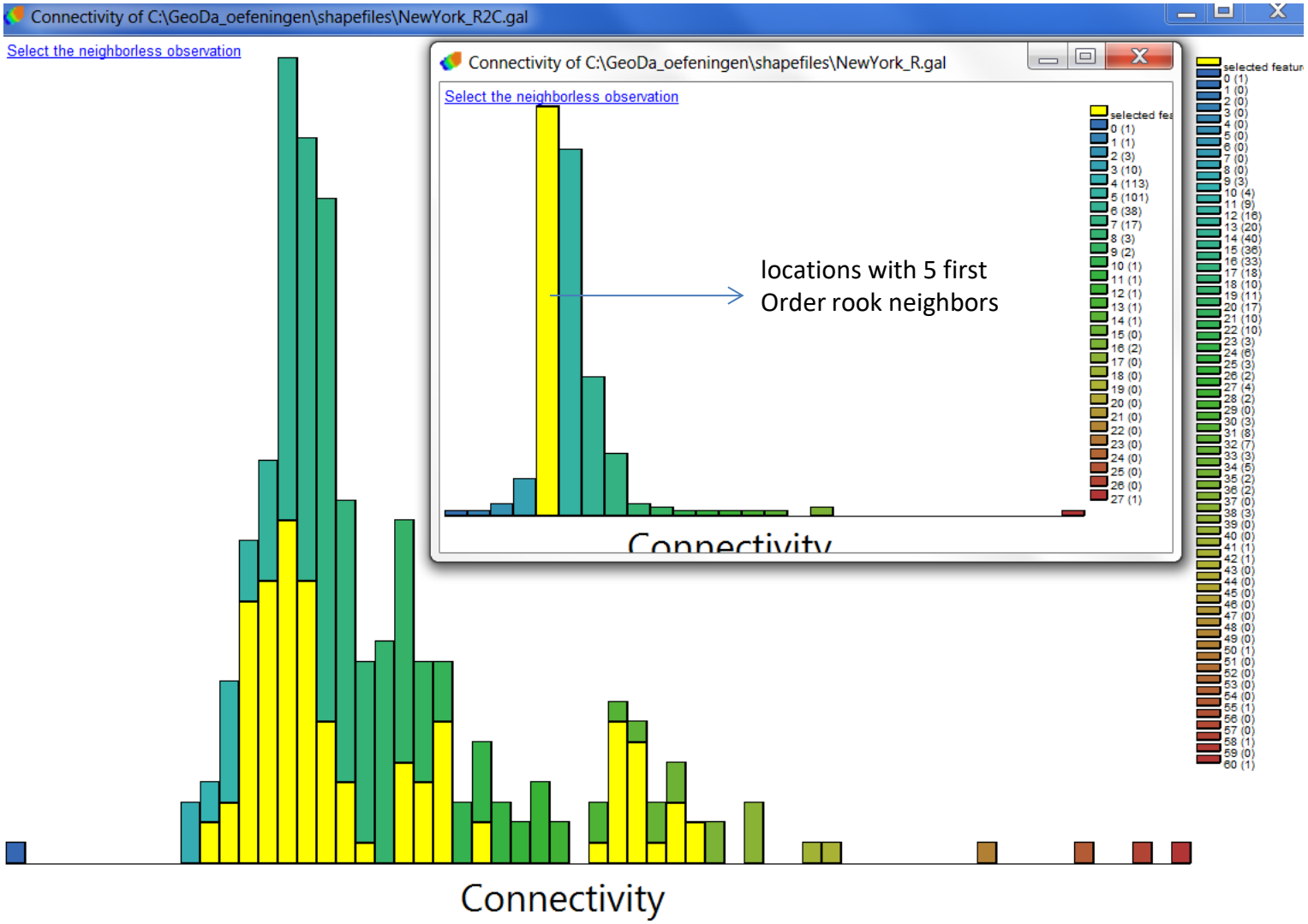


Pure 2nd order Rooks Contiguity

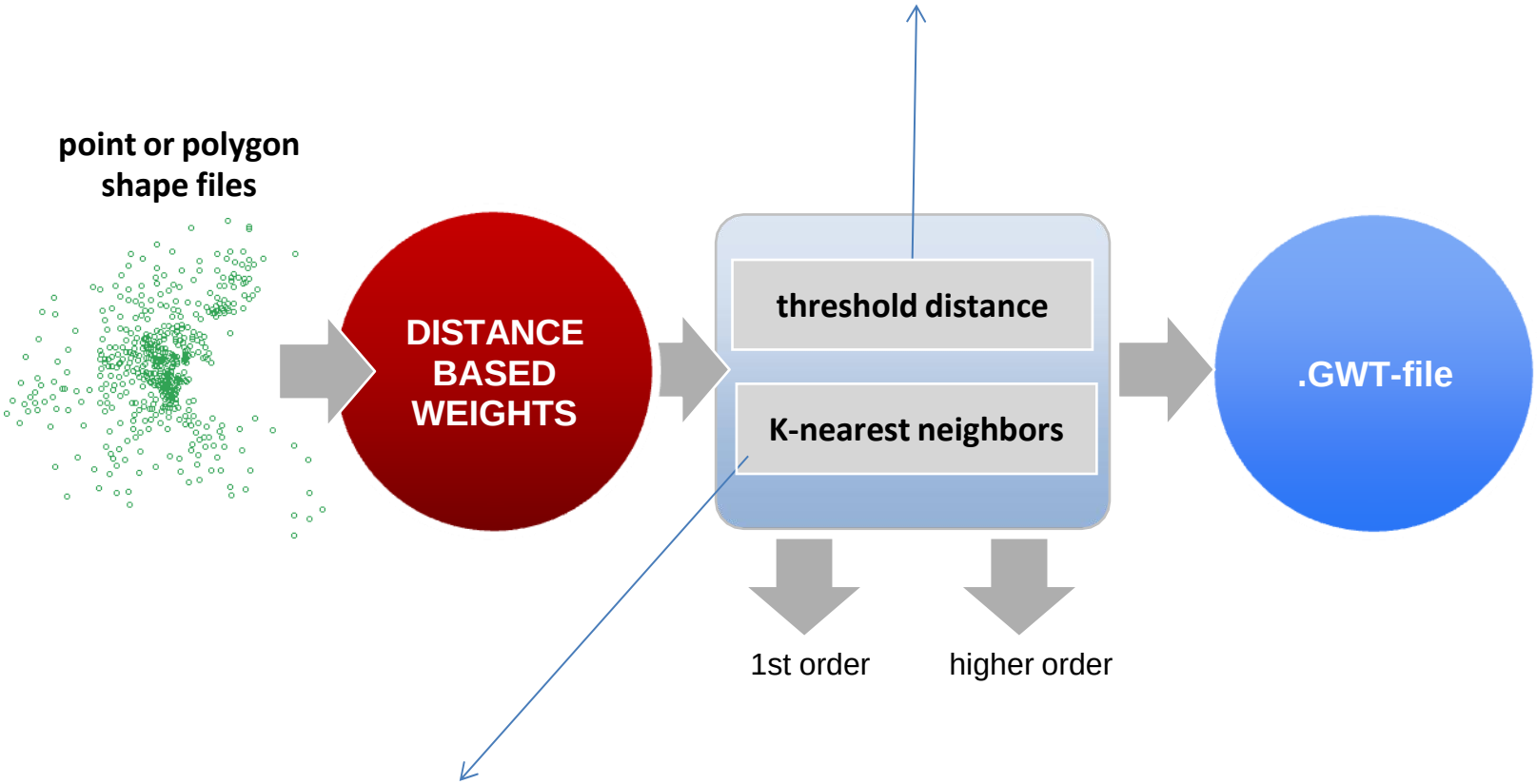


Cumulative 2nd order Rooks Contiguity



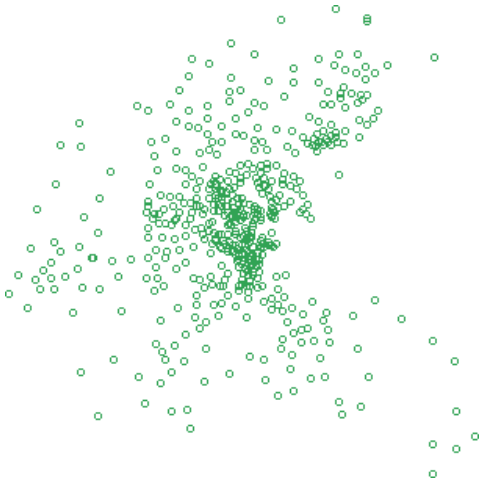

Connectivity of C:\GeoDa_oefeningen\shapefiles\NewYork_R2C.gal
Select the neighborless observation
Connectivity of C:\GeoDa_oefeningen\shapefiles\NewYork_R.gal
Select the neighborless observation
locations with 5 first
Order rook neighbors
Connectivity
Connectivity

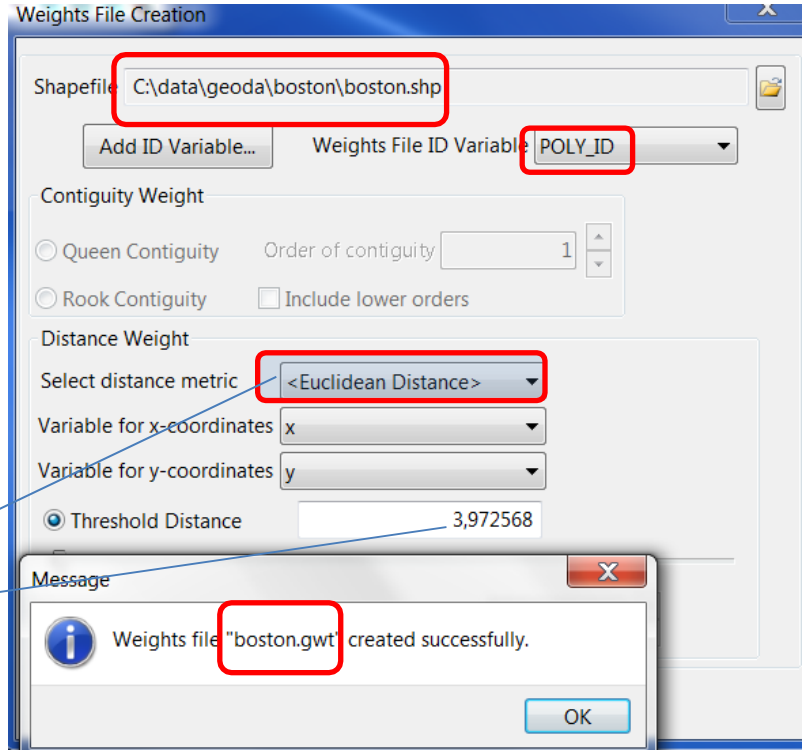
GeoDa calculates the minimum distance required to assure that each observation has at least one neighbor



Spacial weights based on distance threshold can lead to a very unbalenced connectedness structure (esp. In the case when spacial units have very different areas, with small areas having many neighbors while larger ones may have only a few). A commonly used alternative consists of considering the k-nearest neighbors.

In contrast to contiguity weights, distance-based spatial weights can be calculated for both point shape files as well as polygon shape files. For polygon files, if no coordinate variables are specified, the polygon centroids will be used as the basis for distance calculation. When polygon shape files are used, maps must be projected (e.g. UTM) for proper computation of centroids. For unprojected maps, the resulting centroids will only approximate.





the minimum distance required to ensure that each location has at least one neighbor

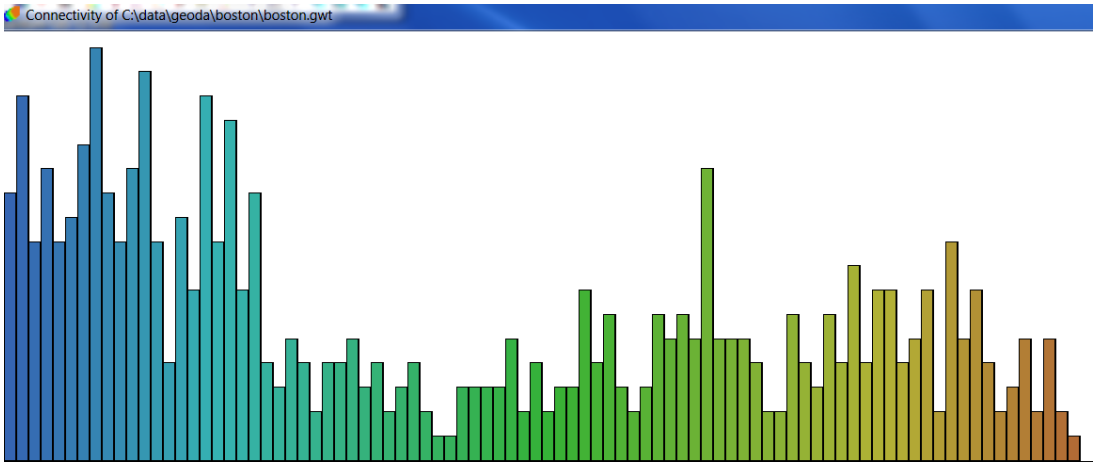
if the points are in latitude and longitude, select the <Arc Distance> option

boston.gwt - Kladblok

Bestand Bewerken Opmaak Beeld Help

```
0 506 boston POLY_ID
1 34          3.07546744
1 33          2.77108282
1 32          2.49721845
1 31          2.83677634
1 28          3.07668653
1 26          3.29875734
1 25          3.18001572
1 24          3.04179223
1 23          3.23945983
1 22          3.64012362
1 21          3.38292773
1 20          3.69871599
1 18          3.83005222
1 15          3.81559956
1 29          2.78179798
1 27          3.57261809
1 500          3.82973889
1 498          3.33241654
1 497          3.80474703
1 35          2.64594029
1 502           3.6840874
1 501          3.58615393
1 2           3.63455637
1 3           3.48058903
1 30          2.56976653
2 45           3.5730799
2 50          3.70194543
2 49          3.23790055
2 48          2.81305883
2 30          1.06569226
2 29          1.01607086
2 27         0.800999376
2 13          2.05039021
2 47          2.53647393
2 46          2.89546197
2 34          2.75036361
```

distance between neighbor pairs



Connectivity for distance-based weights

The distribution has a much broader range compared to contiguity-based weights. Some points are clustered while other are far apart. The minimum threshold needed to avoid islands may be too large for many or most locations in the data set. In such cases, care is needed in the specification of the distance threshold, and the use of K-nearest weights may be more appropriate.

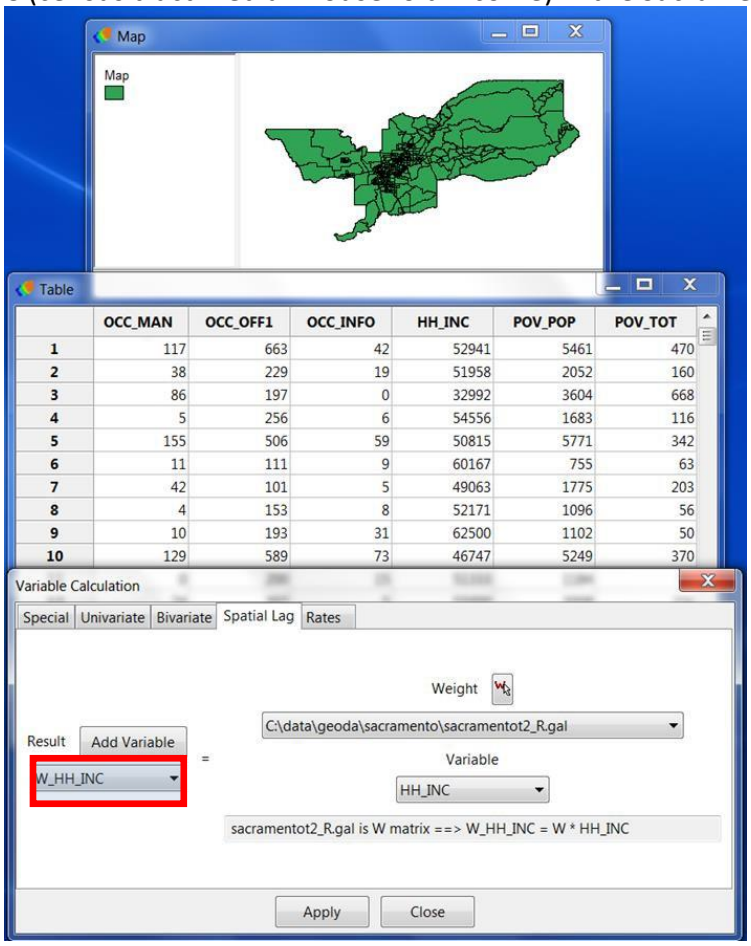
Spatially lagged variables are an essential part of the computation of spatial autocorrelation tests and the specification of spatial regression models. GeoDa computes these variables on the fly, but in some instances it is useful to calculate spatially lagged variables explicitly.

We will calculate a spatially lagged variable for the variable HH_INC (census tract median household income) in the Sacramento file.

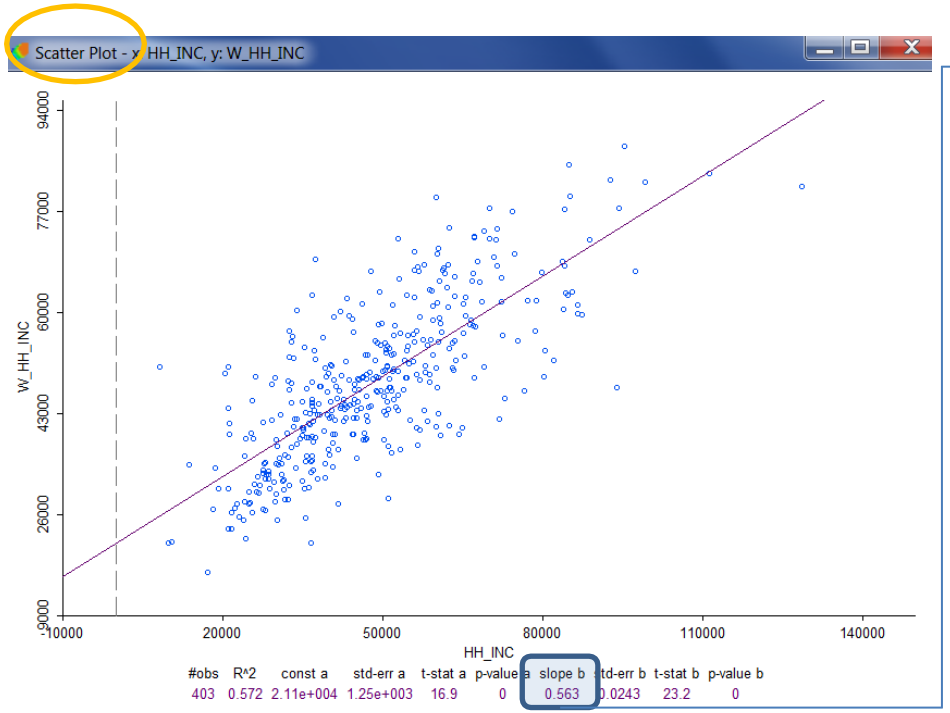
The first thing we do is open the spatial weights file we created. Then we create a new field that is added to the table.

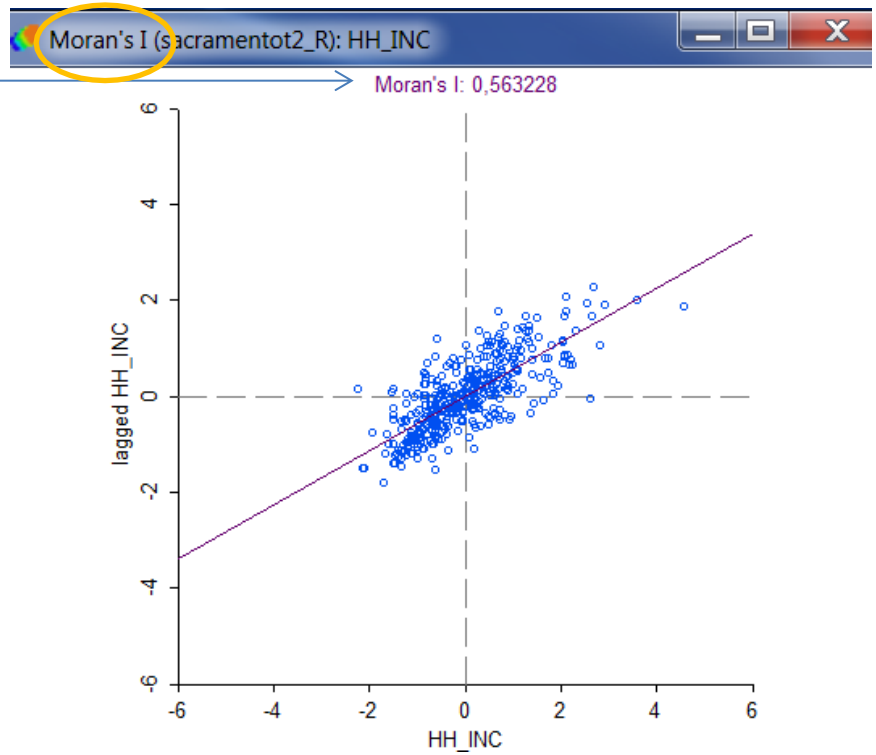
Table

	OCC_INFO	FIPSNO	HH_INC	W_HH_INC	POV_POP	POV_TOT
1	42	6061022001	52941	50100,2500000	5461	470
2	19	6061020106	51958	50164,0000000	2052	160
3	0	6061020107	32992	55022,0000000	3604	668
4	6	6061020105	54556	53532,2500000	1683	116
5	59	6061020200	50815	53165,5000000	5771	342
6	9	6061020101	60167	50673,5000000	755	63
7	5	6061020104	49063	54958,7500000	1775	203
8	8	6061020103	52171	56167,7500000	1096	56
9	31	6061020102	62500	54023,5000000	1102	50
10	73	6061022002	46747	52918,6666667	5249	370
11	15	6017030503	51333	47301,7500000	1184	90
12	5	6061021901	55000	51433,5000000	3008	206
13	57	6061021304	52286	57758,6666667	4692	328
14	84	6061021902	59443	61150,0000000	4286	130
15	30	6061021600	38854	55771,8000000	6866	673
16	31	6017030603	41982	46819,7142857	2759	229
17	44	6061021801	67300	49008,6666667	4026	147



The value of the spatially lagged variable "W_HH_INC" for this location is the mean of its neighbors

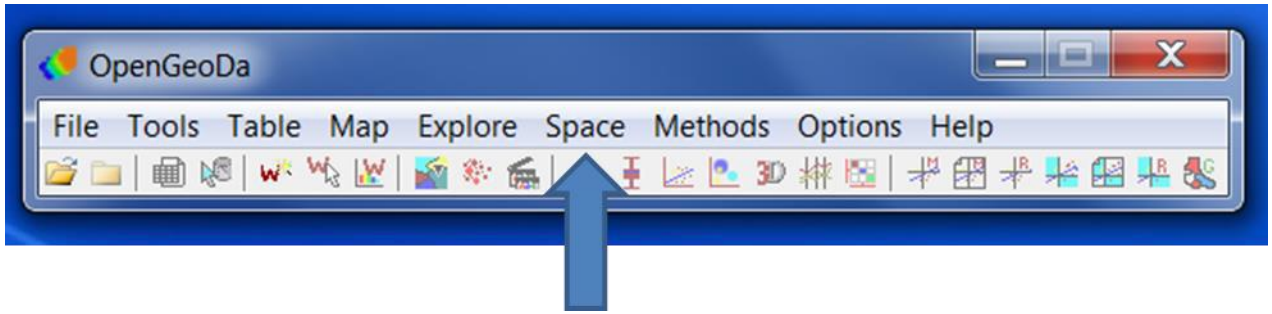

Scatter Plot - x: HH_INC, y: W_HH_INC
W_HH_INC
HH_INC
9000
26000
43000
60000
77000
94000
10000
20000
50000
80000
110000
140000
#obs R^2 const a std-err a t-stat a p-value a slope b std-err b t-stat b p-value b
403 0.572 2.11e+004 1.25e+003 16.9 0 0.563 0.0243 23.2 0

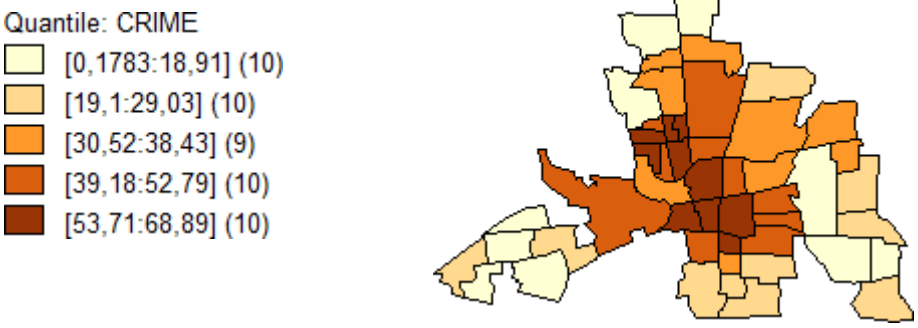

Moran's I (sacramentot2_R): HH_INC
Moran's I: 0,563228
lagged HH_INC
HH_INC
-6
-4
-2
0
2
4
6

- **Global spacial autocorrelation** is handled in GeoDa by means of **Moran's I spatial autocorrelation statistic** and its visualization in the form of a scatterplot.

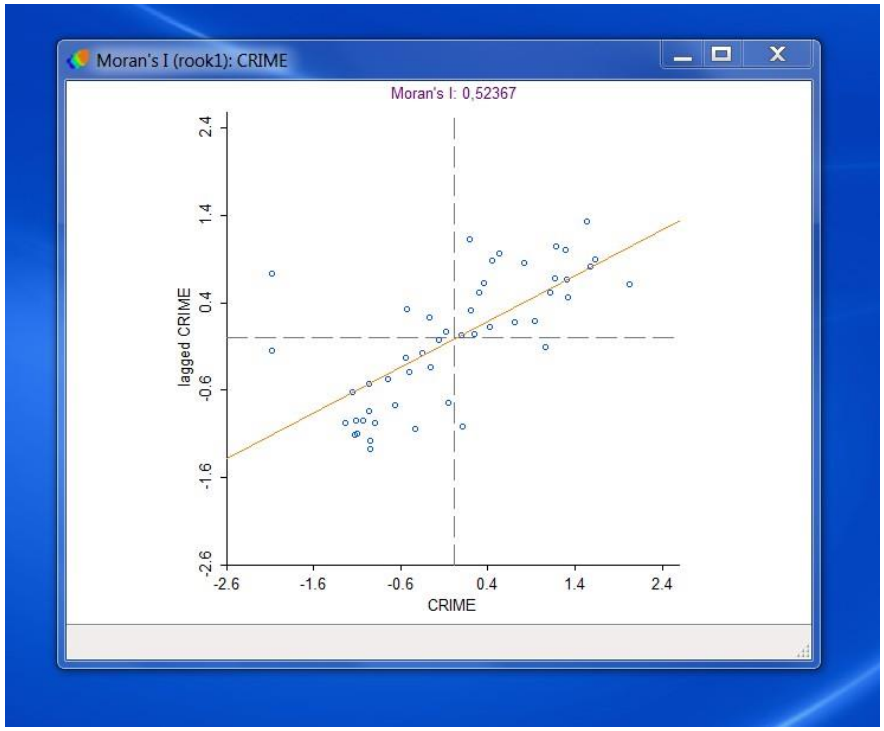
- Global spacial autocorrelation requires a spatial weights file and a variable must be specified.

- Spacial autocorrelation analysis is implemented in its traditional **univariate form** as well in a **bivariate form**.



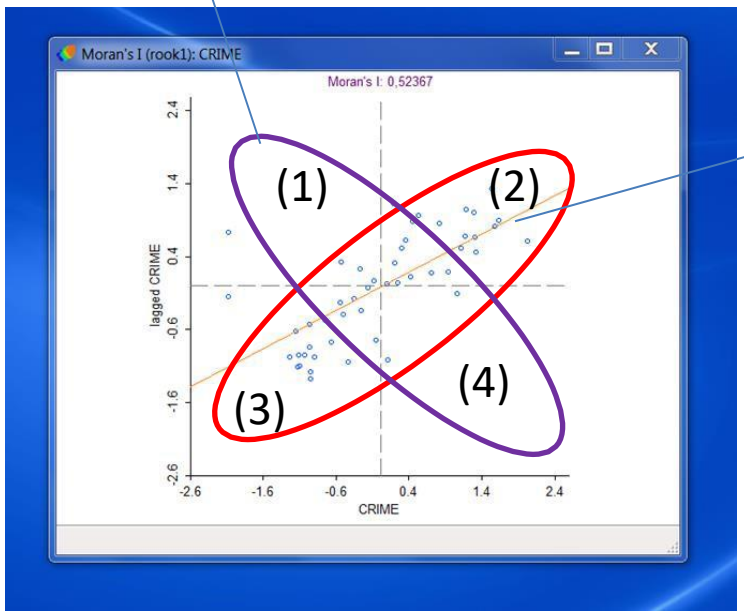


Moran's I for Columbus data (variable = crime ; spacial weights file = rooks-based contiguity file)

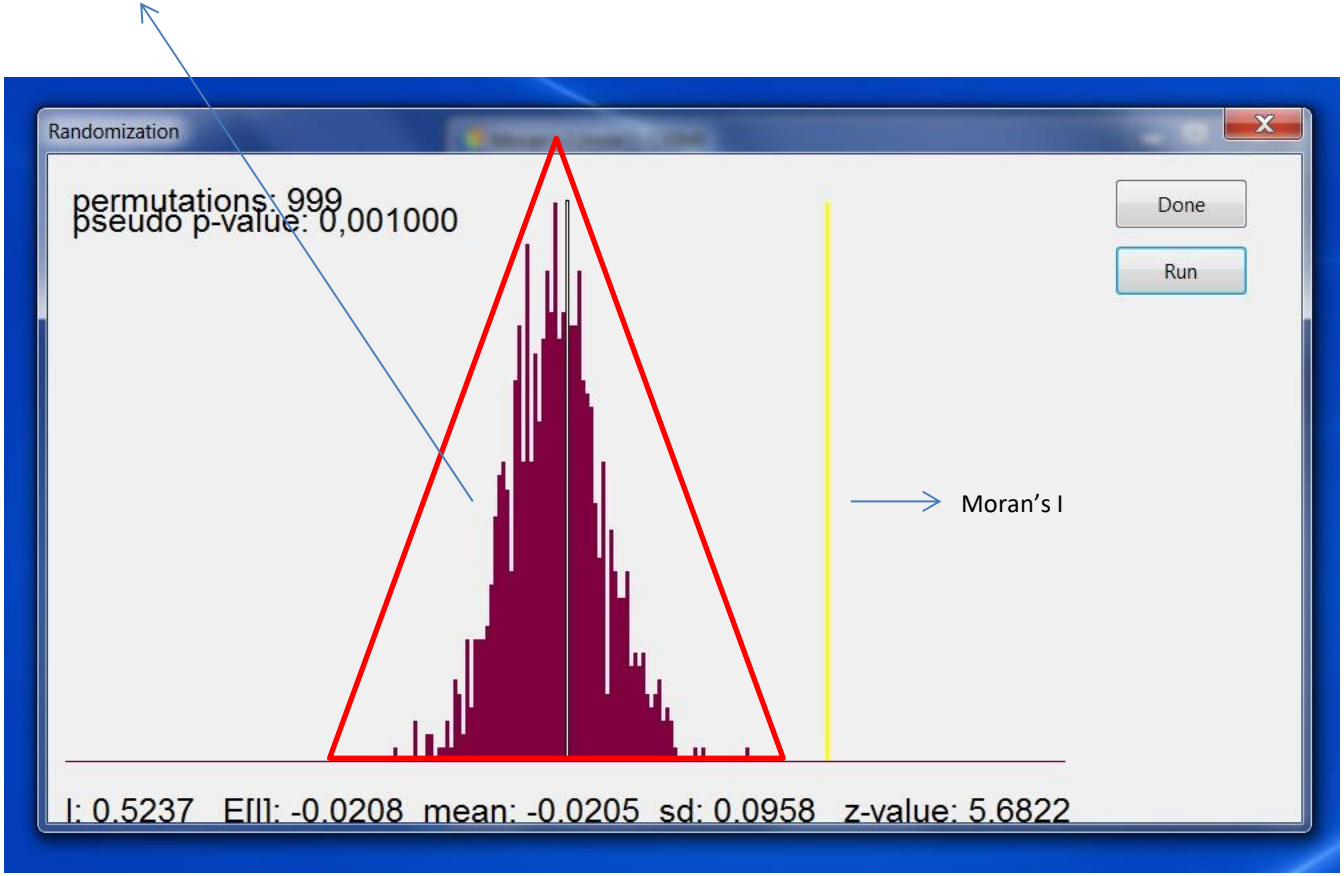


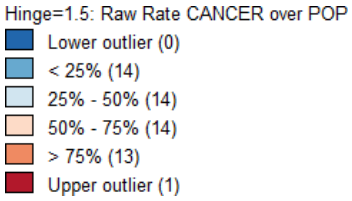
negative autocorrelation

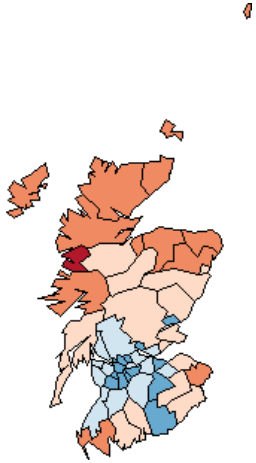
positive autocorrelation

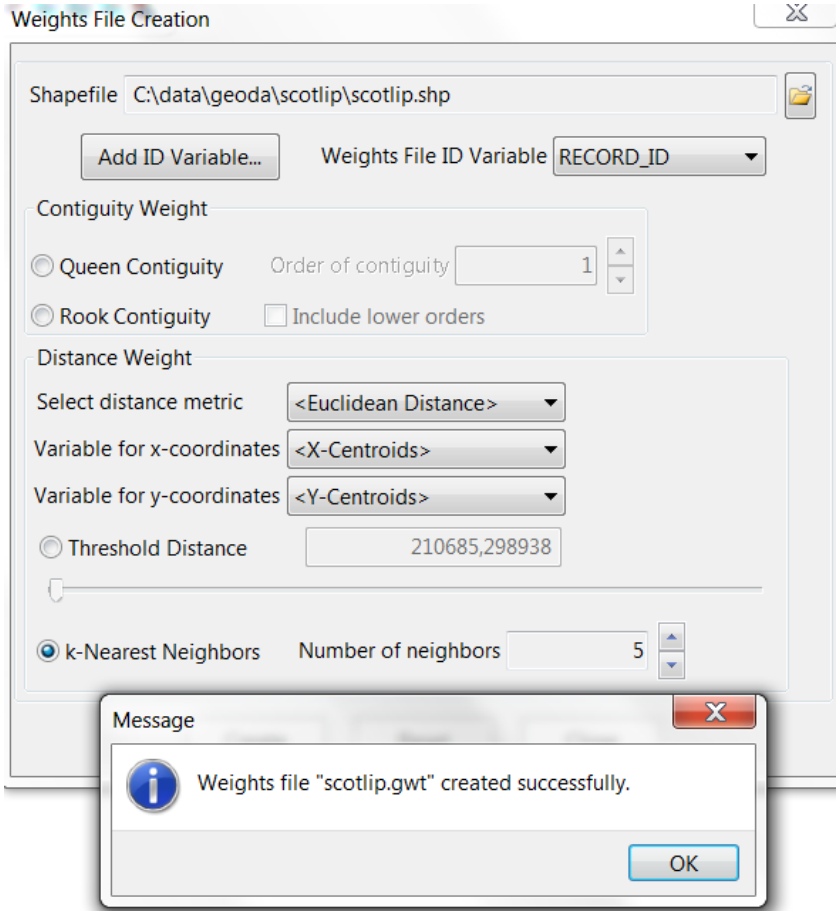


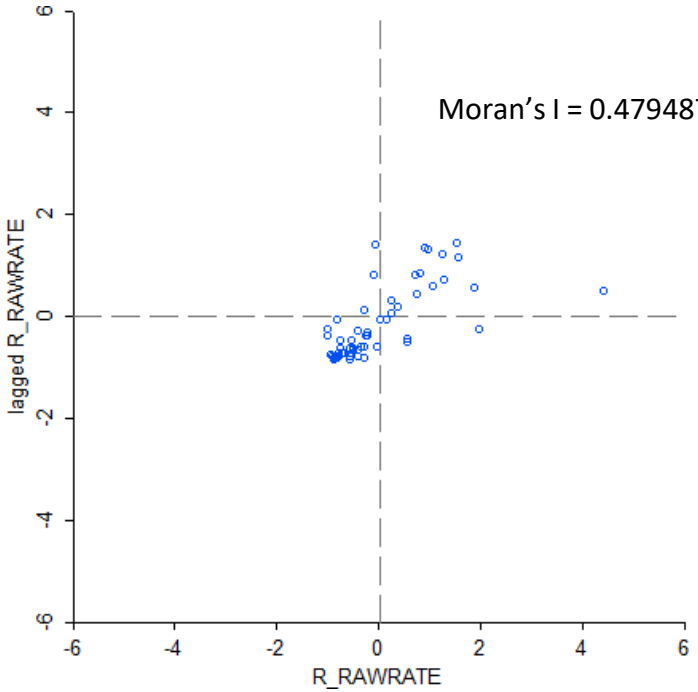
reference distribution calculated for spatially random layouts with the same data as observed
(none of the simulated values is larger than the observed 0.52)

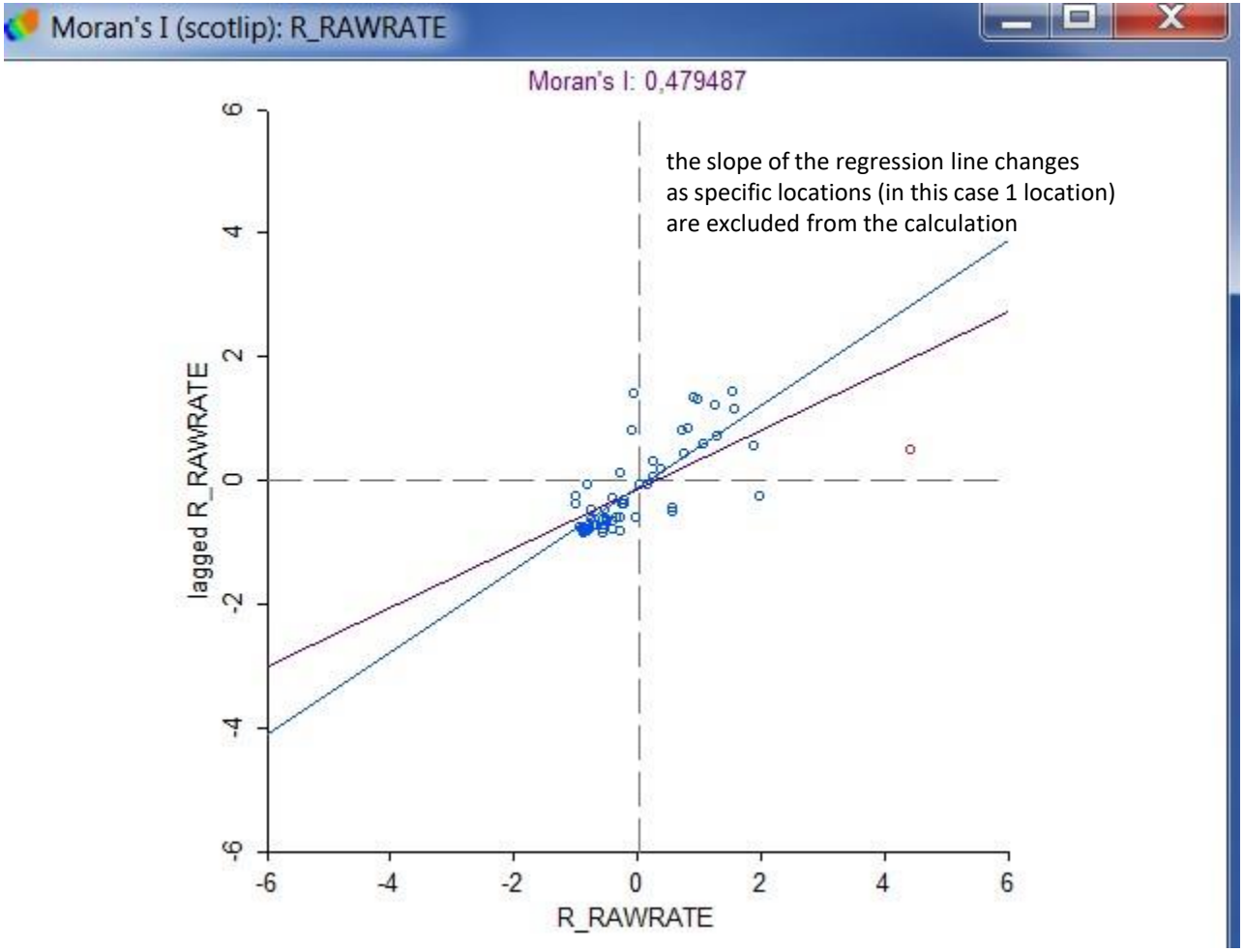



Hinge=1.5: Raw Rate CANCER over POP
Lower outlier (0)
< 25% (14)
25% - 50% (14)
50% - 75% (14)
> 75% (13)
Upper outlier (1)

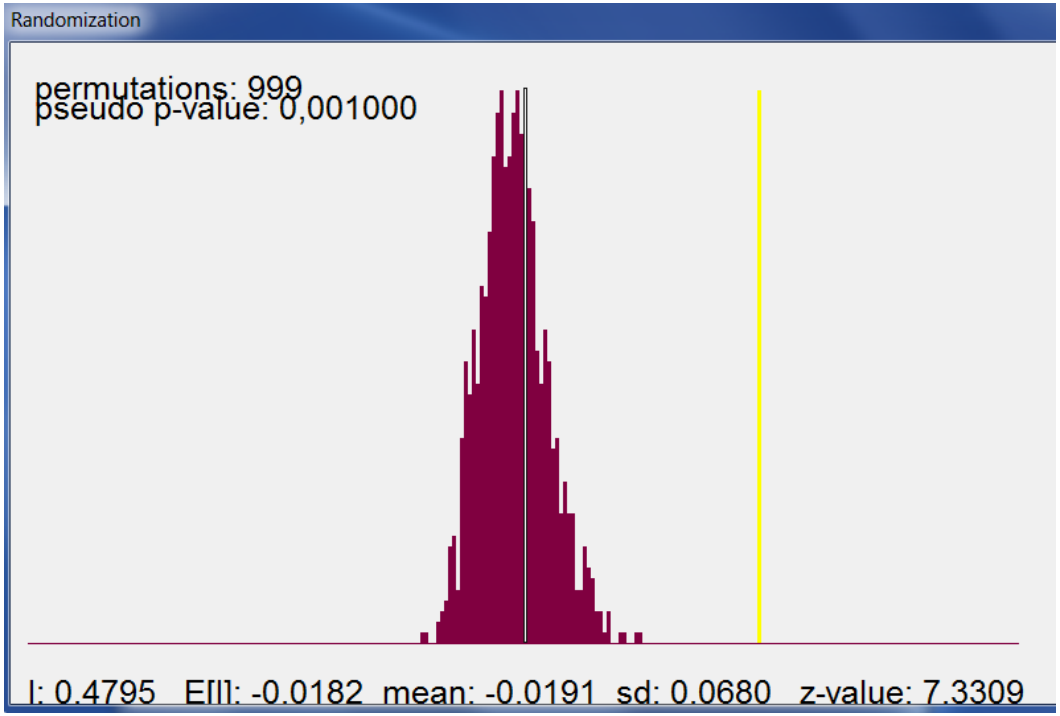



Weights File Creation
Shapefile C:\data\geoda\scotlip\scotlip.shp
Add ID Variable...
Weights File ID Variable RECORD_ID
Contiguity Weight
Queen Contiguity
Order of contiguity 1
Rook Contiguity
Include lower orders
Distance Weight
Select distance metric <Euclidean Distance>
Variable for x-coordinates <X-Centroids>
Variable for y-coordinates <Y-Centroids>
Threshold Distance 210685,298938
k-Nearest Neighbors
Number of neighbors 5
Message
Weights file "scotlip.gwt" created successfully.
OK


Moran's I = 0.479487
lagged R_RAWRATE
R_RAWRATE
6
4
2
0
-2
-4
-6


Moran's I (scotlip): R_RAWRATE
Moran's I: 0,479487
the slope of the regression line changes
as specific locations (in this case 1 location)
are excluded from the calculation
lagged R_RAWRATE
R_RAWRATE
6
4
2
0
-2
-4
-6
-6
-4
-2
0
2
4
6

Inference for Moran's I is based on a random permutation procedure, which recalculates the statistic many times to generate a reference distribution. The obtained statistic is then compared to this reference distribution and a pseudo significance level is computed.



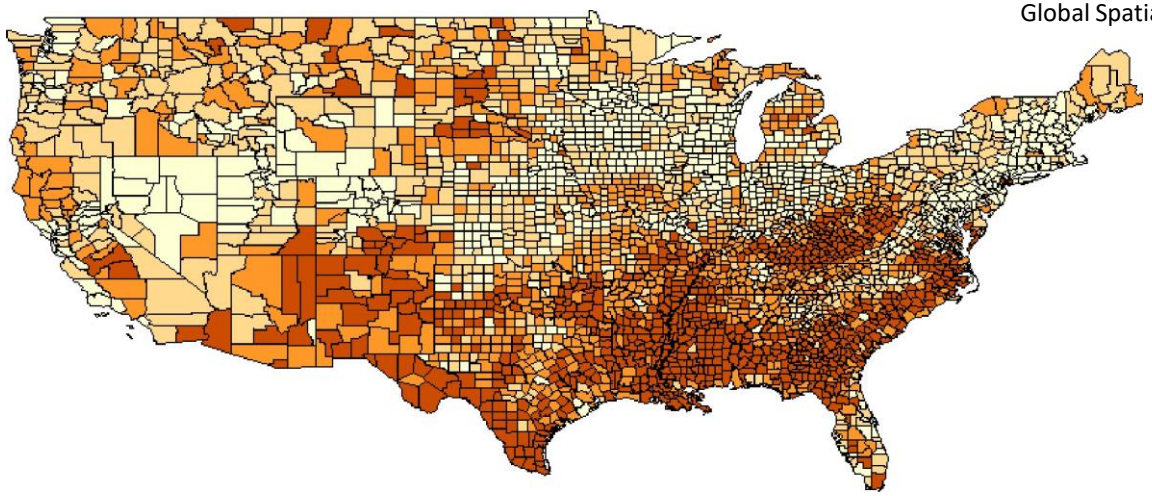
Ouantiil e: R090

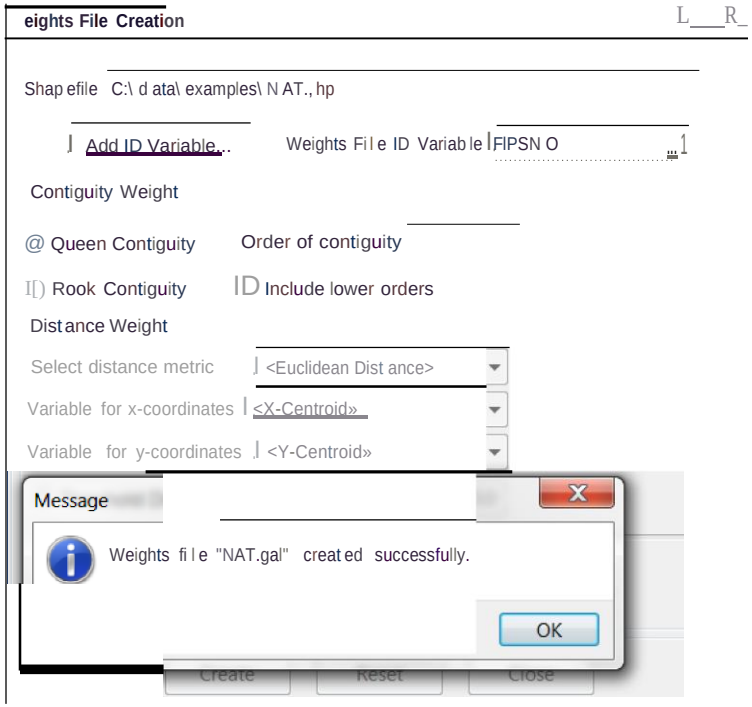
0 [-2,41:-0,i6668] ,(771)

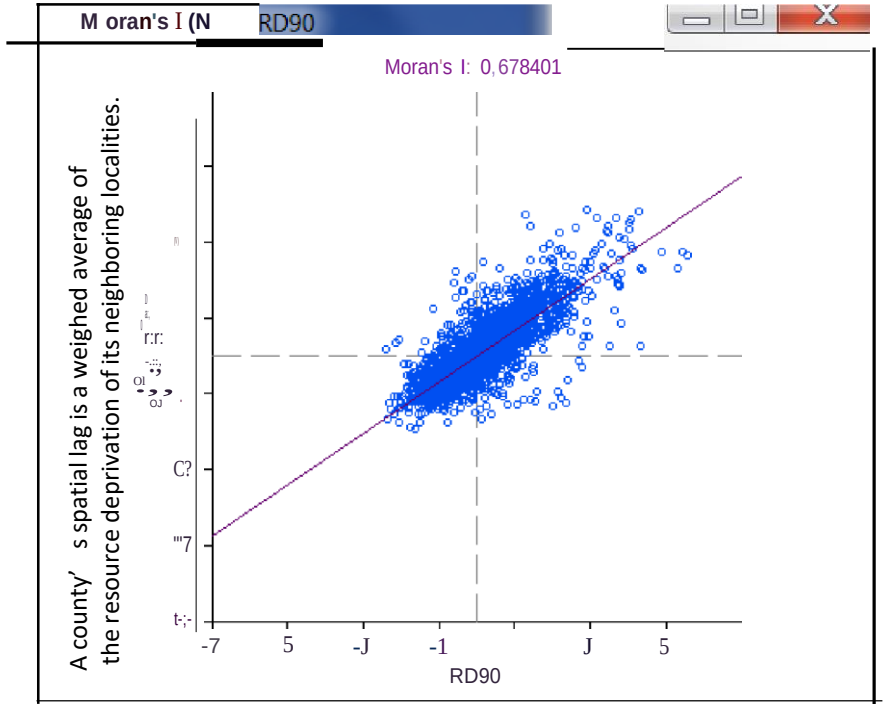
0 [-0,i6667:-0,2:01i6] ,(772)

0 [-0,2:()07:0,4393] ,(771)

• [0,4411:5,583] ,(771)







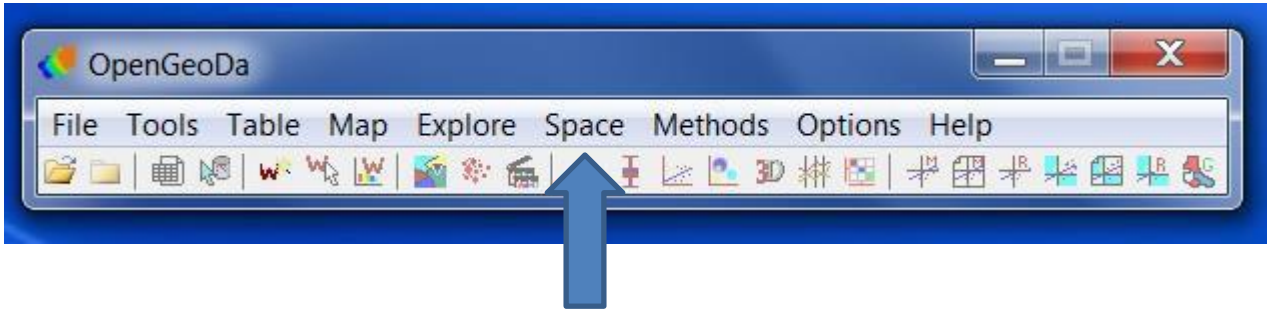
- Global measures : **global spatial autocorrelation** (Moran's I) : a single value which applies to the entire data set (the same pattern or process occurs over the entire geographical area ; and average for the entire area).

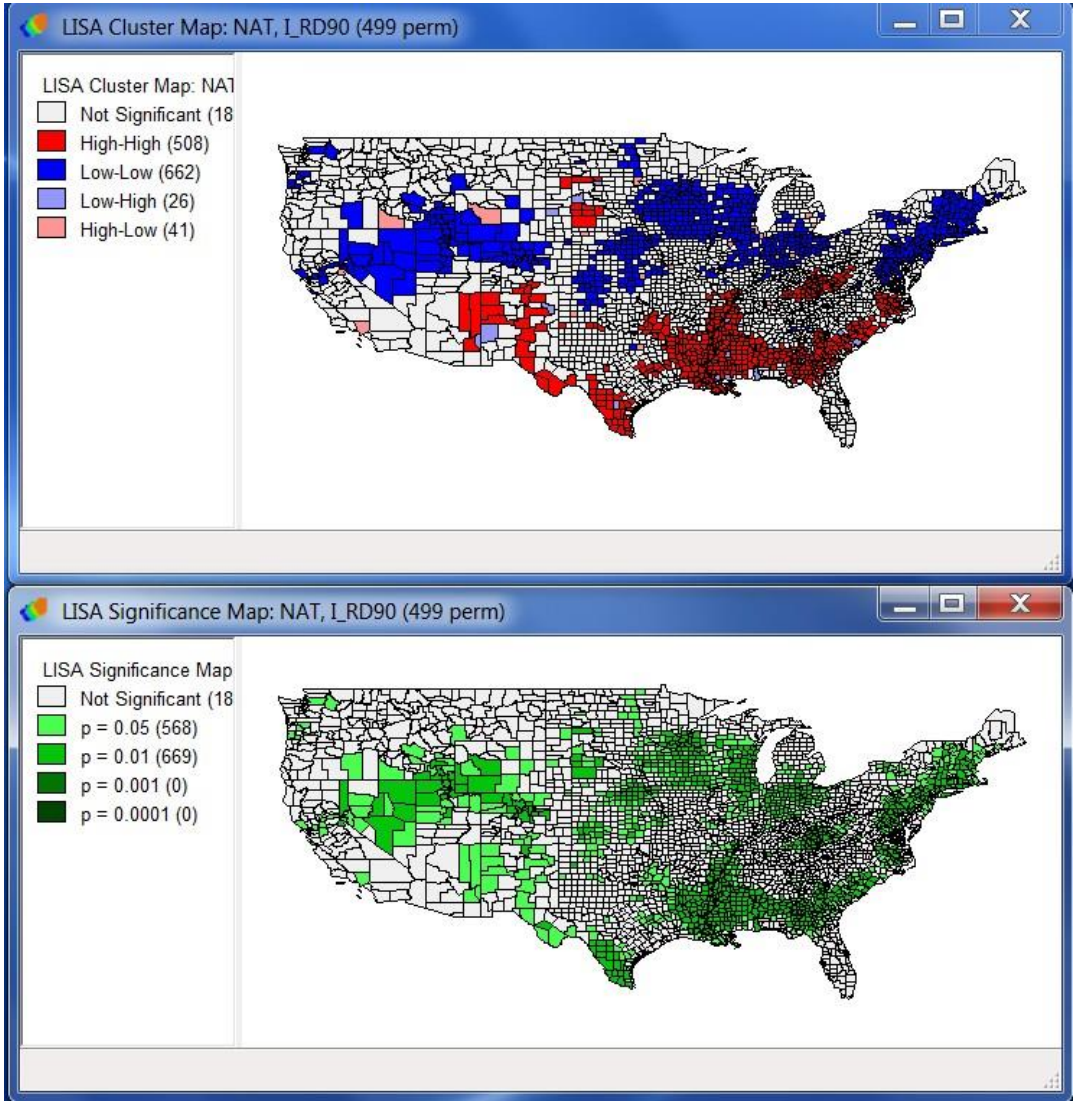
- Local measures : **local spatial autocorrelation** (Lisa) : a value calculated for each observation unit (different patterns of processes may occur in different parts of the region ; a unique number for each location).

- **Local spatial autocorrelation** is based on **local Moran LISA statistics**. This yields a measure of spatial autocorrelation for each individual location.

- Both univariate and multivariate LISA are included in GeoDa.

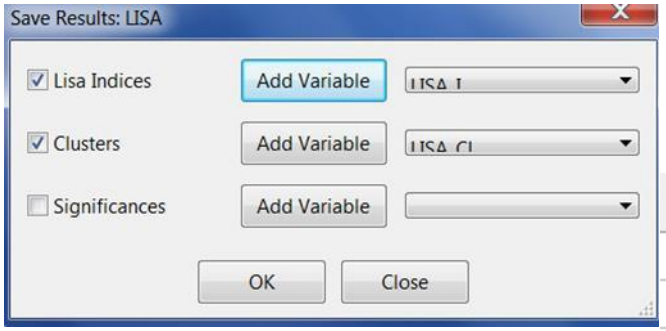
- The input needed for local spatial autocorrelation is the same as for global spatial autocorrelation.





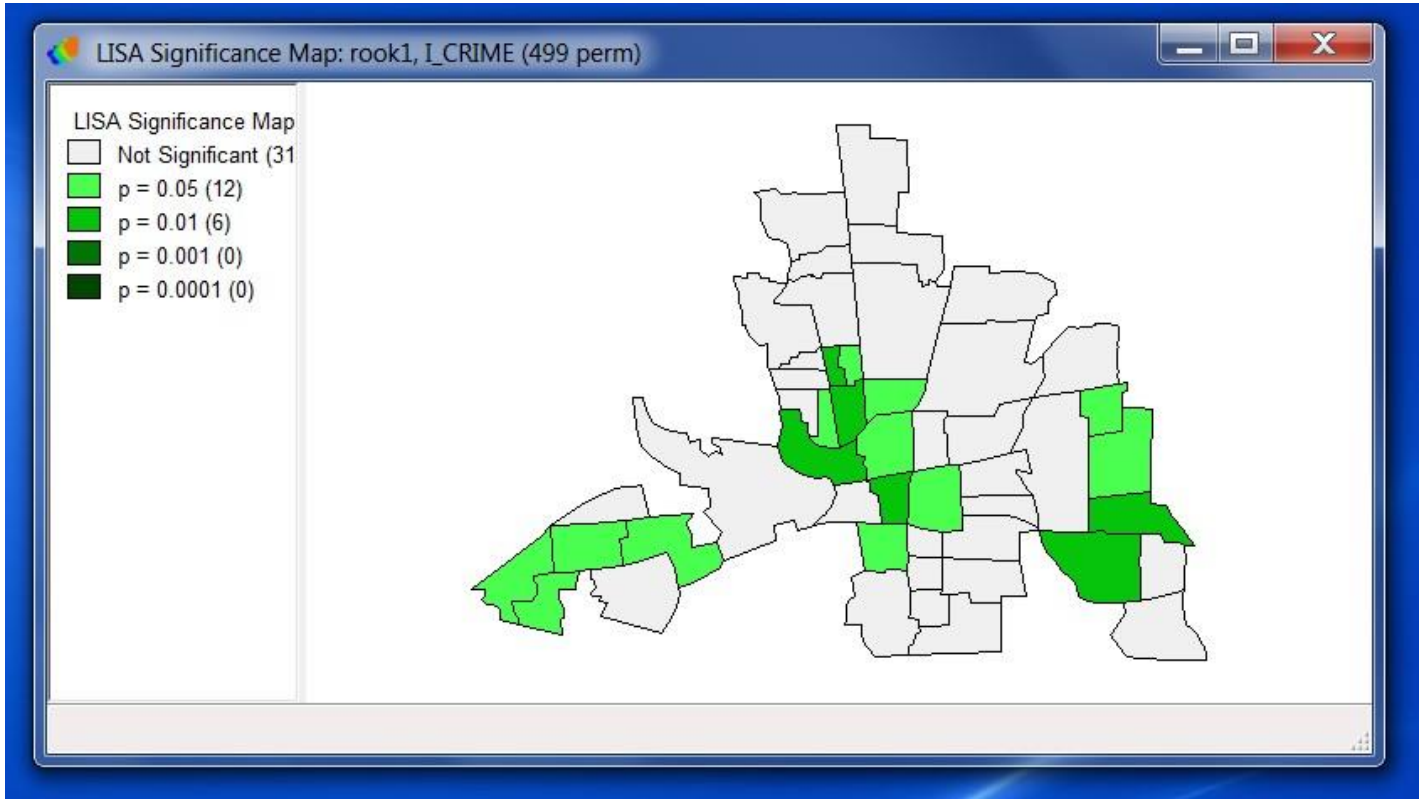
the high-high and low-low locations (positive local spatial autocorrelation) are typically referred to as **spatial clusters**, while the low-high and high-low are termed **spatial outliers** (while outliers are single locations by definition, this is not the case for clusters)

the significance map shows the locations with significant local Moran statistics



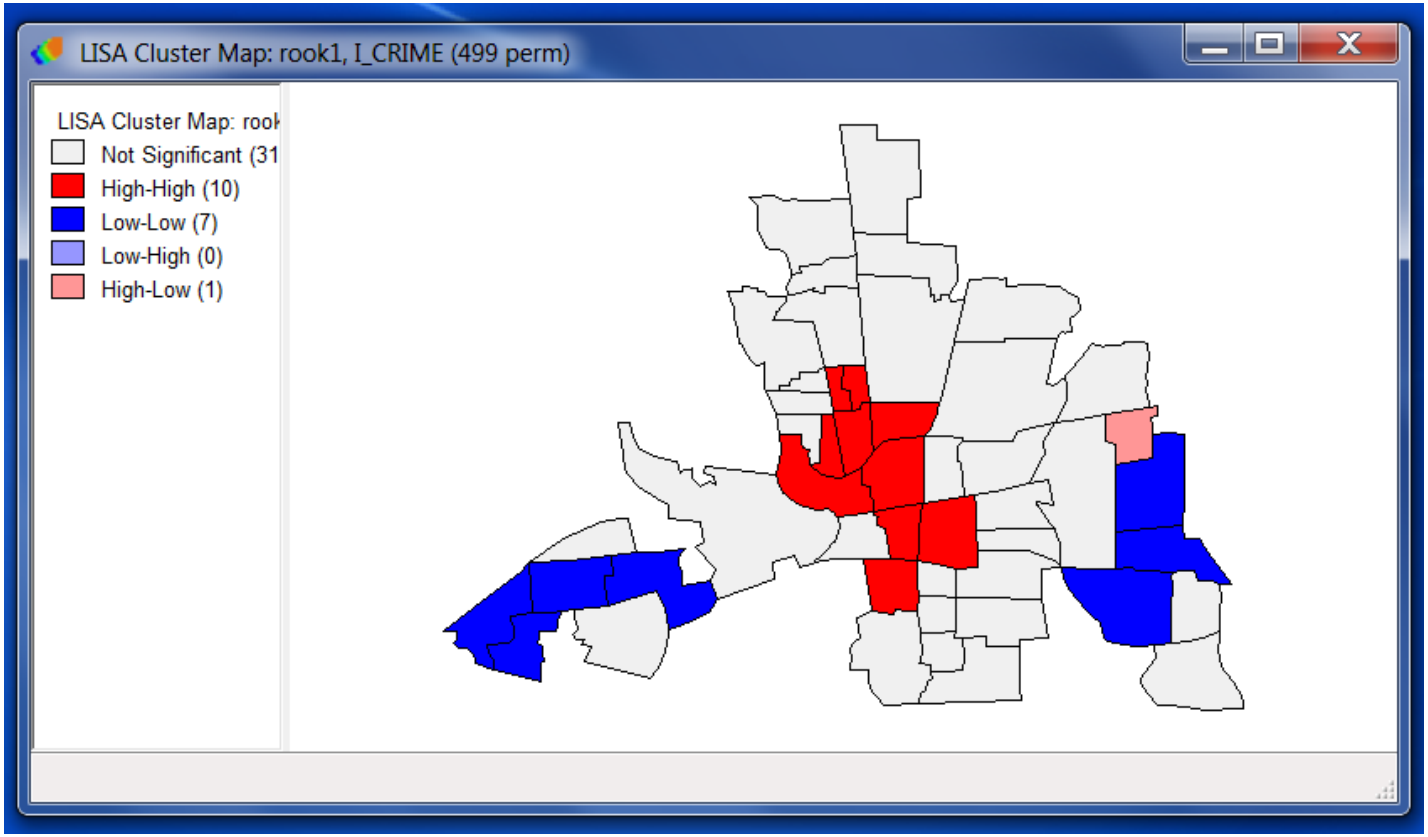
LISA_CL	LISA_I	NAME	STATE_NAME
2,0000000	0,4063892	Cumberland	Illinois
2,0000000	1,1044196	Lake	Colorado
0,0000000	-0,0126622	Highland	Ohio
4,0000000	-2,8967002	Baltimore City	Maryland
0,0000000	0,1267298	Kent	Delaware
2,0000000	2,7504658	Howard	Maryland
0,0000000	0,0376040	Mesa	Colorado
2,0000000	1,1310488	Pitkin	Colorado
2,0000000	2,2066823	Montgomery	Maryland
0,0000000	0,0924545	Audrain	Missouri
0,0000000	0,0992801	Howard	Missouri
2,0000000	0,8703036	Bartholomew	Indiana
2,0000000	0,9870944	Brown	Indiana
2,0000000	0,2193181	Monroe	Indiana
0,0000000	0,0038556	Grant	West Virginia
0,0000000	0,1064690	Cape May	New Jersey

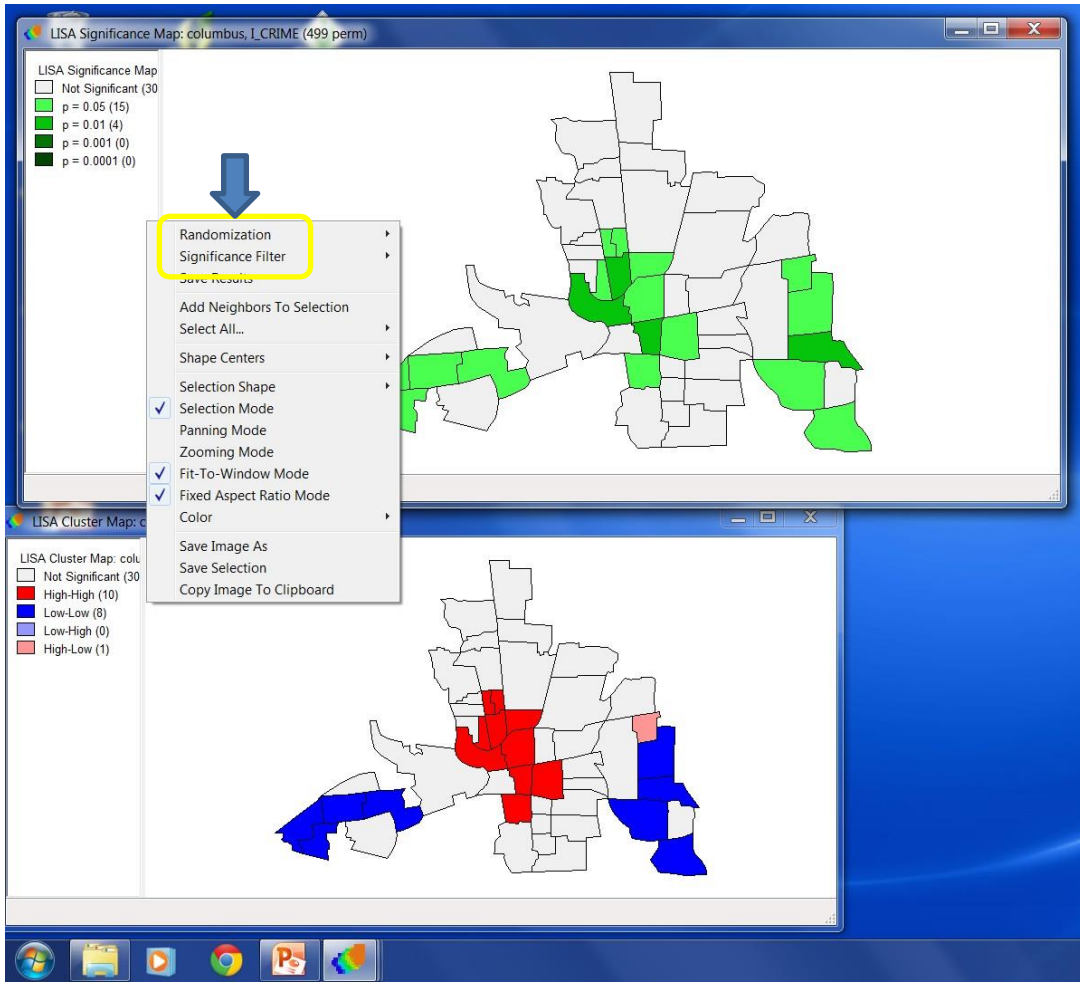
The result for univariate LISA is a special chloropleth map showing those locations with a significant local Moran statistic (depending on the significance level). In the map blow**, the significance map** is shown for the CRIME variable in the Columbus Data set, using rook contiguity.



The result of the cluster map is a special choropleth map showing those locations with a significant local Moran statistic Classified by type of spatial correlation : bright red for the high-high association and bright blue for low-low.

The high-high and low-low locations suggest clustering of similar values, while the high-low and low-high locations Indicate spatial outliers.



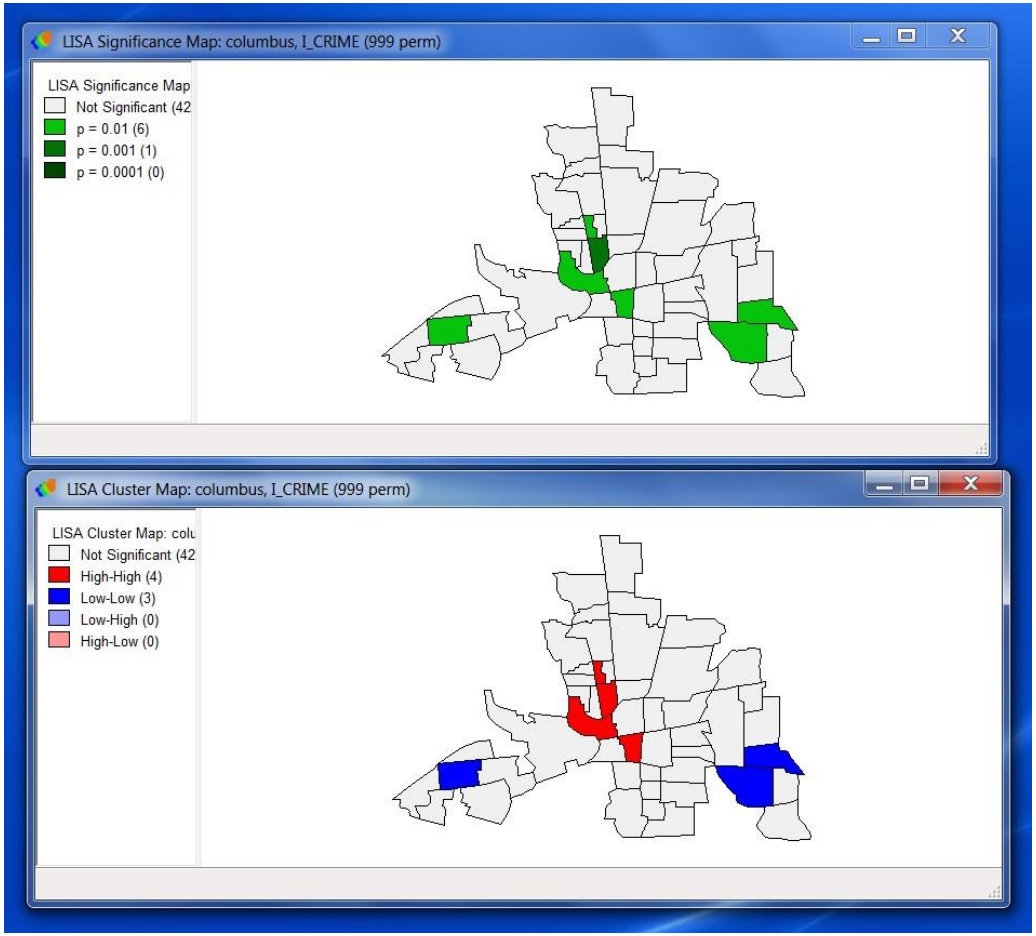


It is strongly recommended that sensitivity analysis be carried out before interpreting results of LISA maps as "significant" clusters.

The randomization option provides a way to address numerical stability of the results.

The significance filter is designed to assess how conclusions depend on the chosen significance level.

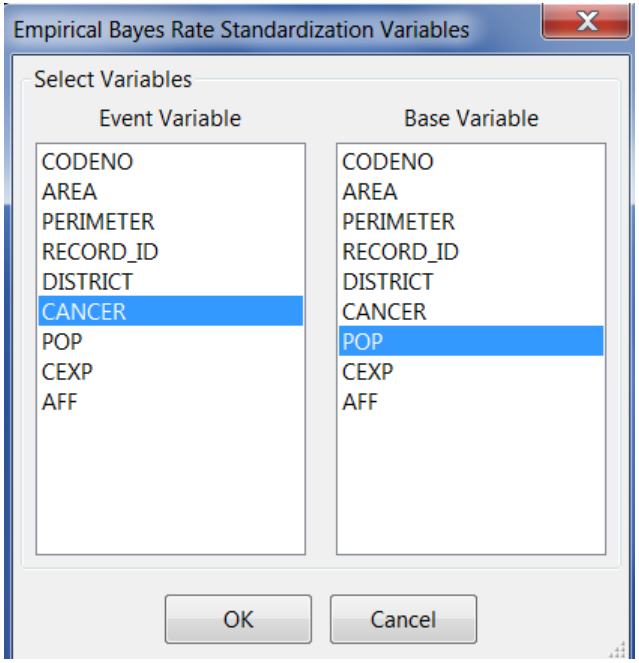
LISA maps after applying a significance filter.

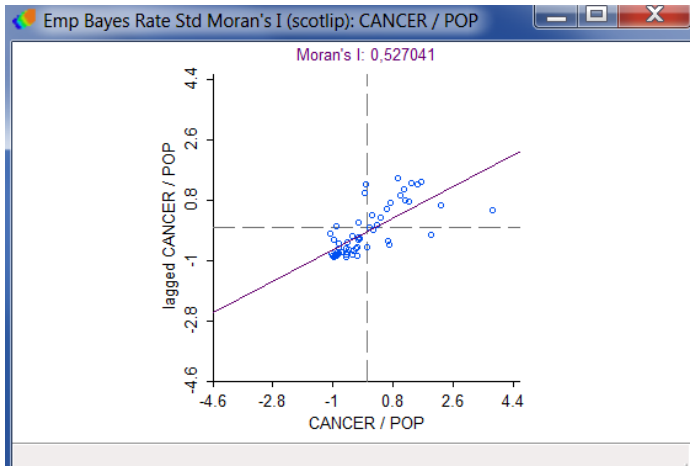


When Moran's I statistic is calculated for rates or proportions, the underlying assumption of stationarity may be Violated by the instrinsic instability of rates. The latter follows when the population at risk (the base) varies Considerably across observations. The variance instavility mat lead to spurious inferences for Moran's I.

To correct for this, GeoDa implements the **Empirical Bayes (EB) standardization**. This is implemented for both the global (Moran scatter plot) and local spatial autocorrelation statistics.

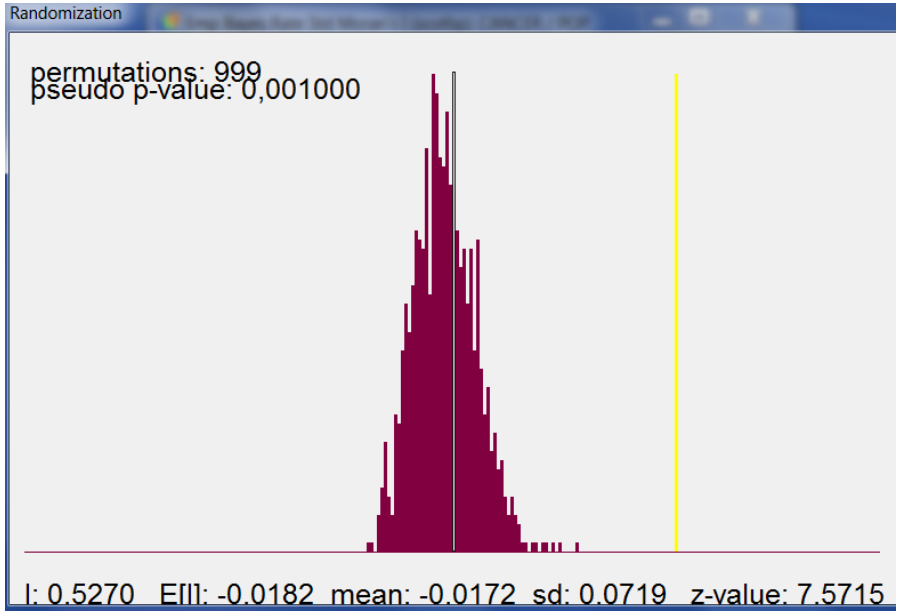
To illustrate this, we will use the Scottish lip cancer data set and associated weights file to compare the results of calculating Moran's I based on the non-standardized rates with the results of the EB standardization.





The value for Moran's I of 0.527 differs somewhat from the statistic for the unstandardized rates (0.479).

More important is to assess whether or not inference is affected. The resulting permulation distribution still suggests a highly significant statistic.

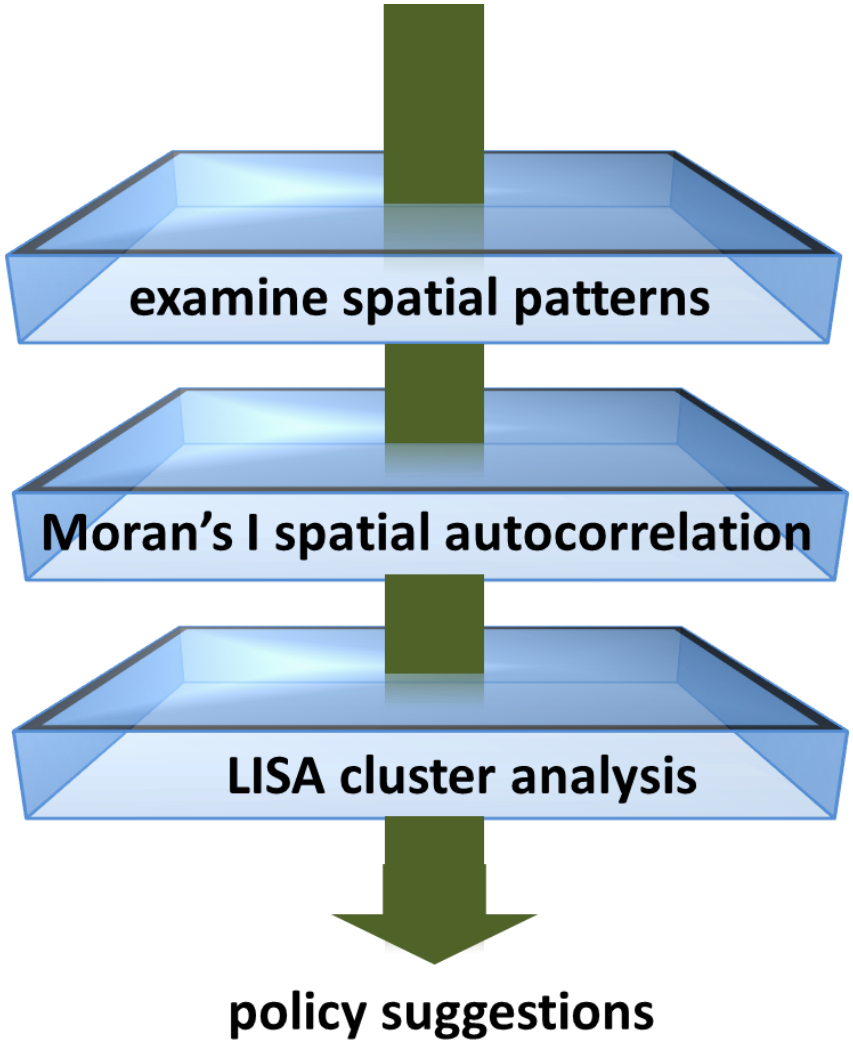


- **Practice** : ***Spatial patterns of rural poverty : An exploratory analysis in the São Fransisco River Bassin***, Brazil (Nove Economia_Belo_Horizonte_21 (1), 45-66_janeiro-abril de 2011).

This study uses recently released municipio-level data on rural poverty in Brazil to identify and analyze spatial patterns of rural poverty in the SFRB.

Moran's I statistics are generated and used to test for spatial autocorrelation, and to prepare cluster maps that locate rural poverty "hot spots" and "cold spots".

The results indicate that poverty reduction in the SFRB should take into account the spatial distribution of poverty. Not only is poverty in the SFRB clustered spatially, but the bulk of the bassin's poor resides in municipios that comprise the poverty "hot spots" the study identifies. These clusters did not correspond to state-level boundaries, so scope may exist for geographically refocusing poverty reduction efforts to make them more efficient.

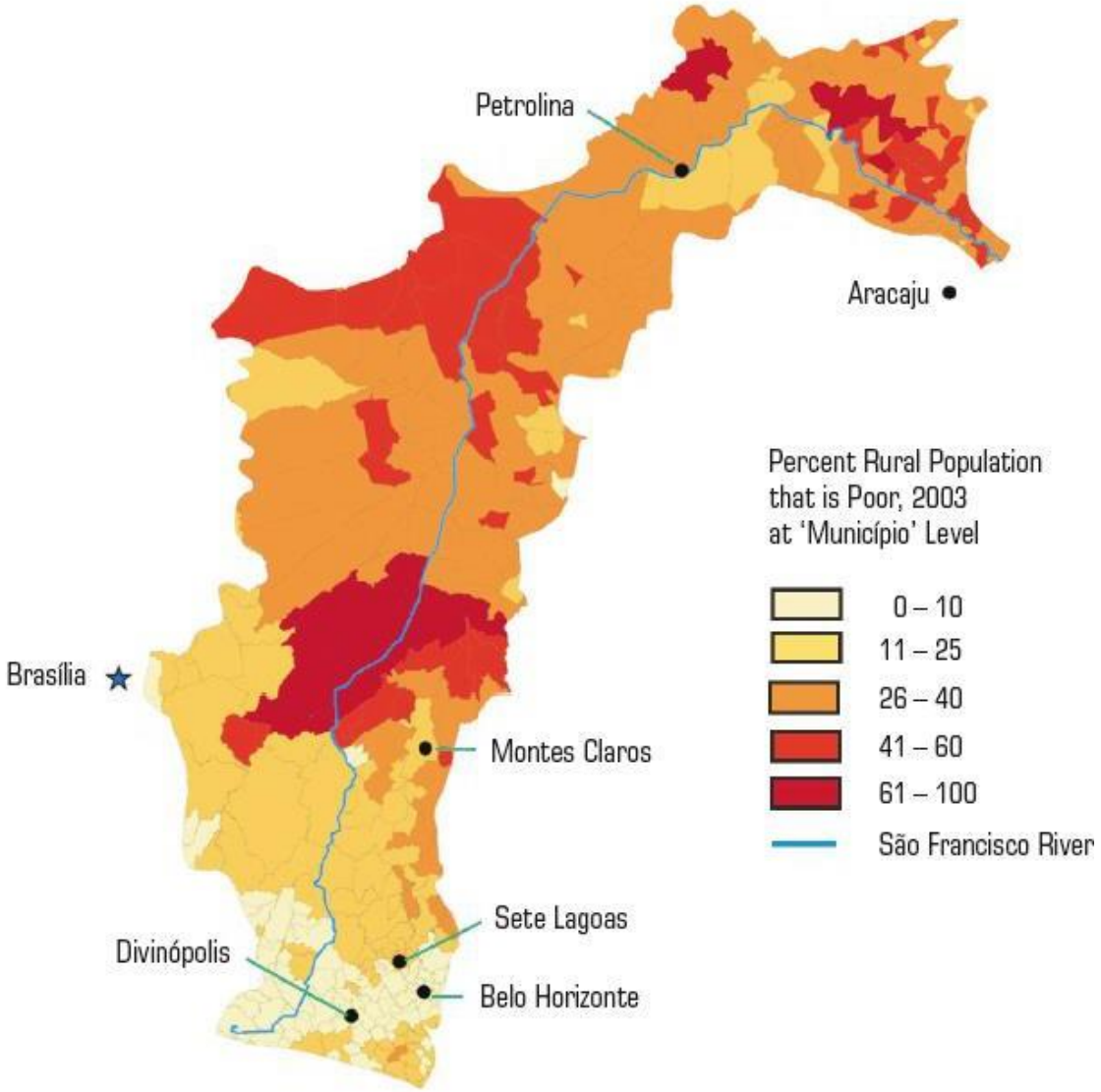


Construct maps to examine the distribution of rural poverty across the basin.

Statistically confirm the spatial patterns of rural poverty.

Redesign poverty programs to make them more effective.

Figure 1_ São Francisco River Basin: Percent Rural Population that is Poor, 2003



Data Source: Azzoni *et al.* (2006). Data aggregated to match *município* boundaries as of 1991.

- Information on spatial patterns of rural poverty in the SFRB may shed light on **the importance of location as a causal factor per se**. Municipios may be more likely to have high (or low) rural poverty rates depending on where they are located geographically :

  - one obvious reason is the stock of natural resources (natural resources are not evely distributed across space) : for farm activities, for example, good soils and easy access to water may improve agricultural conditions, productivity and income ;
  - job and income providers such as firms and service-oriented businesses tend to concentrate in space in order to benefit from large markets (economies of scale) and the availability of specialized skilled labor.

- The value of Moran's I is equal to 0.72, which suggests a strong postitive spatial autocorrelation of rural poverty. This number suggests that for the SFRB, there are more locations wich high (low) rural poverty rates surrounded by locations with high (low) rural poverty rates than would be the case if poverty were distributed randomly.

- The value of Moran's I also suggests that poverty in the SRFB is spatially distributed in clusters and also suggests that poverty in neighboring areas increases the likelihood of poverty in its neighbors. However, the value of Moran's I does not tell us where rural poverty clusters might be, but rather suggests that **the spatial pattern of poverty is not random** (there is more similarity in poverty (or the absence of its) than would be expected if the pattern were random).

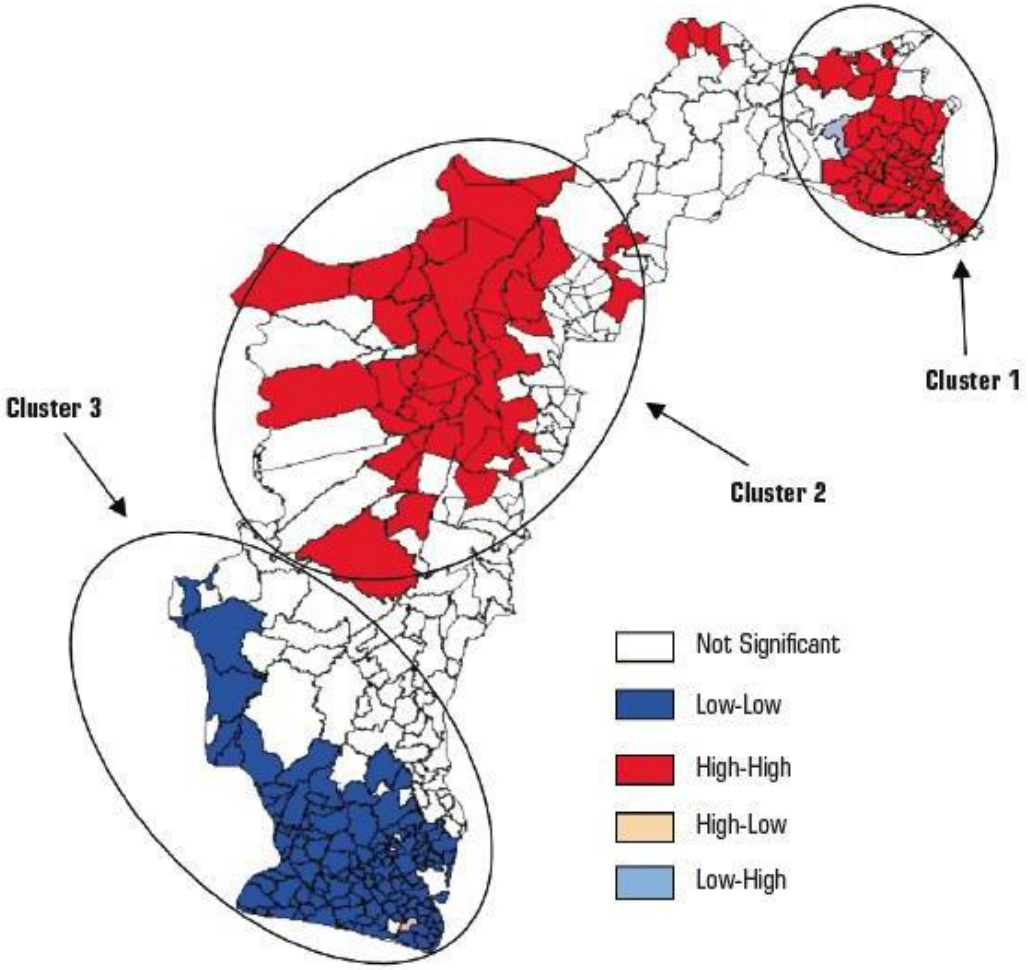
- Making use of **EB-standardization to reduce variance instability**, delivers a coefficient of 0.83 compared to the initial calculation of Moran's I. This indicates that the correlation between rural poverty rates in location i and neighboring locations is stronger when rates are standardized. Hence, increasing the precision with which rural poverty is measured will likely increase the spatial correlation among rural poverty rates in the SFRB.

- Although a Moran I of 0.83 strongly shows that the spatial distribution of rural poverty is not random, it does not locate poverty clusters.

- To locate “hot spots” and “cold spots”, local indicators of spatial autocorrelation must be used (LISA). LISA provides location-specific information and estimates the extent of spatial autocorrelation between the value of a given variable (rural poverty) in a particular location and the values of the same variable in locations around it. This makes it possible to **identify spatial clusters of rural poverty**.

- 3 clusters of rural poverty in the SFRB are detected by LISA. Clusters 1 and 2 are rural poverty “**hot spots**” and correspond to positive and high-high spatial autocorrelation, indicating spatial clusters of locations with above-average rural poverty rates. Cluster 3 is a “**cold spot**” and also corresponds to a positive, but low-low spatial autocorrelation, indicating a cluster of locations with below -average rural poverty rates.

Figure 3_ Local spatial clusters of rural poverty across the "municípios" in the São Francisco River Basin



Data Source: Data from Azzoni *et al.* (2006). Map developed by the authors.

- As mentioned before, the clusters of rural poverty may be attributable to several reasons. But further analysis is required to determine the causes of spatial patterns of rural poverty in the SFRB. **Multivariate regression analysis** that takes into account the variables that may explain poverty is the appropriate approach to the analysis of the spacial determinants of patterns of rural poverty in the SFRB.

- The results of this study suggest that poverty reduction policies in the SFRB should take into account the spatial distribution of poverty. The analysis suggests that location as a causal factor per se is important and locations are indeed more likely to have high (or low) rural poverty rates depending in where they are located in the basin. This may be due to obvious reasons such as stock of natural resources, soil quality, access to water, etc.

- More importantly, the analysis shows that poverty in one location is affected by (or affects) poverty in neighboring locations. That is, there are **spillovers**, either positive or negative externalities that make locations more or less likely to get out of poverty. These spillovers may be associated with the concentration (or lack of concentration) of firms, technology and knowledge. These results set the stage for identifying factors that influence rural poverty in the SFRB, factors that may themselves be spatially correlated.

4. Spatial regression

- When moving from simple descriptive analyses to data modeling, analysts turn to **multivariate regression modeling** to account for variability in attribute values among geographic units by identifying other covariates of the attribute of interest.

- Attributes of spatially referenced data generally violate at least one of the assumptions underlying the standard regression model, which necessitates both caution regarding these violations and attention to methods designed to correct for them.

- **Spatial variation : spatial heterogeneity versus spatial dependence**

- When undertaking initial EDA of spatial data, it is worthwhile to develop a sense of the spatial distribution of the attribute values. By mapping the distributions of variables across space, a distinction can be made between **two types of spatial dependence**.

- **Spatial heterogeneity** : *large-scale regional differentiation* (among attribute values) is an important component of spatial variation. Spatial heterogeneity is the lack of stability across space of one or more attribute values. Heterogeneity gives recognition to the common observation that values of a variable are not the same across space.

- Spatial heterogeneity follows from the intrinsic uniqueness of each location. Spatial heterogeneity is consistent with the description of how places are particular moments of intersecting social relations. The unique combination of social forces together in one place may produce effects which would not happen otherwise. These social forces include nonmaterial forces (e.g. cultural and/or historical processes) that cannot easily or always be quantified, yet these forces shape otherwise measurable social relationships. The spacial regime approach permits the analyst to move beyond geography per se, by focusing on social, economic and demographic factors - or, combined , sociological factors – that comprise the context of place. This approach is intended to enable the analyst to address the “so what” question : what is it about a place that distinguishes it from other places ?

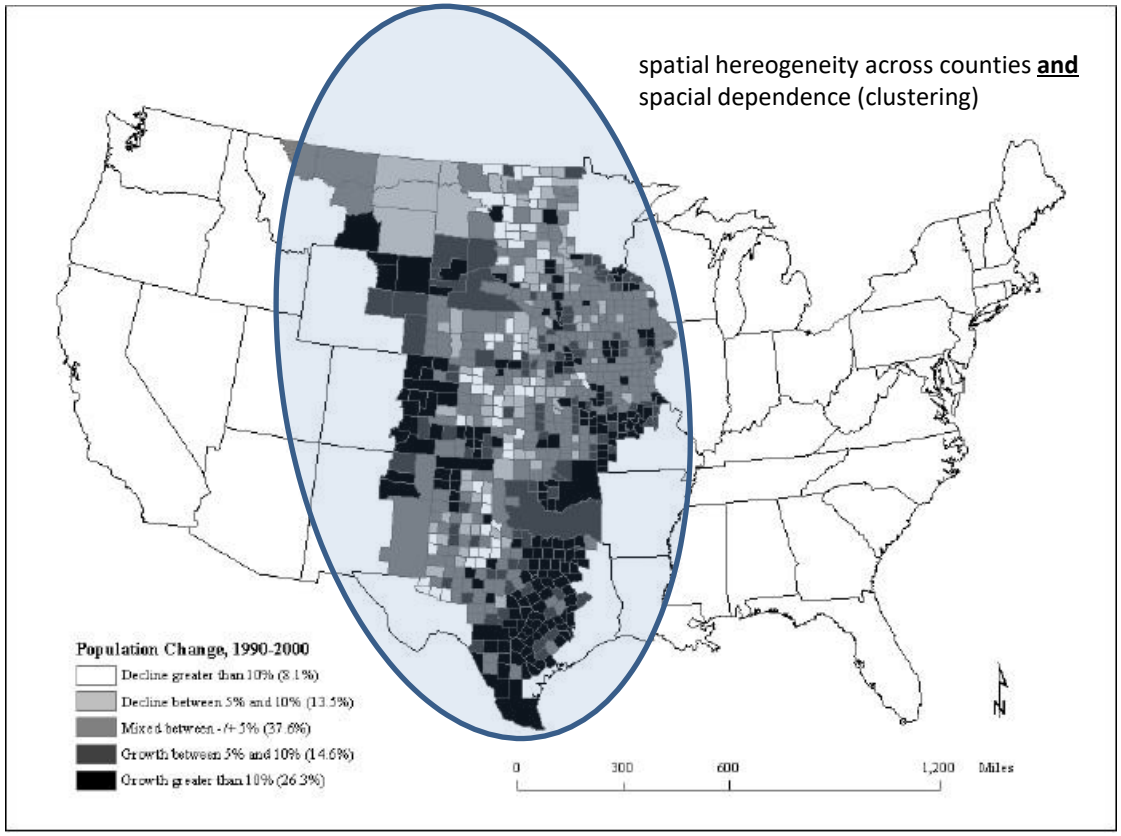
- **Spatial dependence** refers to *small-scale spatial effects* that manifest a lack of independence among observations (spatial clustering). The assumption is that dependence among the observations derives from spatial interaction among the units of analysis which can be defended theoretically and which can be statistically captured by a spatially lagged "neighborhood" effect.

- Two forms of spacial models are commonly used to improve regressions on spatially correlated data :

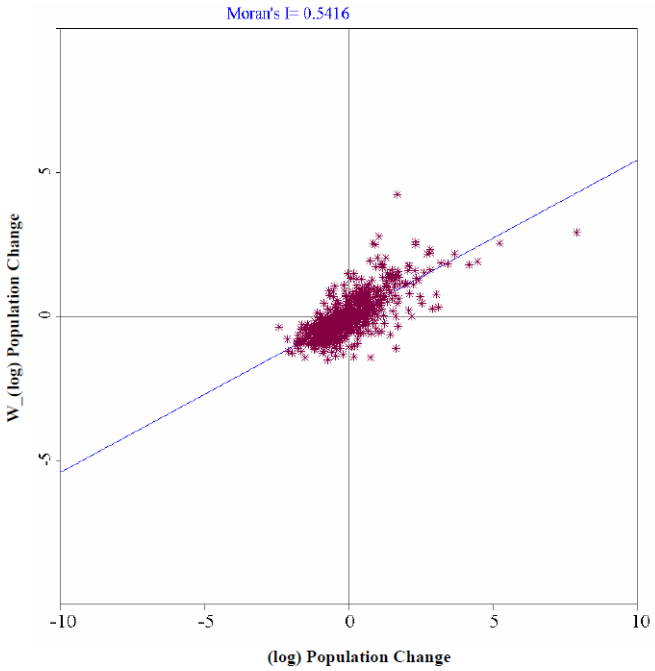
  - **The spacial lag model** : if two locations are adjacent, the value of the dependent variable of the first locations can be influenced by the value of the dependent variable of the other. This means that there is a contagion or dispersion effect, represented best by a spatial lag model.

  - **The spacial error model** : if the error residuals of locations are influenced by one another, this means that the phenomenon under study is not analysed at the correct geographical level, or that there might be an unobserved variable correlated with the spatial structure of the data. This would imply a clustering effect and this has to be studied by a spatial error model.

- A spatial lag model is appropriate if neighboring locations influence one another ; the spatial error model documents that locations geographically cluster but for an unknown reason.



Spatial distribution of population change among Great Plains Counties, 1990-2000



Moran scatterplot of population change

Source : P.R. Voss, K.J. Curtis White & R.B. Hammer : Explorations in spatial demography, in W.A. Kandel & D.L. Brown, Population change and rural society, Springer, 2006, pp. 407-429)

- A model with **spatial lags** is able to borrow information from neighborhood observations because of **the spatial autocorrelation among the units of analysis**. The units of analysis likely fail a formal statistical test of randomness and thus fail to meet a key assumption of classical statistics : *independence among observations*. With respect to statistical techniques that presume such independence (e.g. standard regression analysis), positive autocorrelation means that the spatially autocorrelated observations bring less information to the model estimation process than would the same number of independent observations.

- A carefully selected variable can account for spatial heterogeneity in the data and might boost the explanatory value of the model and largely remove the large-scale spatial process, but spatial autocorrelation would persist if a spatial dependence process were also indicated. There would remain in the data a more complicated, interactive spatial relationship among neighbors that suggests the requirement of some type of autoregressive term in the regression specification.

- The aim of the researcher is to specify and estimate a model that reasonably accounts for or incorporates that spatial effects present in the data. These effects can be modeled as spatial heteregeneity and spatial dependence. When first examining a spatial relationship, the reseacher must ask whether the association appears to be a reaction to some geophysical, cultural, social or economic force that works to create spatial patterning (spatial heterogeneity), or an interaction, indicative of spatial dependence.

- If the association is merely a reaction to some general force, then a modeling strategy with a standard regression structure may be appropriate.

- If, on the other hand, the association is an interaction suggesting some type of formal dependency among units, then a modeling strategy with a spatial dependent covariance structure is the way to proceed. In this instance, heterogeneity likely will not fully remove the spatial effects within the data. An alternative is needed – **a spatially oriented approach that formally incorporates a spatially lagged dependent variable or spatially lagged error term**.

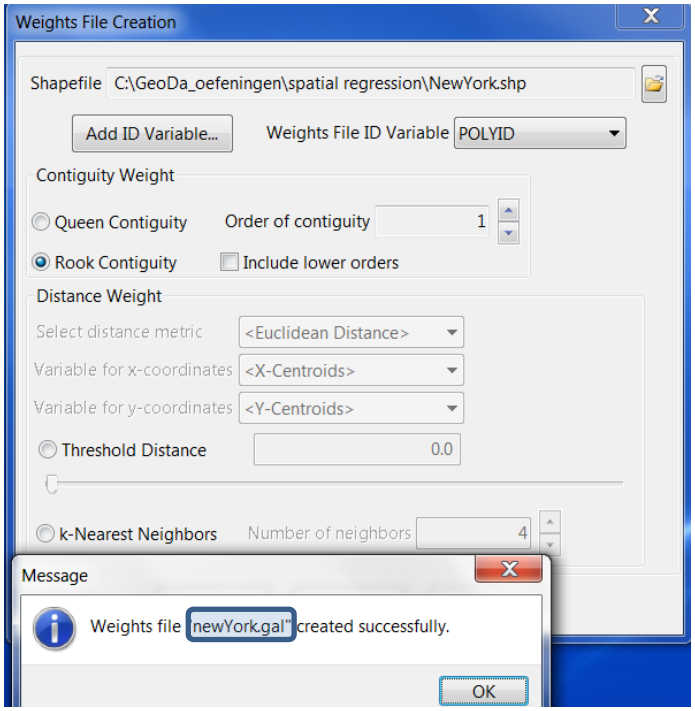
- ## Spatial dependency modeling : example 1

- The shapefile *newyork.shp* is the map of Manhattan in New York City with Census 2000 data* . These are socioeconomic attributes for 297 Census tracts. It includes the following variables:

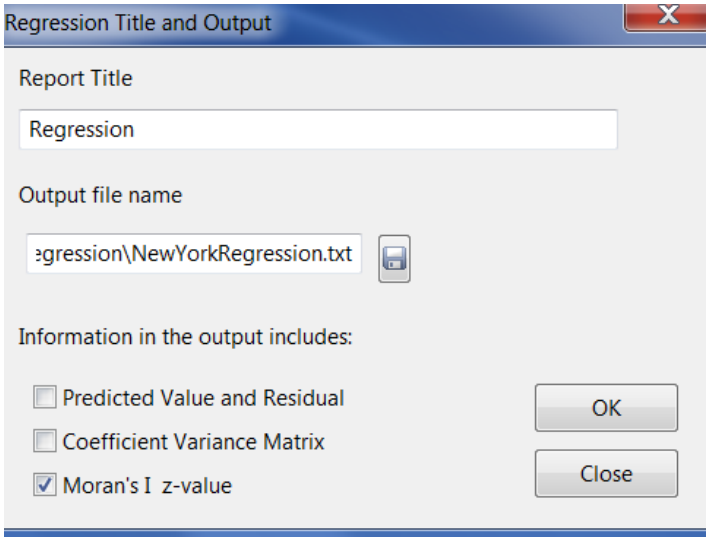
POLYID Polygon ID
STATE State FIPS
COUNTY County FIPS
TRACT Census Tract ID
sctrct00 FIPSID
hvalue Median housing value
t0_pop Total population
pctnhw Percent non-Hispanic white persons
pctnhb Percent non-Hispanic black persons
pcthsp Percent Hispanic persons
pctasn Percent Asian persons
t0p_own Percent homeowners
t0p_coll Percent college educated
t0p_prf Percent of people employed in professional/managerial occupations
t0p_uemp Percent of people unemployed
t0p_for Percent foreign born persons
t0p_rec Percent recent immigrants
t0_minc Median household income
t0p_poor Percent total population below poverty

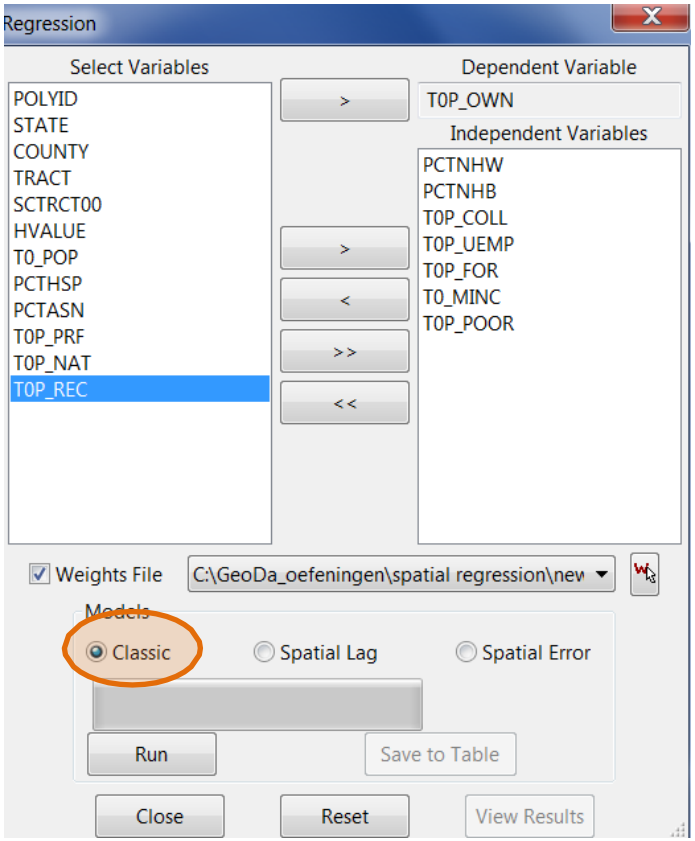
* Source : http://www.s4.brown.edu/S4/Training/Modul2/GeoDa3FINAL.pdf

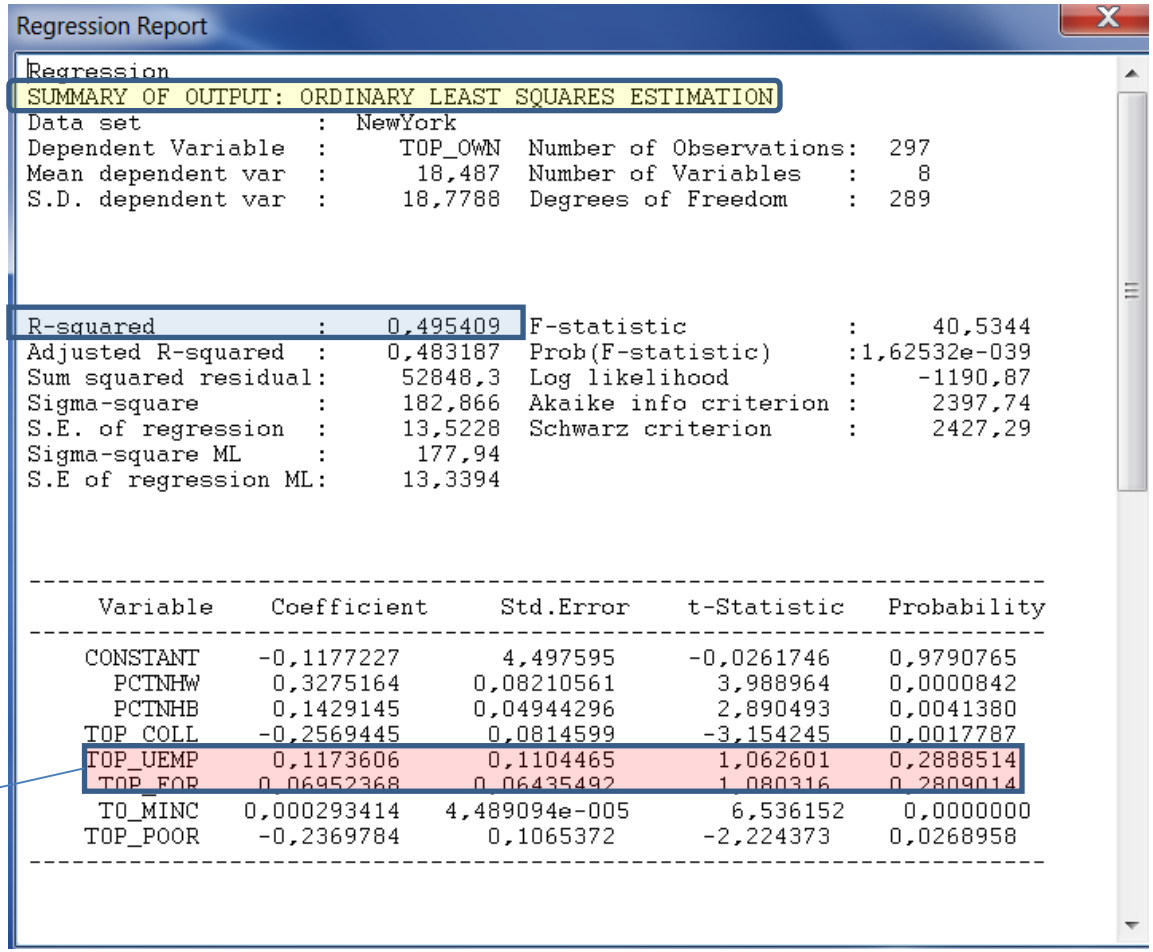
- Before starting a regression, create a weights file :



- In this example, we will predict neighborhood homeownership with several indicators :

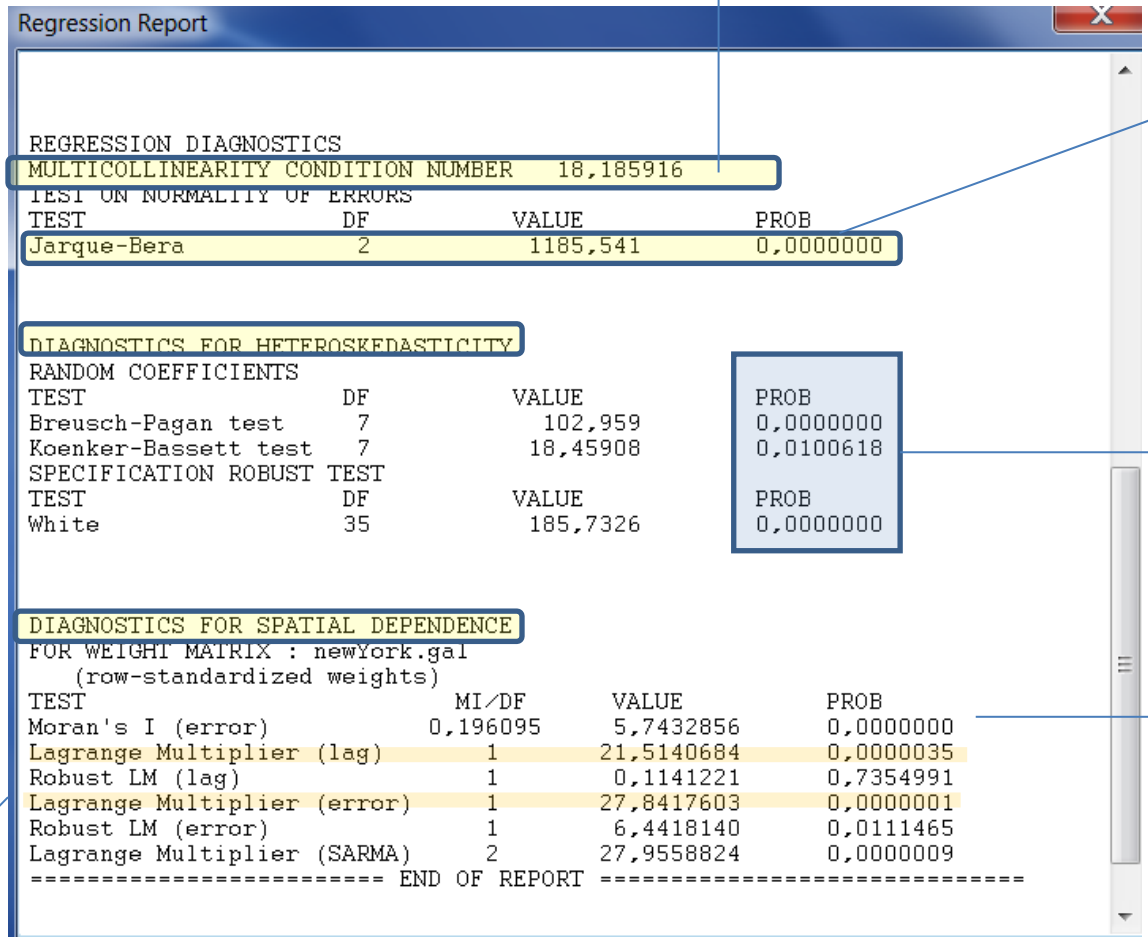






insignificant effects

Test of multicollinearity of the model : one should be alarmed when the condition number is greater than 20.



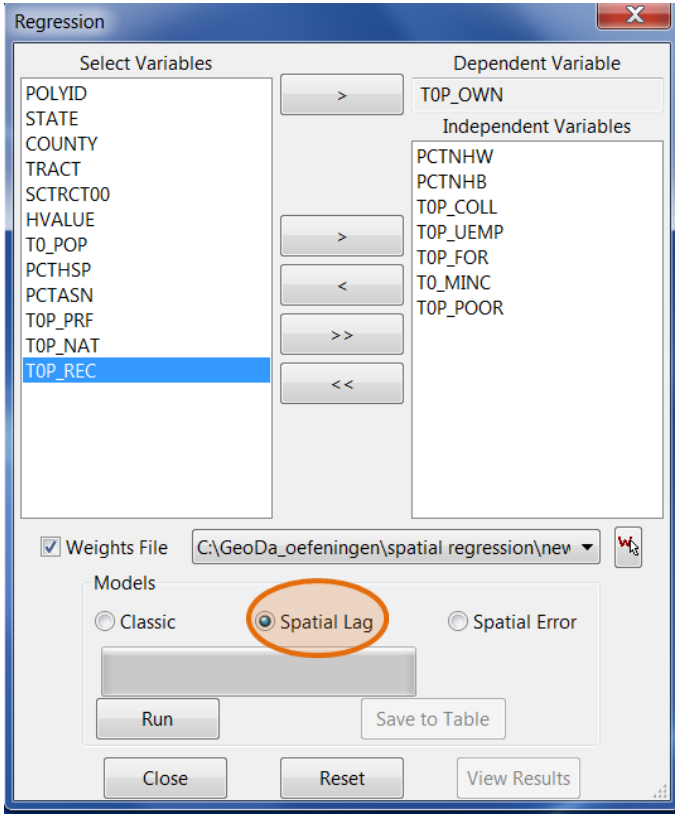
Jarque-Bara test is used to examine the normality of the distribution of the errors. The low probability of the test score suggests non-normal distribution of the error term.

The low probabilities of the three tests point to the existence of heteroscedasticity. Error variance can be affected by spatial dependence in the data.

Moran's I suggests spatial autocorrelation of the residuals.

Both tests of the lag and error are significant, indicating presence of spatial dependence.
The robust test help us understand what type of spatial dependence may be at work. The robust measure for error is still significant, but the robust lag test becomes insignificant, which means that when the lagged dependent variable is present the error dependence disappears.

- After identifying the presence of spatial dependence, we will use GeoDa to re-estimate the model when controlling for spatial dependence.



Coefficient Rho reflects the spatial dependence in the sample data, measuring the average influence on observations by their neighboring observations.

The spatial lag term of homeownership (W_TOP_OWN) appears as an additional indicator. It has a positive effect and is highly significant. As a result, the model fit is improved (higher R-square).

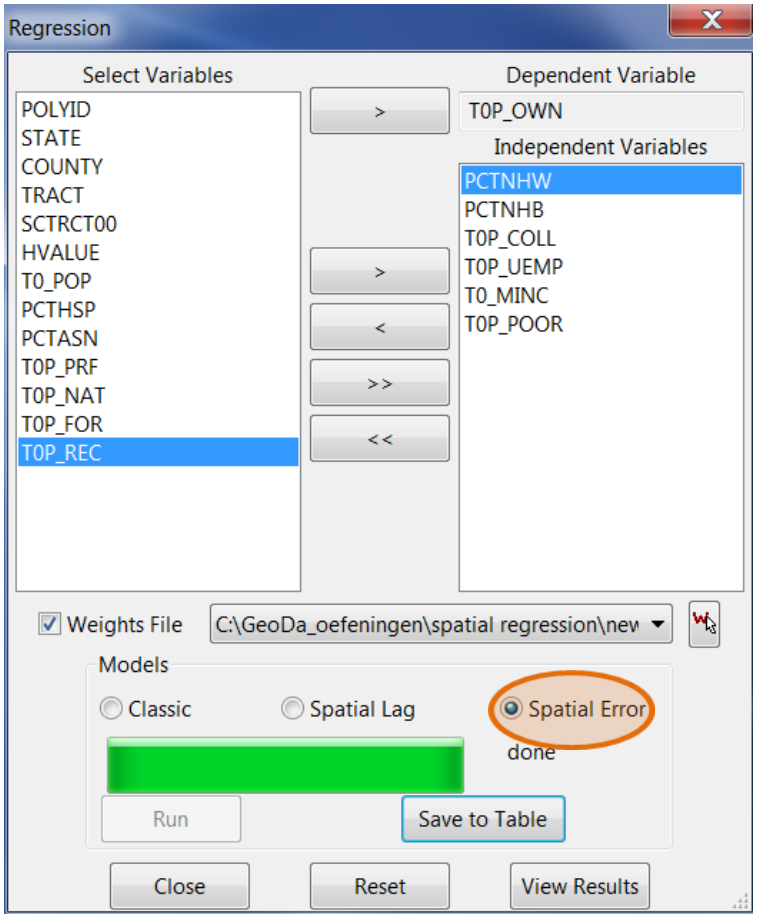
Although the introduction of the spacial lag term improved the model fit , it didn't make the spacial effects go away.

Regression Report

```
Regression
SUMMARY OF OUTPUT: SPATIAL LAG MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : NewYork
Spatial Weight      : newYork.gal
Dependent Variable  :     TOP_OWN  Number of Observations:  297
Mean dependent var  :      18,487  Number of Variables   :    9
S.D. dependent var  :     18,7788  Degrees of Freedom    :  288
Lag coeff.   (Rho)  :    0,244394

R-squared           :    0,526518  Log likelihood        :    -1183,23
Sq. Correlation     : -            Akaike info criterion :     2384,45
Sigma-square        :      166,97  Schwarz criterion     :      2417,7
S.E of regression   :     12,9217

------------------------------------------------------------------------
    Variable     Coefficient     Std.Error    z-value      Probability
------------------------------------------------------------------------
   W_TOP_OWN      0,2443939      0,0710086      3,44175    0,0005781
    CONSTANT      -2,432615       4,334781    -0,5611852    0,5746712
      PCTNHW      0,2890793     0,08031199      3,599454    0,0003190
      PCTNHB      0,1438108     0,04733031       3,03845    0,0023781
    TOP_COLL     -0,2382829     0,07828846     -3,043653    0,0023374
    TOP_UEMP      0,1286291      0,1055537      1,218613    0,2229911
     TOP_FOR     0,08309245     0,06151569      1,350752    0,1767749
     TO_MINC  0,0002428165   4,391116e-005      5,529721    0,0000000
    TOP_POOR      -0,222484      0,1020332     -2,180506    0,0292199
------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                                    DF      VALUE          PROB
Breusch-Pagan test                      7       94,56192       0,0000000

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL LAG DEPENDENCE FOR WEIGHT MATRIX : newYork.gal
TEST                                    DF      VALUE          PROB
Likelihood Ratio Test                   1       15,28552       0,0000924
========================= END OF REPORT =============================
```

cfr. R^2= 0.495 with OLS regression

- Now let's review the results for the spatial error model.



Coefficient of spatially correlated errors is positive. The model fit is improved (higher R^2).

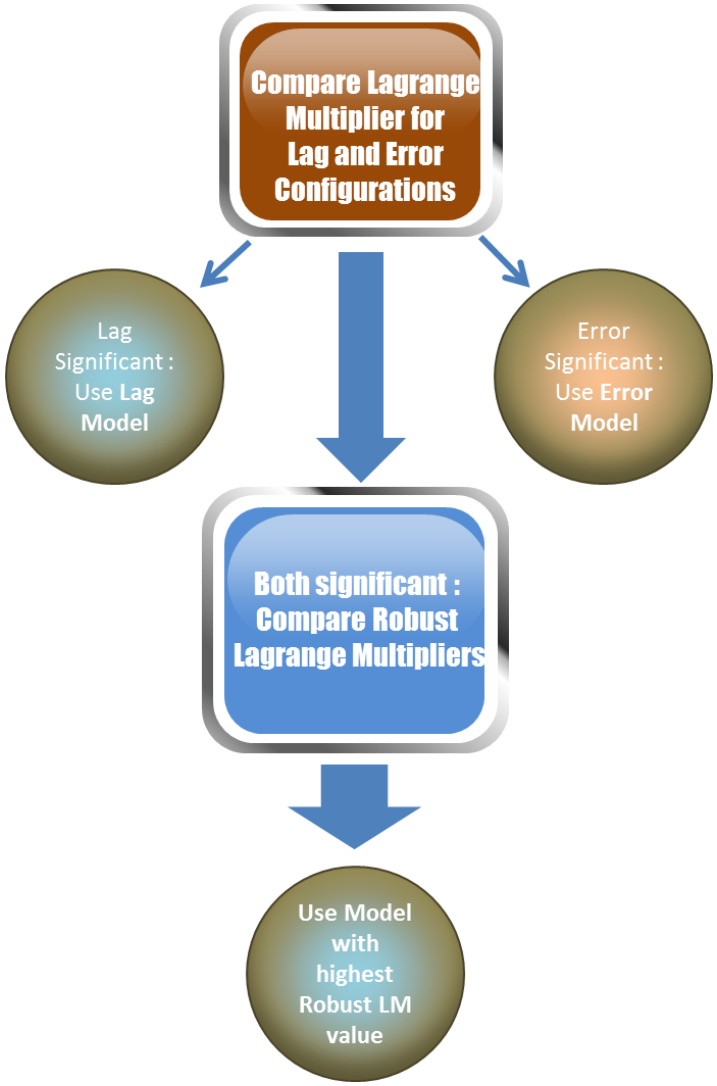
Heteroscedasticity remais significant. Also, spatial error stays significant. Although allowing the error terms to be spatially correlated improved the model fit, it didn't make the spatial effects go away.

Regression Report

```
Regression
SUMMARY OF OUTPUT: SPATIAL ERROR MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : NewYork
Spatial Weight      : newYork.gal
Dependent Variable  :     TOP_OWN   Number of Observations:  297
Mean dependent var  :   18,487003  Number of Variables   :    7
S.D. dependent var  :   18,778784  Degrees of Freedom    :  290
Lag coeff. (Lambda) :    0,291555

R-squared           :    0,530663  R-squared (BUSE)      : -
Sq. Correlation     : -            Log likelihood        :-1182,719769
Sigma-square        :     165,508  Akaike info criterion :     2379,44
S.E of regression   :      12,865  Schwarz criterion     :      2405,3

------------------------------------------------------------------------
    Variable     Coefficient     Std.Error    z-value      Probability
------------------------------------------------------------------------
    CONSTANT       2,52046       3,817573      0,6602258    0,5091088
      PCTNHW     0,3307332     0,07759524       4,262288    0,0000202
      PCTNHB     0,1398923      0,0508603       2,750521    0,0059502
    TOP_COLL    -0,2540283     0,08263879      -3,073959    0,0021125
    TOP_UEMP     0,0869458      0,1036599      0,8387599    0,4016040
     TO_MINC   0,0002603595     4,57898e-005       5,685971    0,0000000
    TOP_POOR    -0,2237273     0,09935776      -2,251735    0,0243390
      LAMBDA     0,2915555      0,0808238       3,607297    0,0003095
------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                                    DF      VALUE          PROB
Breusch-Pagan test                      6       38,44156       0,0000009

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL ERROR DEPENDENCE FOR WEIGHT MATRIX : newYork.gal
TEST                                    DF      VALUE          PROB
Likelihood Ratio Test                   1       17,49713       0,0000288
========================= END OF REPORT ==============================
```

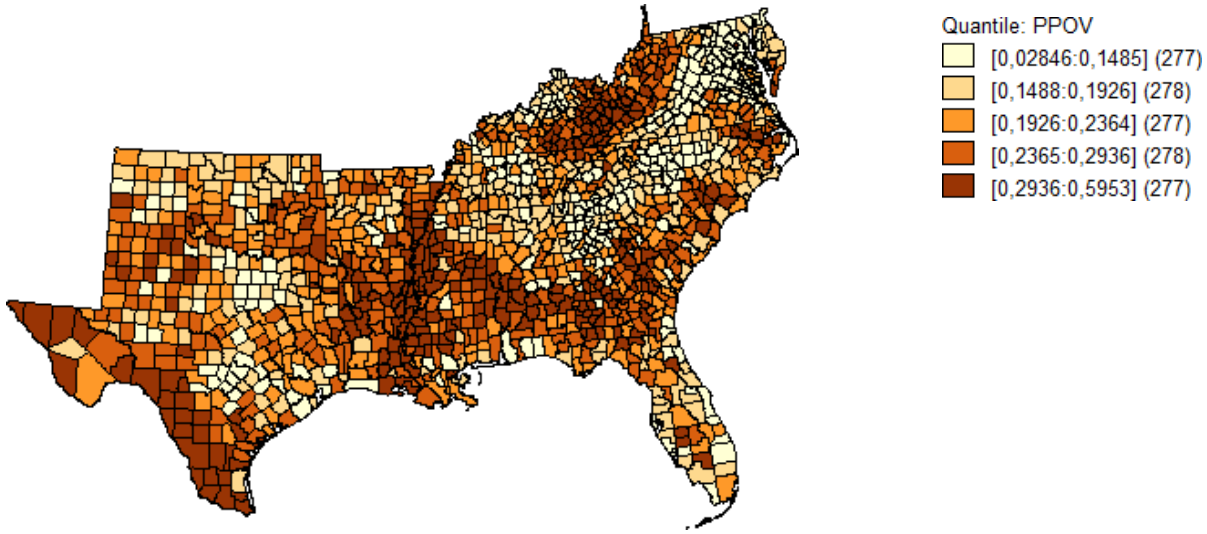
- Comparing the spatial lag and spatial error models, we can see that both models yield improvement to the original OLS model. Therefore, controlling spatial dependence improves model performance.

- Now the question is which of the two models is better ? To some extent, this is an open question. The general advice is first to look for a theoretical basis to inform your choice. When it is not so clear theoretically, you can compare the model performance parameters : the R-squared and log likelihood. In this example, the spatial error model has greater R-squared and log likelihood values. That provides a statistical basis to adopt this solution.

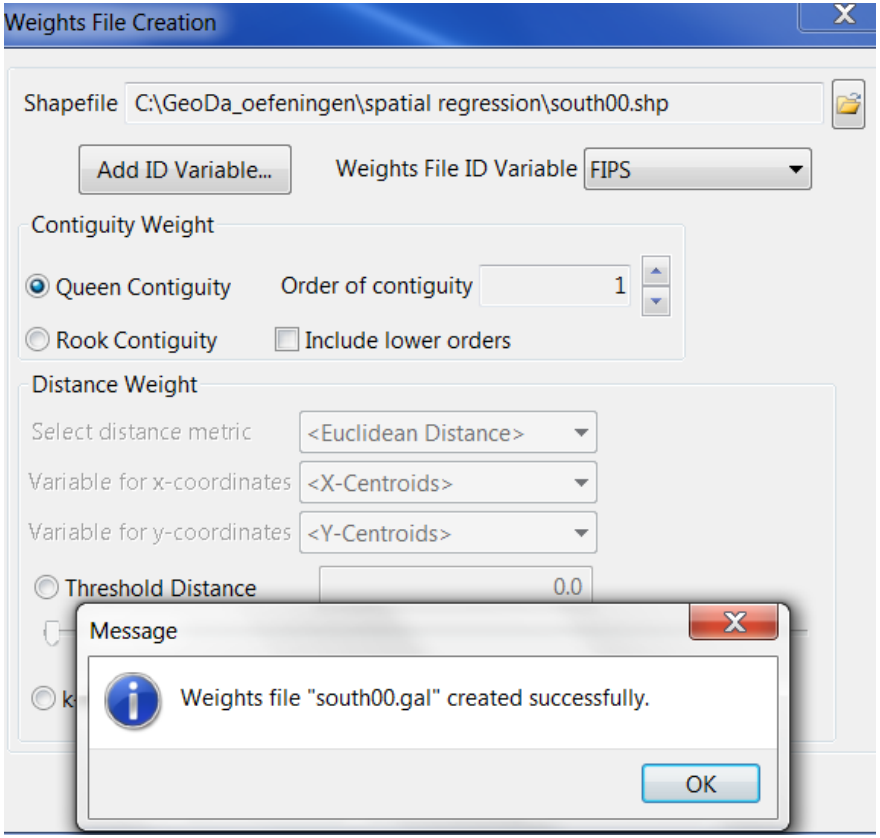

Compare Lagrange Multiplier for Lag and Error Configurations
Lag Significant : Use Lag Model
Error Significant : Use Error Model
Both significant : Compare Robust Lagrange Multipliers
Use Model with highest Robust LM value

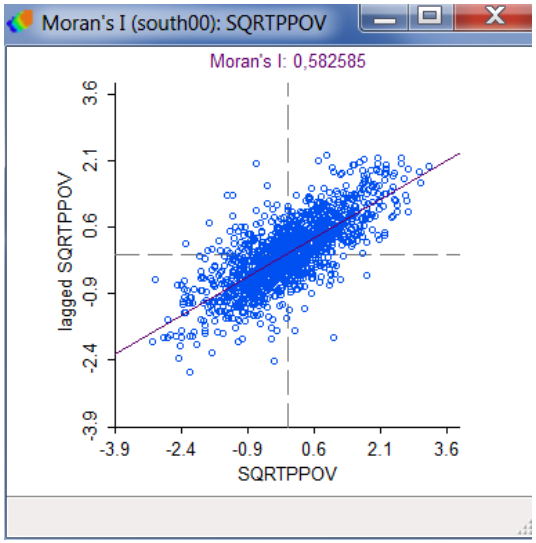
- **Spatial dependency modeling : example 2**

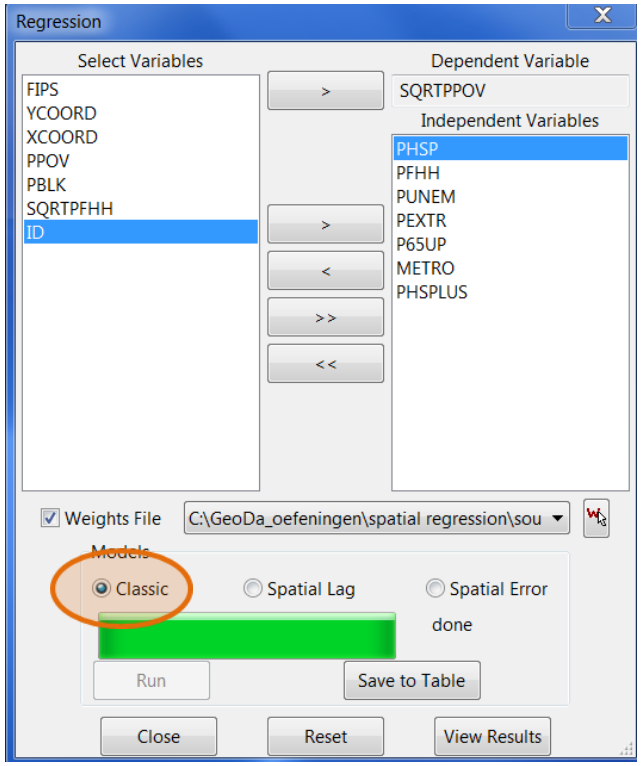
- Analysis of poverty in the U.S. *



Source : http://csde.washington.edu/services/gis/workshops/SPREG.html


Weights File Creation
Shapefile C:\GeoDa_oefeningen\spatial regression\south00.shp
Add ID Variable...
Weights File ID Variable FIPS
Contiguity Weight
Queen Contiguity
Order of contiguity 1
Rook Contiguity
Include lower orders
Distance Weight
Select distance metric <Euclidean Distance>
Variable for x-coordinates <X-Centroids>
Variable for y-coordinates <Y-Centroids>
Threshold Distance 0.0
Message
Weights file "south00.gal" created successfully.
OK

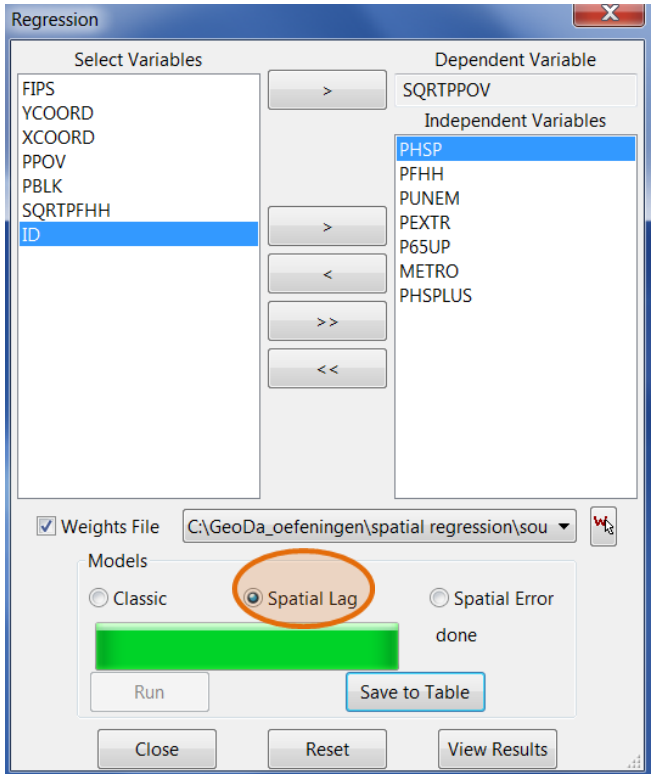

Moran's I (south00): SQRTPPOV
Moran's I: 0,582585
lagged SQRTPPOV
SQRTPPOV
-3.9
-2.4
-0.9
0.6
2.1
3.6



violation of regression assumptions

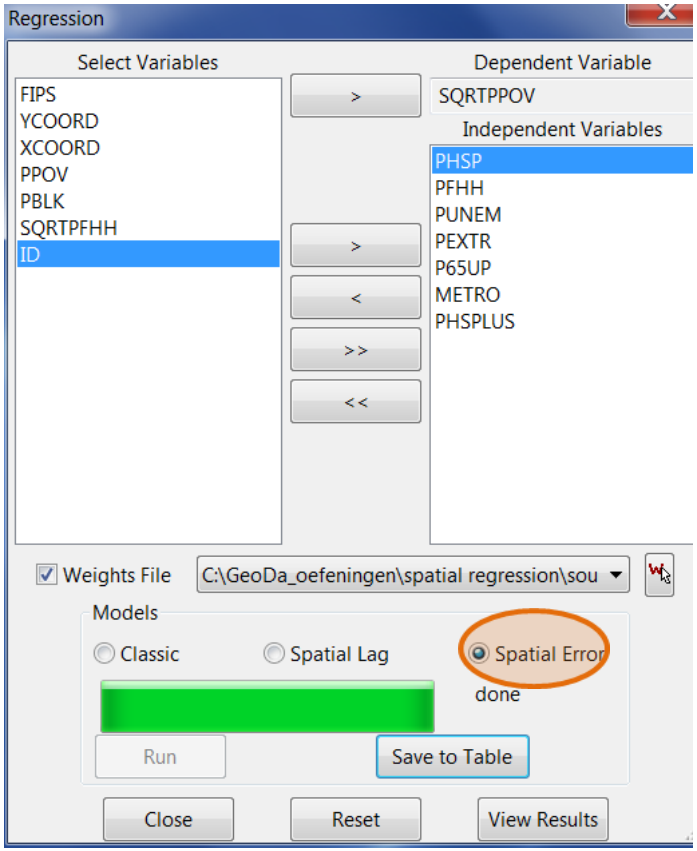
Regression Report

```
R-squared           :     0,779727  F-statistic           :      697,343
Adjusted R-squared  :     0,778608  Prob(F-statistic)     :            0
Sum squared residual:      2,84725  Log likelihood        :      2323,69
Sigma-square        :   0,00206472  Akaike info criterion :     -4631,38
S.E. of regression  :    0,0454392  Schwarz criterion     :     -4589,5
Sigma-square ML     :   0,00205281
S.E of regression ML:     0,045308


------------------------------------------------------------------------
     Variable     Coefficient      Std.Error    t-Statistic   Probability
------------------------------------------------------------------------
    CONSTANT      0,3093621     0,009790428        31,59842     0,0000000
        PHSP     0,07107947     0,009903332        7,177329     0,0000000
        PFHH      0,5292559      0,02109148        25,09335     0,0000000
       PUNEM       1,460965      0,06340116        23,04319     0,0000000
       PEXTR      0,3446715      0,02537929        13,58082     0,0000000
       P65UP        0,22194      0,03574824        6,208416     0,0000000
       METRO    -0,01047755     0,003193159       -3,28125     0,0010593
     PHSPLUS     -0,2835089      0,01369503       -20,70159     0,0000000
------------------------------------------------------------------------


REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER    21,917102
TEST ON NORMALITY OF ERRORS
TEST                     DF          VALUE            PROB
Jarque-Bera               2          140,0158         0,0000000


DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                     DF          VALUE            PROB
Breusch-Pagan test        7          220,1356         0,0000000
Koenker-Bassett test      7          124,4715         0,0000000
SPECIFICATION ROBUST TEST
TEST                     DF          VALUE            PROB
White                    35            N/A            N/A


DIAGNOSTICS FOR SPATIAL DEPENDENCE
FOR WEIGHT MATRIX : south00.gal
   (row-standardized weights)
TEST                           MI/DF       VALUE           PROB
Moran's I (error)           0,308056     19,3728246      0,0000000
Lagrange Multiplier (lag)          1      300,8632892     0,0000000
Robust LM (lag)                    1       71,0395569     0,0000000
Lagrange Multiplier (error)        1      362,2829911     0,0000000
Robust LM (error)                  1      132,4592589     0,0000000
Lagrange Multiplier (SARMA)        2      433,3225481     0,0000000
======================== END OF REPORT =============================
```



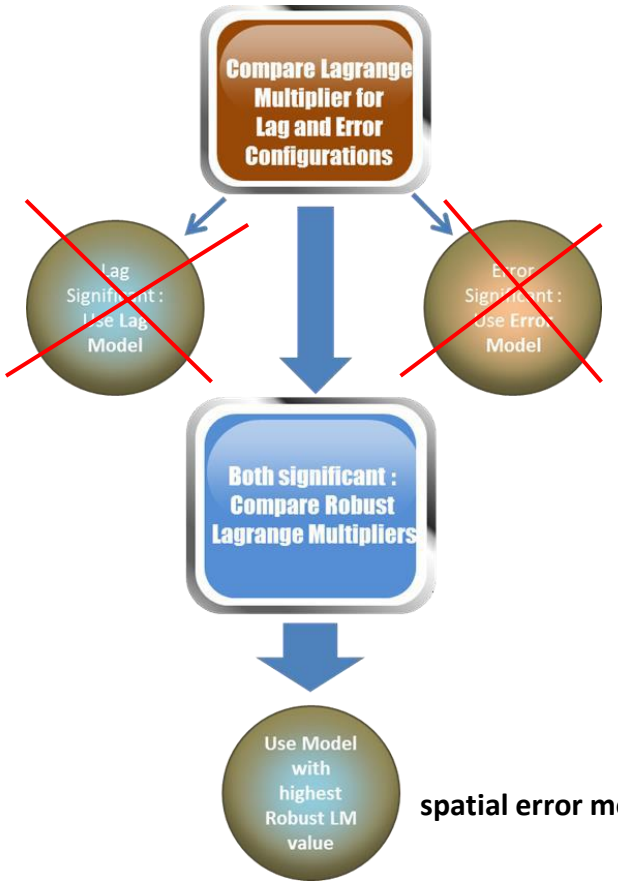
Regression Report

```
Regression
SUMMARY OF OUTPUT: SPATIAL LAG MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : south00
Spatial Weight      : south00.gal
Dependent Variable  :    SQRTPPOV  Number of Observations: 1387
Mean dependent var  :    0,464095  Number of Variables   :    9
S.D. dependent var  :   0,0965369  Degrees of Freedom    : 1378
Lag coeff.   (Rho)  :     0,33423


R-squared           :     0,822248  Log likelihood        :     2457,37
Sq. Correlation     : -             Akaike info criterion :    -4896,74
Sigma-square        :   0,00165654  Schwarz criterion     :    -4849,62
S.E of regression   :    0,0407006

------------------------------------------------------------------------
    Variable     Coefficient      Std.Error     z-value       Probability
------------------------------------------------------------------------
  W_SQRTPPOV      0,3342297       0,0200363       16,6812     0,0000000
    CONSTANT      0,1850717      0,01117582       16,56002    0,0000000
        PHSP     0,06413139     0,008918838       7,190554    0,0000000
        PFHH      0,4595375      0,02030913       22,62714    0,0000000
       PUNEM       1,061432      0,05982543       17,74215    0,0000000
       PEXTR      0,2397121      0,02361957       10,14887    0,0000000
       P65UP      0,2189879      0,03202753       6,837492    0,0000000
       METRO  -0,007535908     0,002861533      -2,633521    0,0084505
     PHSPLUS    -0,2428268      0,01253364      -19,37401    0,0000000
------------------------------------------------------------------------


REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                                    DF      VALUE          PROB
Breusch-Pagan test                      7        307,5235      0,0000000


DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL LAG DEPENDENCE FOR WEIGHT MATRIX : south00.gal
TEST                                    DF      VALUE          PROB
Likelihood Ratio Test                   1        267,362       0,0000000
========================= END OF REPORT ==============================
```



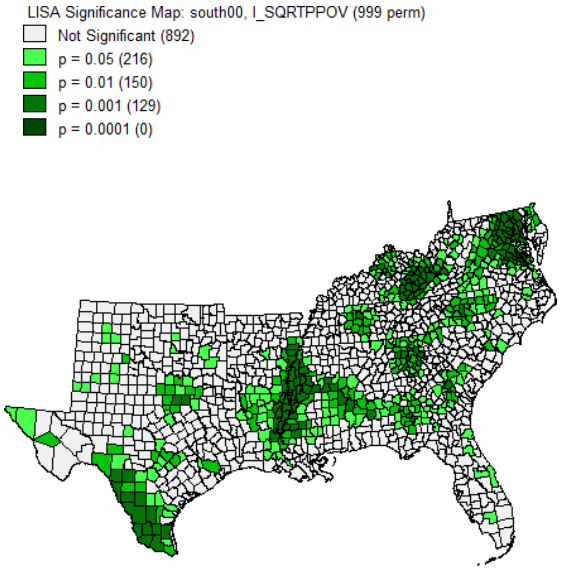
Regression Report

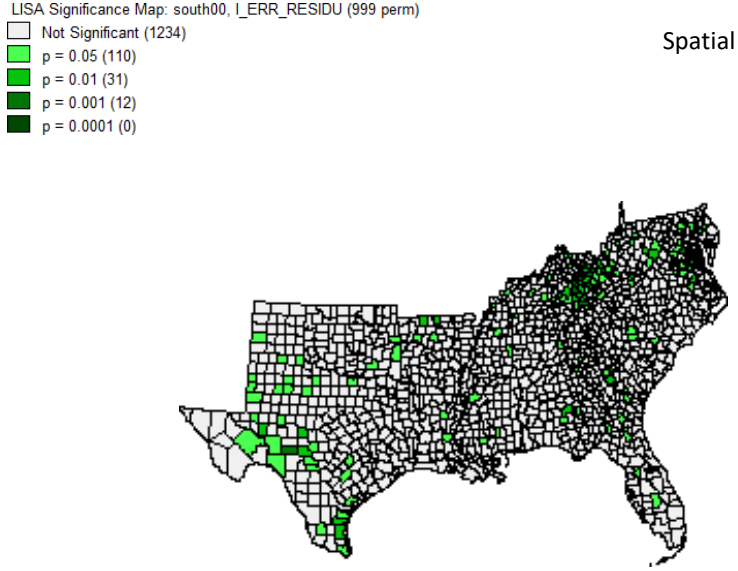
```
Regression
SUMMARY OF OUTPUT: SPATIAL ERROR MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : south00
Spatial Weight      : south00.gal
Dependent Variable  :    SQRTPPOV   Number of Observations: 1387
Mean dependent var  :    0,464095  Number of Variables   :    8
S.D. dependent var  :    0,096537  Degrees of Freedom    : 1379
Lag coeff. (Lambda) :    0,660223



R-squared           :    0,846531  R-squared (BUSE)      : -
Sq. Correlation     : -            Log likelihood        : 2504,640854
Sigma-square        :  0,00143024  Akaike info criterion :    -4993,28
S.E of regression   :   0,0378185  Schwarz criterion     :     -4951,4


------------------------------------------------------------------------
    Variable     Coefficient     Std.Error    z-value      Probability
------------------------------------------------------------------------
    CONSTANT      0,300464      0,0107058       28,06554    0,0000000
        PHSP     0,0992286      0,01622671       6,115139    0,0000000
        PFHH     0,6589877      0,02298421       28,67133    0,0000000
       PUNEM     0,8935399      0,06233104       14,33539    0,0000000
       PEXTR     0,3092319      0,02751795       11,23746    0,0000000
       P65UP      0,213199      0,03870122       5,508843    0,0000000
       METRO   -0,004556235      0,002838814      -1,604978    0,1084986
     PHSPLUS    -0,2476567      0,01297538      -19,08666    0,0000000
      LAMBDA     0,6602234      0,0259337       25,45813    0,0000000
------------------------------------------------------------------------


REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                                    DF      VALUE          PROB
Breusch-Pagan test                      7       443,7842       0,0000000


DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL ERROR DEPENDENCE FOR WEIGHT MATRIX : south00.gal
TEST                                    DF      VALUE          PROB
Likelihood Ratio Test                   1       361,9057       0,0000000
========================= END OF REPORT ==============================
```

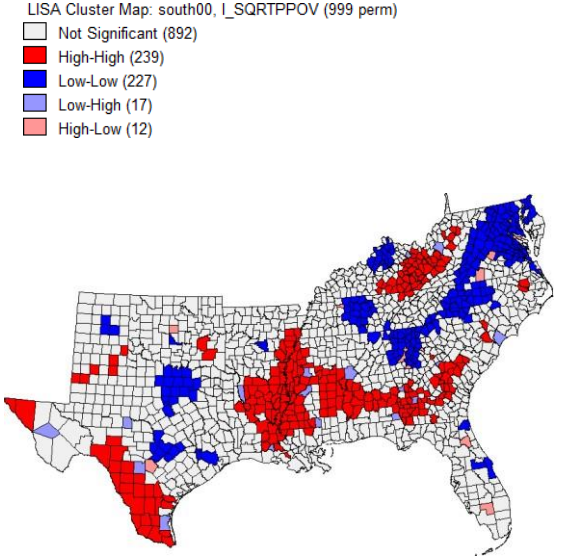


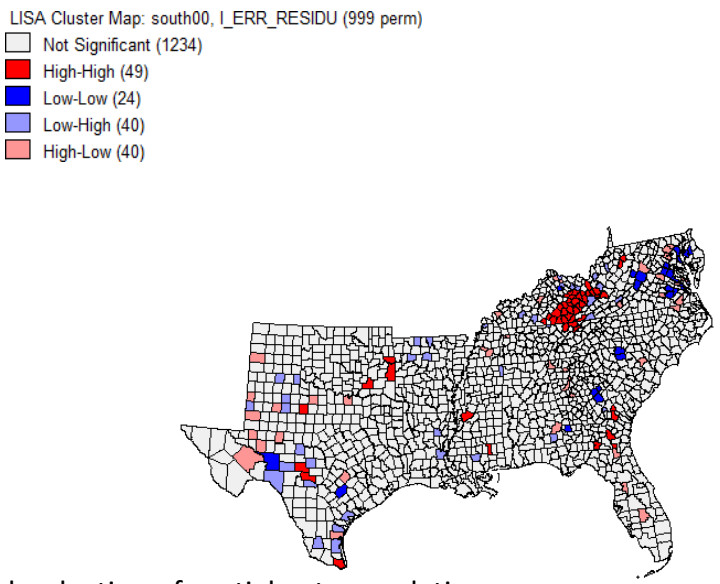
spatial error model

Model	R^2	Log Likelihood
OLS	0,780	2323,69
Spatial Lag	0,822	2457,37
Spatial Error	0,847	2504,64









The spatial error form results in a substantial reduction of spatial autocorrelation.

Part 2

Analyzing spatial heterogeneity with geographically weighted regression

- Traditional regression analysis describes a modelled relationship between a dependent variable and a set of independent variables. When applied to spatial data, the regression analysis often assumes that the modelled relationship is stationary over space and produces **a global model** which is supposed to describe the relationship at every location in the study area. This would be misleading, however, if relationships being modelled are intrinsically different across space. One of the spatial statistical methods that attempts to solve this problem and **explain local variation** in complex relationships is **Geographically Weighted Regression (GWR)**.

- In a global regression model, the dependent variable is often modelled as a linear combination of independent variables, where a parameter belonging to each variable is assumed to be stationary over the whole area (i.e. the model returns one value for each parameter). GWR extends this framework by dropping the stationarity assumption: the parameters are assumed to be continuous functions of location. The result of the GWR analysis is a set of continuous localised parameter estimate surfaces, which describe the geography of the parameter space. These estimates are usually mapped or analysed statistically to examine the plausibility of the stationarity assumption of the traditional regression and different possible causes of non-stationarity.

The definitive text on GWR is : Fotheringham, A.S., Brunsdon, C. & Charlton, M.E., *Geographically Weighted Regression : The Analysis of Spatially Varying Relationships*, Chichester, Wiley, 2002.

Differences between local and global statistics

Global	Local
Summarize data for whole region	Local disaggregation of global statistics
Single-valued statistic	Multi-values statistic
Non-mappable	Mappable
GIS unfriendly	GIS friendly
Aspatial or spatially limited	Spatial
Emphasize similarities across space	Emphasize differences across space
Search for regularities or 'laws'	Search for exceptions or local 'hot-spots'

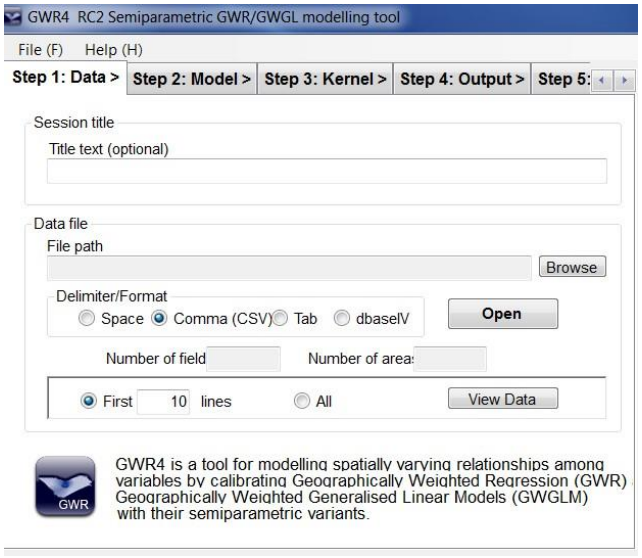
(*Source* : Fotheringham, Brunsdon and Charlton, 2002, pp. 6)

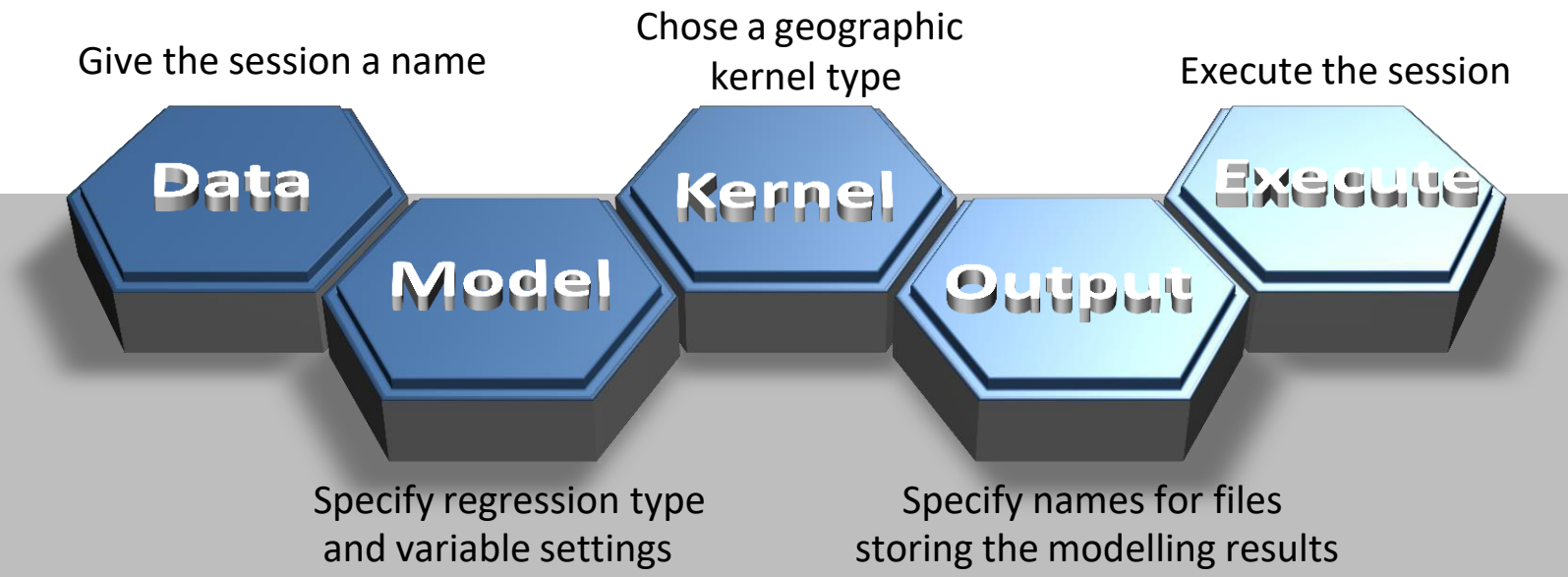
- The use of linear regression is common in many areas of science. Ordinary linear regression implicitly assumes spatial stationarity of the regression-model that is, the relationships between the variables remain constant over geographical space. We refer to a model in which the parameter estimates for every observation in the sample are identical as **a global model**.

- Spatial non-stationarity occurs when a relationship (or pattern) that applies in one region does not apply in another. Global models are statements about processes or patterns which are assumed to be stationary and as such are local independent, i.e. are assumed to apply to all locations. In contrast local models are spatial disaggregations of global models, the results of which are location-specific. The template of the model is the same : the model is a linear regression model with certain variables, but the coefficients alter geographically. If the parameter estimates are allowed to vary across the study area such that every observation has its own separate set of parameter estimates we have **a local model**.

- GWR does not assume the relationships between independent and dependent variables are constant across space. Instead, GWR explores whether the relationships between a set of predictors and an outcome vary by geographical location. GWR is suggested to be a powerful tool for investigating **spatial non-stationarity** in the relationship between predictors and the outcome variable.

- **GWR4** is new release of a Microsoft Windows based application for calibrating geographically weighted regression models, which can be used to explore geographically varying relationships between dependent/response variables and independent/explanatory variables.





For an extensive review of these 5 steps, see T. Nakaya, *GWR4 User Manual*, update 7 may 2012.

. An introduction to macro-level spatial nonstationarity : A geographically weighted regression analysis of diabetes and poverty

- Type II diabetes is a growing health problem. Because the burden of diabetes falls disproportionally on less advantaged individuals, poverty is one of the most important risk factors for diabetes.

- Micro-level (individual-level) research has consistently found positive associations between diabetes and poverty. Poverty and diabetes may be related because economic disadvantage may limit people to poorer diets and more sedentary lifestyles.

- Macro-level (context-level) investigations have also found a positive association between diabetes and poverty. Rates of diabetes are higher in areas with higher economic deprivation.

- What follows, provides a study of **the geographical variability in the relationship between poverty and diabetes**. We first show how a classical ordinary least squares regression captures the "global" and positive relationship between diabetes and poverty (an increase in the concentration in poverty is accompanied with an increase in the prevalence of diabetes). We then make use of an exploratory geographically weighted regression to specify a local modal. The findings reveal that the diabetes-poverty relationship macro-level relationship varies by geographical space

- Theoretically, spatial non-stationarity is based on the concept of **the social construction of space**. The interaction between individuals with each other and their physical environment produces space. Human beings are just as much spatial as temporal beings. By temporal, we mean that we are most influencedby what is immediate in space. What happens near us matters more than non-proximal events. Human's spatiality and temporality are essential and equal powerful in explaining human behavior. Consequently, everything that is social is inherently spatial, just as everything spatial is inherently socialized.

- From this perspective, we analyse how the macro-level relationship between diabetes and poverty unfolds over geographical space.

- Investigations on spatial non-stationarity focus on the phenomenon that two measurements taken from geographically close locations are often more similar than measurements from more widely separated locations (Tobler's law (1970, p. 236) : "Everything is related to everything else, but near things are more related than distant things").

- For this reason, spatial autocorrelation has been developed to deal with the tendency toward interdependence among spatial data. Investigating diabetes prevalence requires we expand our understanding of how macro-level relationships vary as a function of geographical distance.

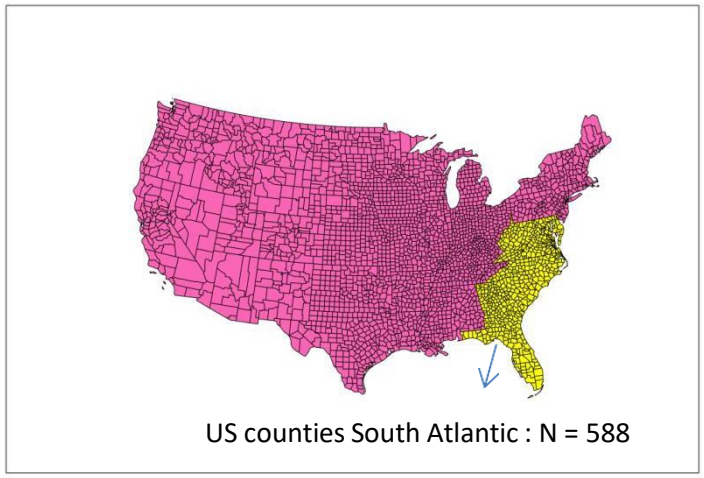
- In a global modal, we can hypothesize that poverty and diabetes are positively related. In a local modal, we can hypothesize that the diabetes-poverty macro-level relationship will be spatial non-stationary.

Tobler, W.R., A computer movie simulating urban growth in the Detroit region, *Economic Geography*, 46, 1970, pp . 234-240.

- Traditionally non-spatial research, including the OLS approach, assumes that the nature of statistical relationships is the same for all points within the entire study area. With GWR, we can explore how the diabetes-poverty relationship varies over space. The OLS results are thus for the "**global model**" findings while the GWR outputs are the "**local**" analyis results.

- We first execute an OLS multivariate regression to show the linear association between diabetes and poverty in US counties in the South Atlantic area (N=588)*. The goal of this "global model" is to verify the positive association found in previous studies. In the OLS model we use the percentage of diabetes in the county as the dependent variable and the percentage in poverty as the independent variable. We control the relationship between poverty and diabetes prevalence for median income of households and the percentage of people who completed high school. We then develop a GWR-model to account for spatial variations. The GWR model contains the same variables used in the OLS regression.

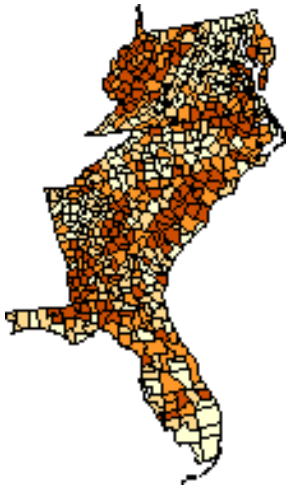
* Source : http://www.ers.usda.gov/data-products/county-level-data-sets.aspx
We focus on the 588 contiguous counties because GWR analysis requires that all polygons be physically adjacent or in near physical proximity to at least one other polygon with data on the variables of interest.



US counties South Atlantic : N = 588

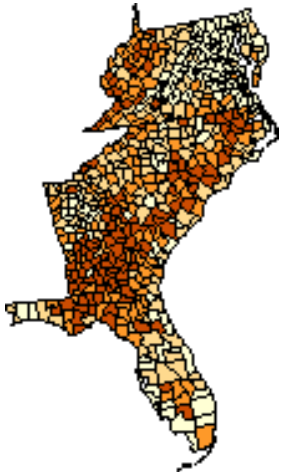
Quantile: pc_diabete

- [6,2:10,2] (142)
- [10,3:11,6] (163)
- [11,7:12,8] (137)
- [12,9:17,3] (146)



Quantile: pc_poverty

- [2:11,9] (146)
- [12:16,3] (148)
- [16,4:20,8] (148)
- [20,9:40,4] (146)



Global results

- Poverty is positively associated with diabetes. The results of OLS-model 1 demonstrate that an increase of one percentage point in the poverty concentration of a county is associated with a 0,15 percent increase in diabetes.

- Model 1 has an R^2 of 0,262. While diabetes prevalence and percent in poverty are statistically significantly related, a substantial proportion of the variation in diabetes prevalence remains unexplained.

- After adding median income of households and the percentage of people who completed high school to the regression equation, the effect of poverty is substantially reduced and no longer significant and even the sign of the coefficient for poverty changes from positive to negative. The R-square value for model 2 achieves a respectable 0,395*.

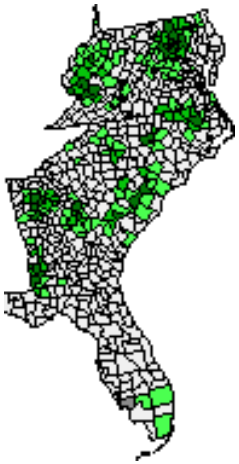
- We also note a problem : the regression equation shows strong spatial autocorrelation (Moran's I = 0,328 ; p < 001)**, a clear indication that the model is in violation with at least one of the assumptions underlying standard linear regression. The Moran test tells us that the residuals are not independent. Moreover, the Koenker-Bassett test for heteroscedasticity indicates that the residuals also are not distributed identically.

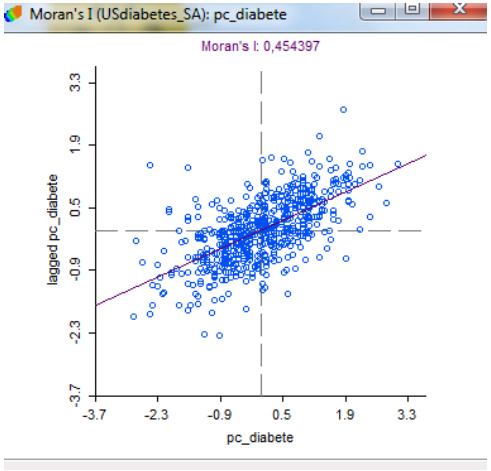
* Collinearity diagnostics were estimated using SPSS 20.0, and no problems of multicollinearity were found among the independent variables. The collinearity diagnostics used were the variance inflation factors (VIF) and tolerances for individual variables. Multicollinearity is said to exist if the VIF is 5 or higher (or equivalently, tolerances of 0,20 or less). The highest VIF in this analysis was 3,314 and the lowest tolerance was 0,302 for median income of households.

** Moran's I is strongly positive, indicating powerful positive autocorrelation (clustering of like values). LISA analysis demonstrates that most counties are found in the high-high and low-low quadrants.

LISA Significance Map: USdiabetes_SA, l_pc_diabete (999 perm)

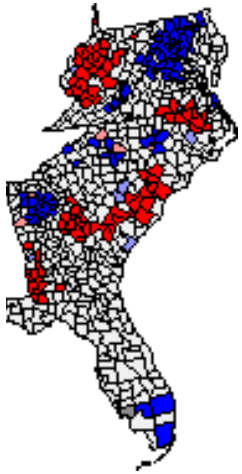
- Not Significant (411)
- p = 0.05 (101)
- p = 0.01 (42)
- p = 0.001 (29)
- p = 0.0001 (0)
- Neighborless (5)





LISA Cluster Map: USdiabetes_SA, l_pc_diabete (999 perm)

- Not Significant (411)
- High-High (83)
- Low-Low (74)
- Low-High (8)
- High-Low (7)
- Neighborless (5)



- Comparing the residual spatial autocorrelation (I = 0,328) with the spatial autocorrelation for the dependent variable (I = 0,454) tells us that spatial autocorrelation in one or more independent variables "explains" a portion of the spatial autocorrelation in the dependent variable*.

- It is frequently the case that the independent variables in a regression model can almost completely account for the spatial autocorrelation in a dependent variable, thus removing a problematic spatially autocorrelated residual. However, in the present case, the regressors have not satisfactorily accounted for spatial dependence in the data, and a correction to the model clearly is necessary. But what type of correction ? Might there be spillover effects among counties that influence the diabetes prevalence of their neighbors (spatial lag model) ? Or does the residual dependence in the model likely stem from omitted variables on the right-hand side of the regression equation, thus suggesting a spatial error model ?

* Moran's I is calculated by specifying a matrix of weights that characterizes the structure of local dependence. In this analysis "neighbors" are defined under the "first-order queen" convention, meaning that the neighbors for any given county "A" are those other counties that share a common boundary with "A" (or single point of contact with "A"). Importantly, "A" is not considered a neighbor of itself and is excluded from the average.

- We used a spatial regression model to control for the spatial autocorrelation. We chose which spatial dependence model to use (spatial lag or spatial error) using Lagrange Multiplier tests. Although both models exhibit significant spatial dependence, we used the model with the highest test statistic, in this case, the spatial error model.

- Aside from the remaining heteroscedasticity, the spatial error model appears to be a plausible alternative to the OLS specification. The AIC score is lower and the explanatory power of the model increases considerably over the OLS regression, with an R^2 of 0,538.

- In contrast with OLS-model 2, the effect of poverty on diabetes is statistically significant, independent from the median income of households and the percentage of people who completed high school.

- It is still not clear if spatial non-stationarity is a concern in our analysis. It is necessary to investigate the homoscedastic assumptions underlying the OLS with local modeling.

OLS and spatial regression models predicting the prevalence of diabetes in US South Atlantic counties (N=588)

	OLS (1)		OLS (2)		Spatial Error	
independent variables	coeff.	std.err.	coeff.	std.err.	coeff.	std.err.
constant	9,066 **	0,185	18,662 **		20,146 **	1,036
% poverty	0,151 **	0,010	-0,007	0,017	-0,040 *	0,016
median income of households			-0,000068 **	0,000008	-0,000077 **	0,000009
% completed high school			-0,051 **	0,012	-0,059 *	0,012
spatial error (Lambda)					0,530 **	
heteroscedasticity	30,240 **□		55,547 **□		48,399 **●	
R^2	0,262		0,395		0,538	
AIC	2233,690		2120,650		2002,780	

Lagrange Multiplier (Lag)	72,872 **
Robust LM (Lag)	1,642
Lagrange Multiplier (Error)	141,604 **
Robust LM (Error)	70,375 **

* p<0,05 ** p<0,01

□ Koenker-Bassettt test for heteroscedasticity

● Breusch-Pagan test for heteroscedasticity

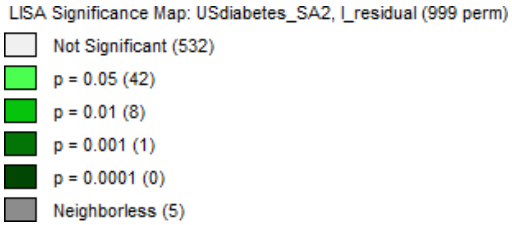
OLS models and the spatial error model are estimated by making use of Open GeoDa 1.2.0 (august 2012) ©Luc Anselin, 2011,2012

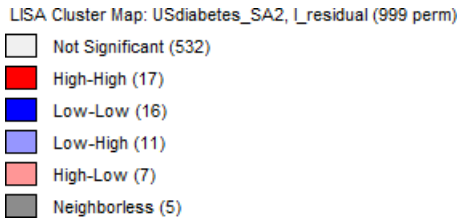
Local results

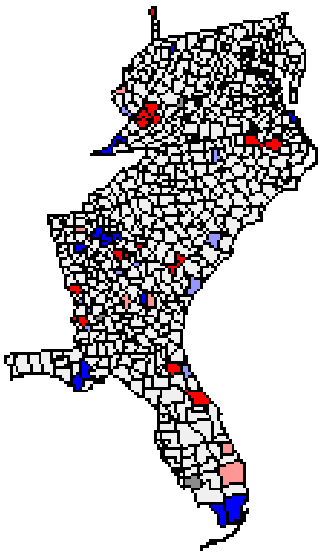
- In using spatial regression models we assume that the spatial process accounting for diabetes levels is the same across the study area. That is, the relationship is spatially stationary. However, few social processes will be found to be so constant over space. Global models will hide potential heterogeneity, or spatial non-stationarity, in the determinants of diabetes. GWR provides a method to access the degree to which the relationship between the potential determinants and the prevalence of diabetes varies across space.

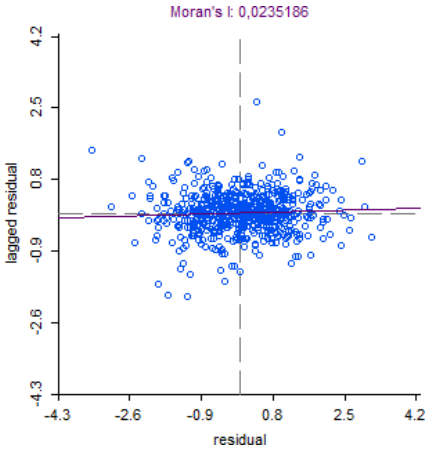
- The spatial non-stationarity of the relationship of each independent variable to the dependent variable can be assessed to determine whether the GWR method offers any improvement over a global regression model. The variability in the observed GWR estimates for the spatial units is compared to the variability of the GWR results from a large number of allocations of the analytical data across the units. Where one finds a significant difference between the variability of an observed estimate to those computed using the randomized data, spatial non-stationarity for that independent variable is indicated.

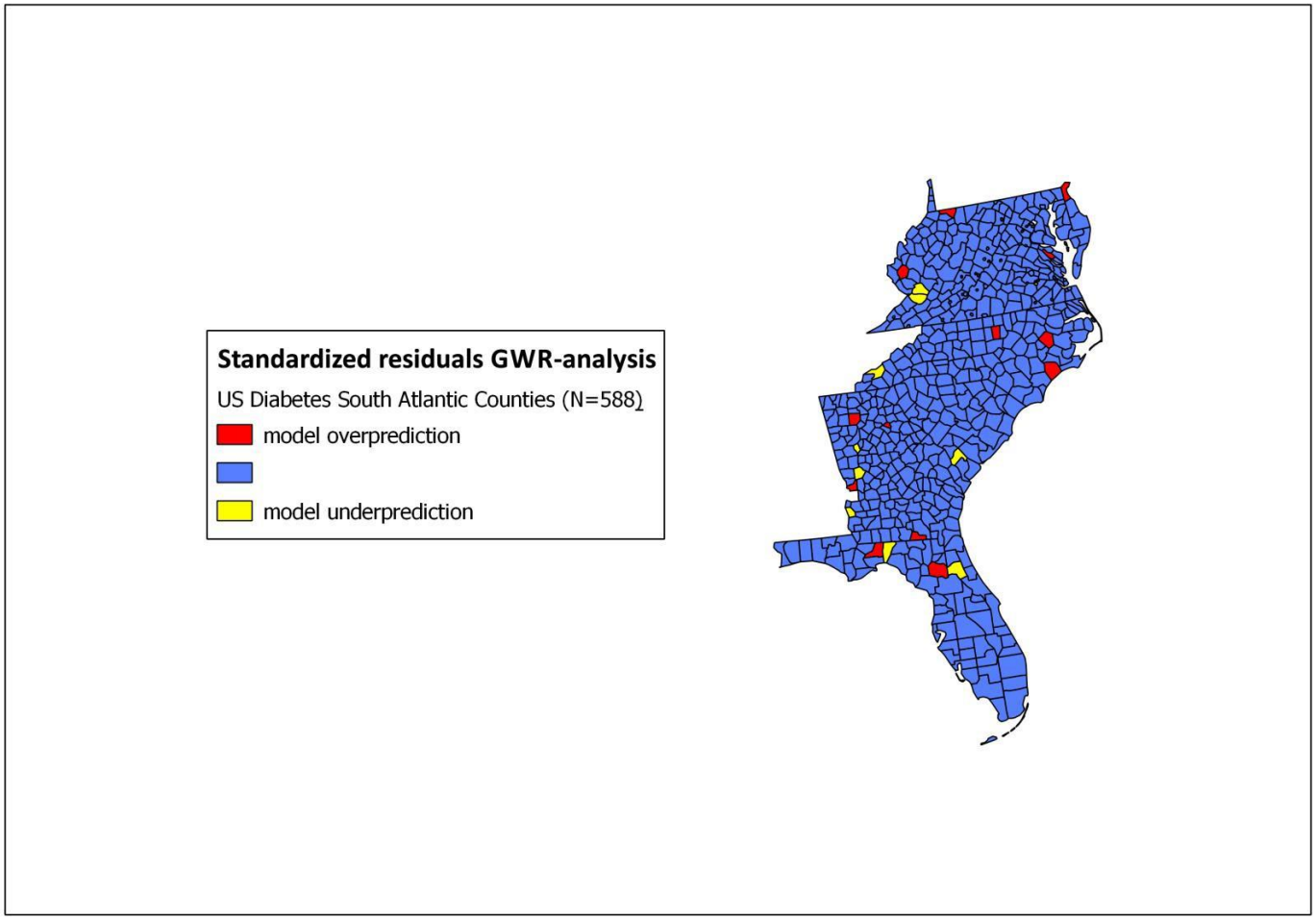
- We first made use of a local Moran's I cluster analysis of the residuals of the GWR model as a diagnostic for the collinearity of the GWR residuals. We found no violations of residual independence.







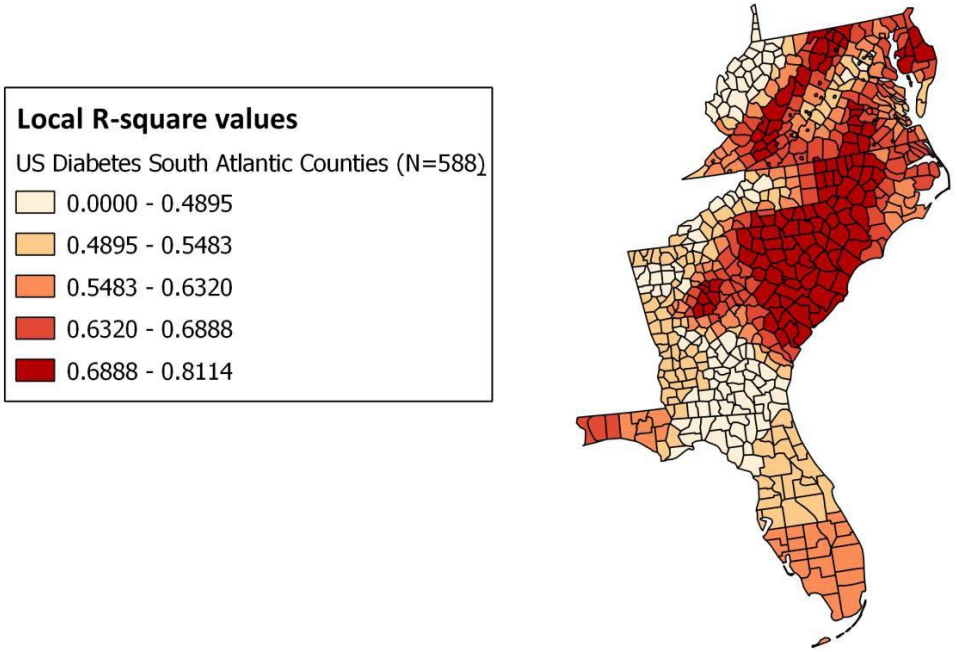



Standardized residuals GWR-analysis
US Diabetes South Atlantic Counties (N=588)
model overprediction
model underprediction

- The GWR results can best be summarized through the maps of the parameter estimates and the Monte Carlo tests. We provide maps of the local R^2 values and for each of the independent variables with a significant Monte Carlo test.

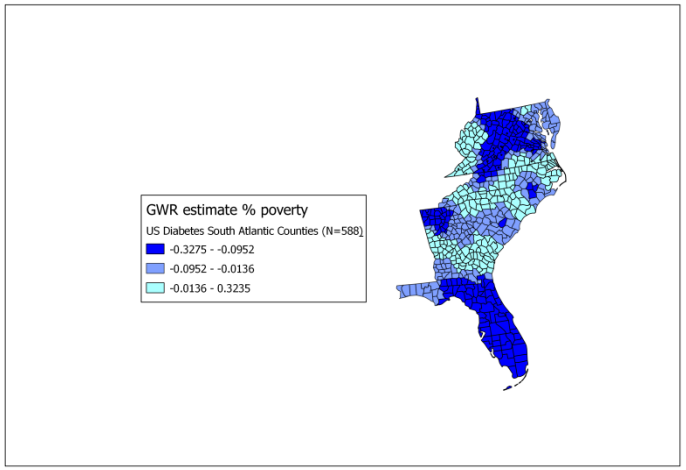
- The Monte Carlo tests for spatial variability of parameters indicate that the associations between the independent variables and diabetes are all non-stationary across space. Explicitly, the associations we found in OLS could not be generalized to anywhere in the South Atlantic region. In contrast to OLS, the GWR model explains 62,2 % of the total variance.

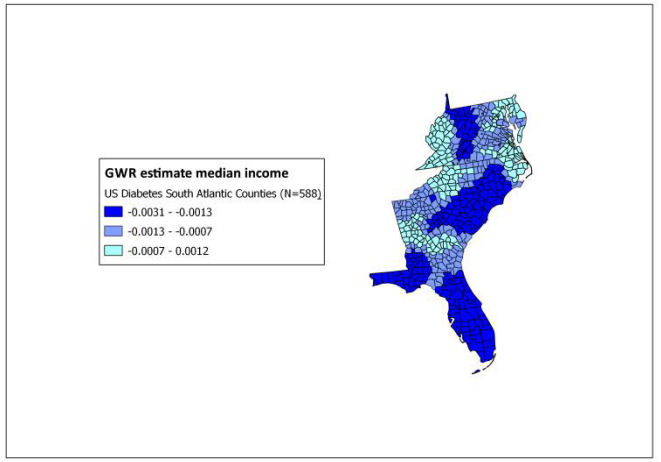
- As shown on the map of the local R^2 values, the total variance explained by the local model ranges from 16,1 % to 81,1%. The model fits the data well in the northern counties. Especially in the southern situated counties, there are areas that may benefit from a model with additional covariates. Herein lies the value of the GWR approach : without the ability to map the local R^2, we would not know where our model could be improved with additional covariates.

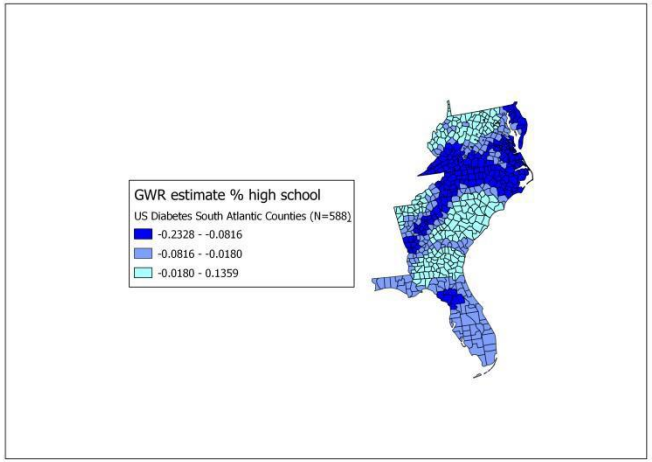

Local R-square values
US Diabetes South Atlantic Counties (N=588)
0.0000 - 0.4895
0.4895 - 0.5483
0.5483 - 0.6320
0.6320 - 0.6888
0.6888 - 0.8114

- The model results of the GWR can be interpreted in two ways. Those interested in a particular area can use the model results for that place to get a multivariate understanding of key local determinants of the diabetes prevalence. We will not do this here. An alternative way to examine the results is by considering for each determinant the varying nature across the counties of the South Atlantic region.

- For example, the GWR coefficient for the percentage of poverty ranges from -0,33 to 0,32 which signals that the poverty-diabetes macro-level association is spatially non-stationary. The blue marked counties indicate areas where an increase in poverty predicts lower diabetes prevalence. The shift to light-blue marked areas captures the spatially non-stationary relationship between poverty and diabetes. The poverty-diabetes relationship fluctuates from negative to positive as a function of geographical location. Similar results exist for the relationship between median household income, resp. educational attainment and diabetes. In short, after accounting for location, we find that macro-level associations between predictor variables and diabetes fluctuate as a function of geography.


GWR estimate % poverty
US Diabetes South Atlantic Counties (N=588)
-0.3275 - -0.0952
-0.0952 - -0.0136
-0.0136 - 0.3235


GWR estimate median income
US Diabetes South Atlantic Counties (N=588)
-0.0031 - -0.0013
-0.0013 - -0.0007
-0.0007 - 0.0012


GWR estimate % high school
US Diabetes South Atlantic Counties (N=588)
-0.2328 - -0.0816
-0.0816 - -0.0180
-0.0180 - 0.1359

- The previous analysis demonstrates that GWR addresses the need for place-specific or place-sensitive forms of analysis.

- Effective locational decision making is essential for properly addressing many socio-economic, demographic and health related concerns. Presently, these decisions are supported by quantitative models, which are potentially powerful tools, but whose estimates are often affected by uncertainty, which reduces their reliability.

- Uncertainty in the model parameters stems from two proporties of geographical phenomena :
  - spatial dependence : near things are more related than distant things ;
  - spatial non-stationarity : variability over space ;

- These two properties are mutually related, and most observed processes exhibit both, simultaneously.

- Advanced spatial analytical methods exist to correct for the effects of each property. However, despite the recognized simultaneity of their occurrence, each advanced spatial method is designed to address only one property. Spatial autoregressive methods address spatial dependence but do not account for non-stationarity ; geographically weighted regression addresses non-stationarity but does not account for spatial dependence.