

# 07 and 08 - Spatial Autocorrelation– Global and Local

---

J.S. Onésimo Sandoval  
Sociology  
Saint Louis University

# Outline

- Spatial Statistics for Spatial Data
- Introduction to Spatial Autocorrelation
- Importance of Spatial Autocorrelation
- Measuring Spatial Autocorrelation
- Conceptualization of Spatial Relationship
- Null Hypothesis and Spatial Autocorrelation
- Global Spatial Autocorrelation Lab

# Spatial Statistics for Spatial Data

---

# Spatial Statistics for Spatial Data

- Spatial statistics play an ever increasing role
- Immense methodological diversity
- “How Much” vs. “How Much is Where”
- One of the key features of spatial data is the autocorrelation of observations in space
- Classical statistical modeling vs. Spatial statistical modeling
- Spatial Patterns

# Introduction to Spatial Autocorrelation

---

# Spatial Autocorrelation

*Many ways to define it!*

1. *Tobler's first law of geography*

2. *Similarity*

3. *Probability*

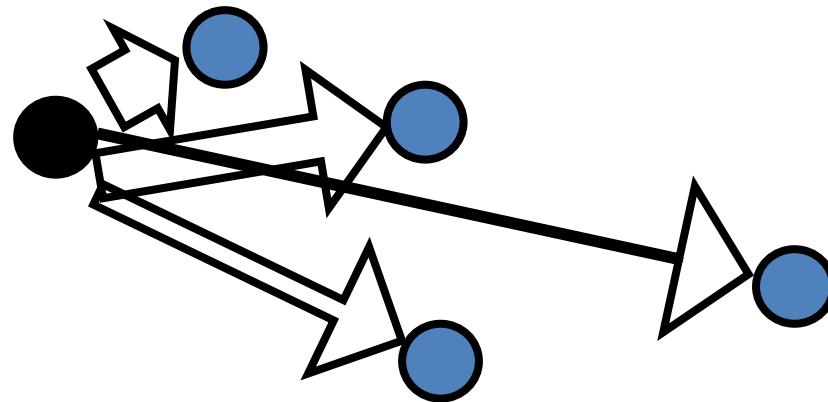
4. *Correlation*

# Spatial Autocorrelation

## 1. Tobler's Law

The confirmation of Tobler's first law of geography\*:

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



*The single most important concept in spatial statistics!*

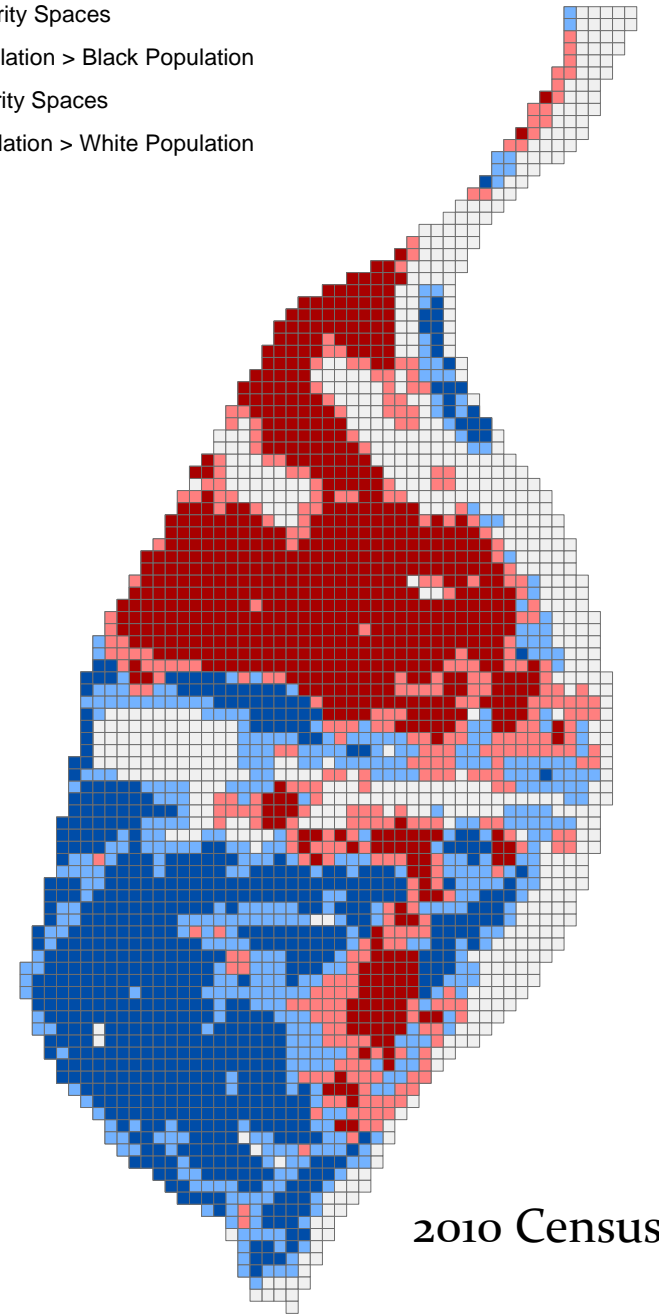
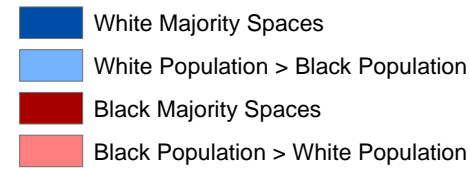
\*Tobler W., (1970) "A computer movie simulating urban growth in the Detroit region". *Economic Geography*, 46(2): 234-240

**Spatial:**  
On a map

**Auto:**  
Self

**Correlation:**  
Degree of  
relative  
similarity

**Saint Louis Color Line**





# Spatial Autocorrelation

## 2. Similarity

<u>Positive spatial autocorrelation</u>	<u>Negative spatial autocorrelation</u>
Similar to	Different from (dissimilar)
- high values surrounded by nearby high values	- high values surrounded by nearby low values
- intermediate values surrounded by nearby intermediate values	- intermediate values surrounded by nearby intermediate values
- low values surrounded by nearby low values	- low values surrounded by nearby high values
- clustering	- checker board pattern
<p><u>Positive</u> spatial autocorrelation much more common than <u>Negative</u> spatial autocorrelation</p>	

# Spatial Autocorrelation

A. Completely separated pattern (+ve)						B. Evenly spaced pattern (-ve)						C. Random pattern					
1	1	1	0	0	0	1	0	1	0	1	0	0	0	1	1	0	1
1	1	1	0	0	0	0	1	0	1	0	1	0	1	0	1	0	0
1	1	1	0	0	0	1	0	1	0	1	0	0	0	0	0	0	1
1	1	1	0	0	0	0	1	0	1	0	1	0	1	1	1	0	0
1	1	1	0	0	0	1	0	1	0	1	0	0	0	1	1	0	1
1	1	1	0	0	0	0	1	0	1	0	1	0	1	1	1	0	1

positive

negative

No SA

# Spatial Autocorrelation

## 3. Probability

- Measure of the extent to which the occurrence of an event in one geographic unit (polygon) makes more probable, or less probable, the occurrence of a similar event in a neighboring unit.

high negative spatial  
autocorrelation

no spatial  
autocorrelation

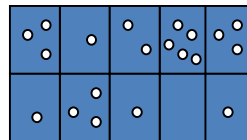
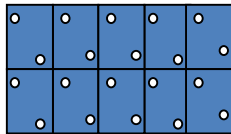
high positive spatial  
autocorrelation

Dispersed Pattern

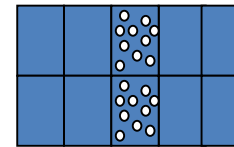
Random Pattern

Clustered Pattern

UNIFORM/  
DISPERSED



CLUSTERED



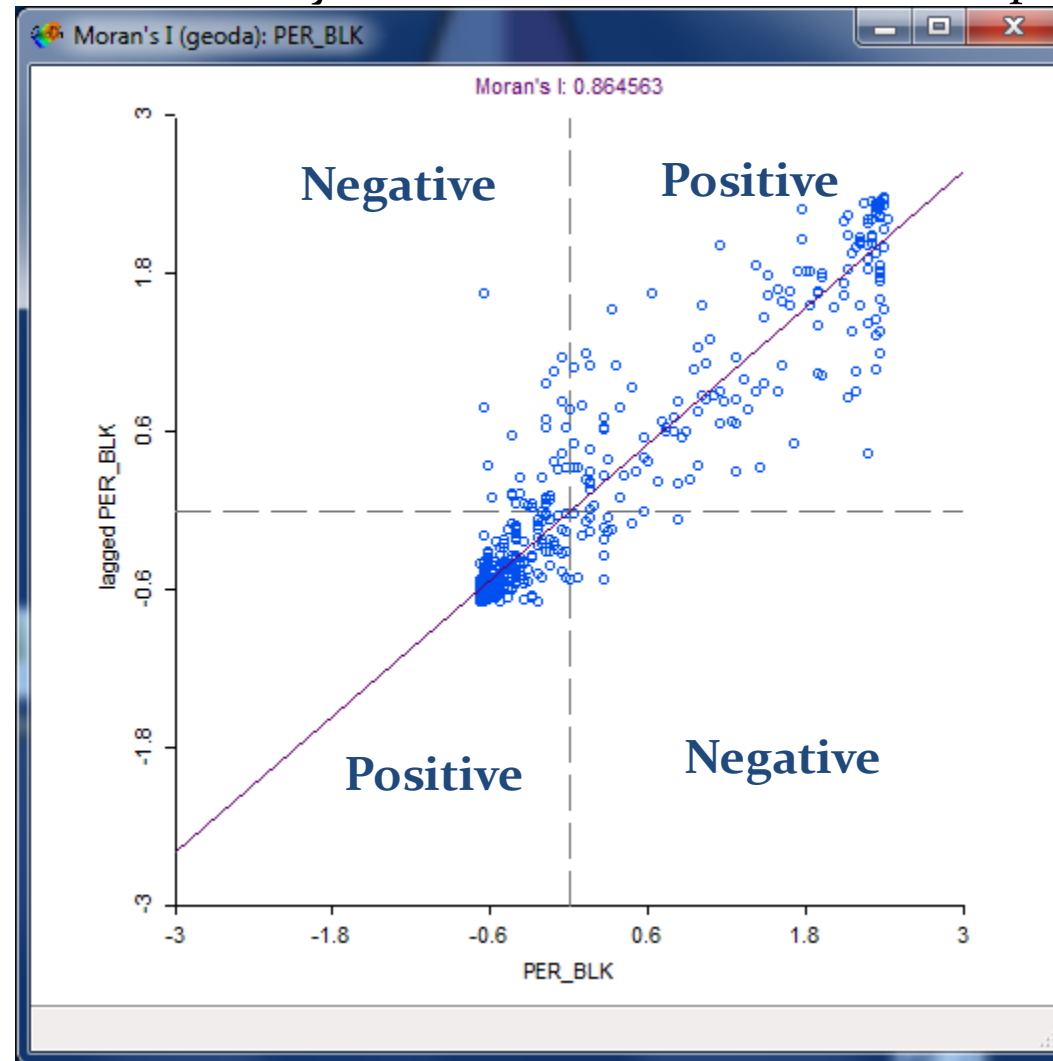
# Spatial Autocorrelation

## 4. *Correlation*

- a) Let's say  $Z(s)$  is the attribute  $Z$  observed in a certain spatial location.
  - Thus, we can define Spatial Autocorrelation as the correlation between  $Z(s_i)$  and  $Z(s_j)$ .
  - In other words, it is the correlation between the same attribute at two locations.
- b) In Layman's terms - Spatial Autocorrelation is the correlation of a variable with itself through space

# Spatial Autocorrelation

*Correlation* = “similarity”, “association”, or “relationship”



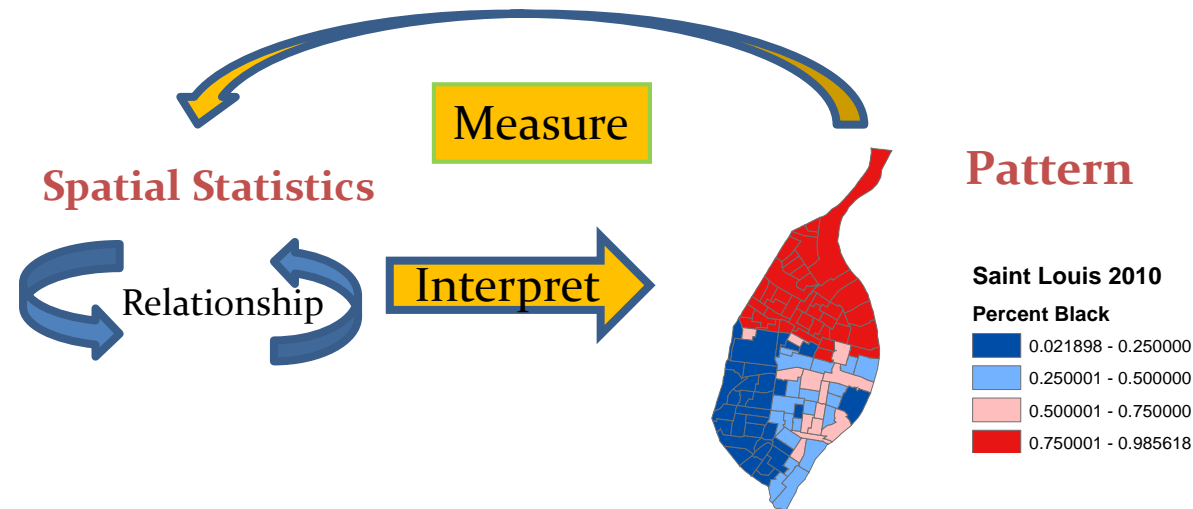
# Importance of Spatial Autocorrelation

---

# Why is Spatial Autocorrelation Important?

## Two reasons

1. Spatial autocorrelation is important because it implies the existence of a spatial process
  - Why are nearby areas similar to each other?
  - Why do high income people live “next door” to each other?
  - Why does racial segregation persist?
    - These are GEOGRAPHICAL questions. They are about location



# Why is Spatial Autocorrelation Important?

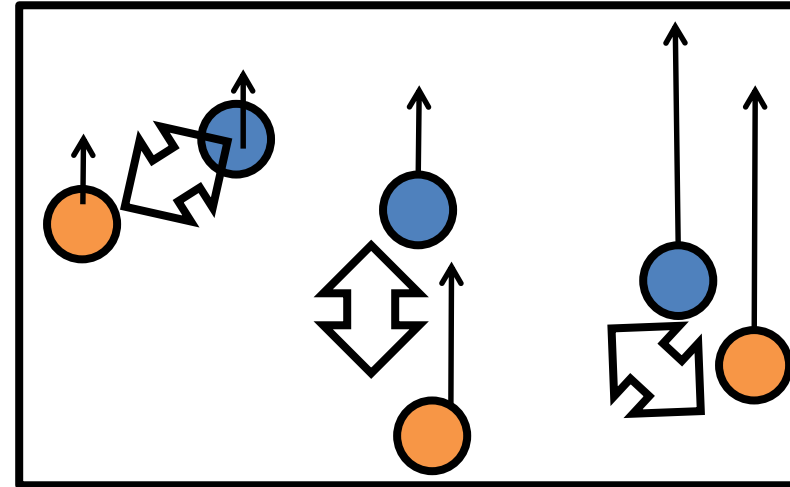
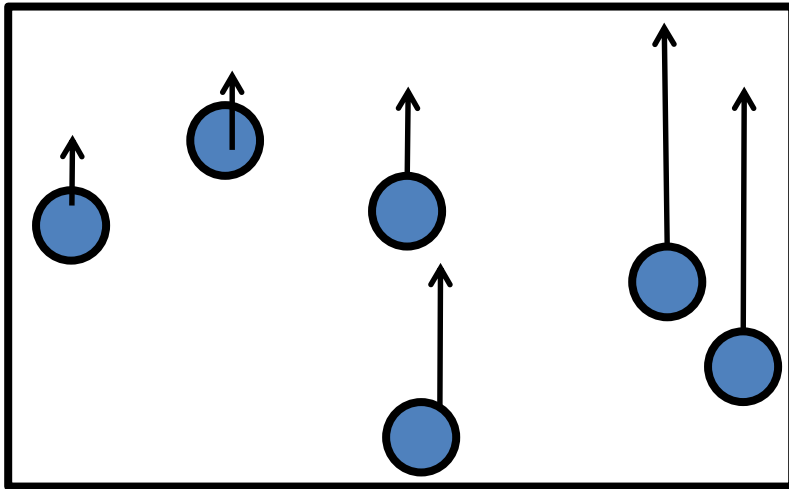
## Two reasons

2. It invalidates most traditional statistical inference tests
  - If SA exists, then the results of standard statistical inference tests may be incorrect (wrong!)
    - We increase the likelihood to make Type I error.
  - We need to use spatial statistics



# Why are standard statistical tests wrong?

- Statistical tests are based on the *assumption* that the values of observations in each sample are independent of one another
- Spatial Autocorrelation violates this
  - samples taken from nearby areas are related to each other and are not independent



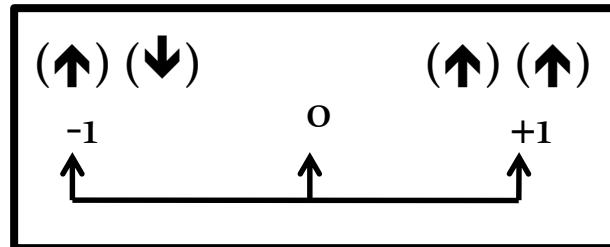
Implies a relationship between nearby observations

# Why are standard statistical tests wrong?

## Example for the *correlation coefficient* ( $r$ )

What is the *correlation coefficient* ( $r$ )?

- The most common statistic in all of science
- Measures the strength of the relationship (or “association”) between two variables e.g. income and education
- Varies on a scale from  $-1$  thru  $0$  to  $+1$ 
  - +1 implies a perfect positive association
    - As values go up (↑) on one, they also go up (↑) on the other
      - income and education
  - 0 implies no association
  - 1 implies perfect negative association
    - As values go up on one (↑), they go down (↓) on the other
      - price and quantity purchased
- Full name is the *Pearson Product Moment correlation coefficient*



# Why are standard statistical tests wrong?

Example for the *correlation coefficient* ( $r$ )

**If Spatial Autocorrelation exists:**

1. Correlation coefficients appear to be bigger than they really are
2. They are more likely to be found “statistically significant”

***We make two mistakes:***

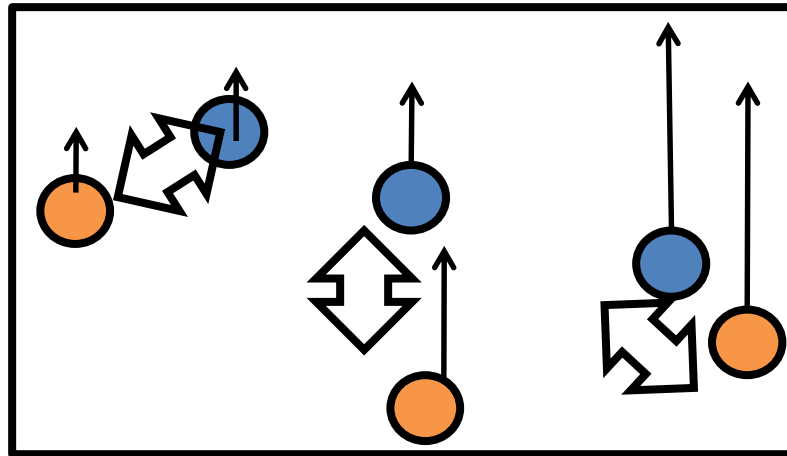
- *We believe that the relationship is stronger than it really is*
- *We are more likely to incorrectly conclude a relationship exists when it does not*

# Why are standard statistical tests wrong?

Example for the *correlation coefficient* ( $r$ )

**If Spatial Autocorrelation exists:**

- Correlation coefficients are bigger than they really are:
  - Because income and education are similar in nearby areas
  - Correlation coefficient is “biased upward”

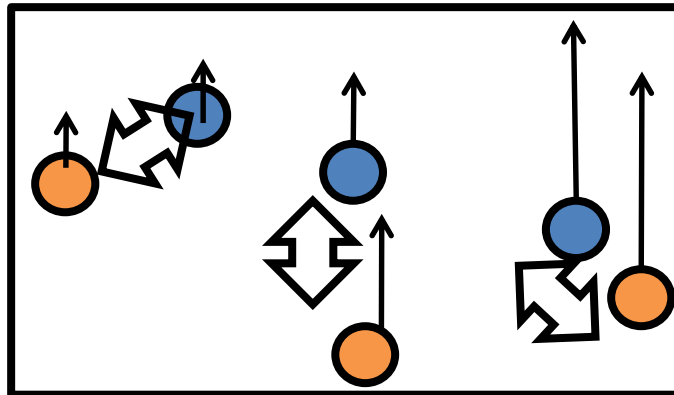


# Why are standard statistical tests wrong?

## Example for the *correlation coefficient* ( $r$ )

### If Spatial Autocorrelation exists:

- More likely to appear “statistically significant”
  - Standard error is smaller because spatial autocorrelation “artificially” reduces variability
- There is actually more variability than it appears
  - “exaggerated precision”



# Null Hypothesis and Spatial Autocorrelation

---

# How to make a decision about Spatial Autocorrelation

- Null Hypothesis
  - Observed spatial pattern of value is equally likely as any other spatial pattern
  - Value at one locations does not depend on values at other locations
  - Under spatial randomness, the location of value may be altered without affecting the information contained in the data.
- $H_0$ =There is no spatial clustering of the values associated with the geographic features in the study area
- $H_1$ = There is spatial clustering of the values associated with the geographic features in the study area

# Steps for Analysis for Spatial Autocorrelation

- Step One – Define the variable for consideration
  - (Typically start with the dependent variable and move to the independent variables and make assumptions)
    - e.g., interval-ratio variable
    - e.g., independent samples
- Step Two – State the Hypothesis
  - $H_0$  = No Spatial Autocorrelation (or  $H_0=0$ )
  - $H_1$  = Spatial Autocorrelation is present (or  $H_1$ =Moran's value)
- Step Three – Establish the Critical Region and Compute your test statistics (Moran's I – edges and corners)
  - Alpha .05, .01, and .001



# Moran's I

---

# Moran's I

- One of the oldest indicators of spatial autocorrelation (Moran, 1950). Still a de facto standard for determining spatial autocorrelation.
- Applied to zones or points with continuous variables associated with them.
- Compares the value of the variable at any one location with the value at all other locations

# Moran's I

- $W_{ij}$  is a contiguity matrix
  - If zone  $j$  is adjacent to zone  $i$ , the interaction receives a weight of 1
  - Another option is to make  $W_{ij}$  a distance-based weight which is the inverse distance between locations  $i$  and  $j$  ( $1/d_{ij}$ )
  - Compares the sum of the cross-products of values at different locations, two at a time weighted by the inverse of the distance between the locations
- Similar to correlation coefficient, it varies between  $-1.0$  and  $+1.0$ 
  - When autocorrelation is high, the coefficient is high
  - A high  $I$  value indicates positive autocorrelation

# Moran's I

$$I = \frac{n}{p} * \frac{\sum_i \sum_j w_{ij} (z_i - \bar{z})(z_j - \bar{z})}{\sum_i (z_i - \bar{z})^2}$$

where

$$p = \sum_i \sum_j w_{ij}$$

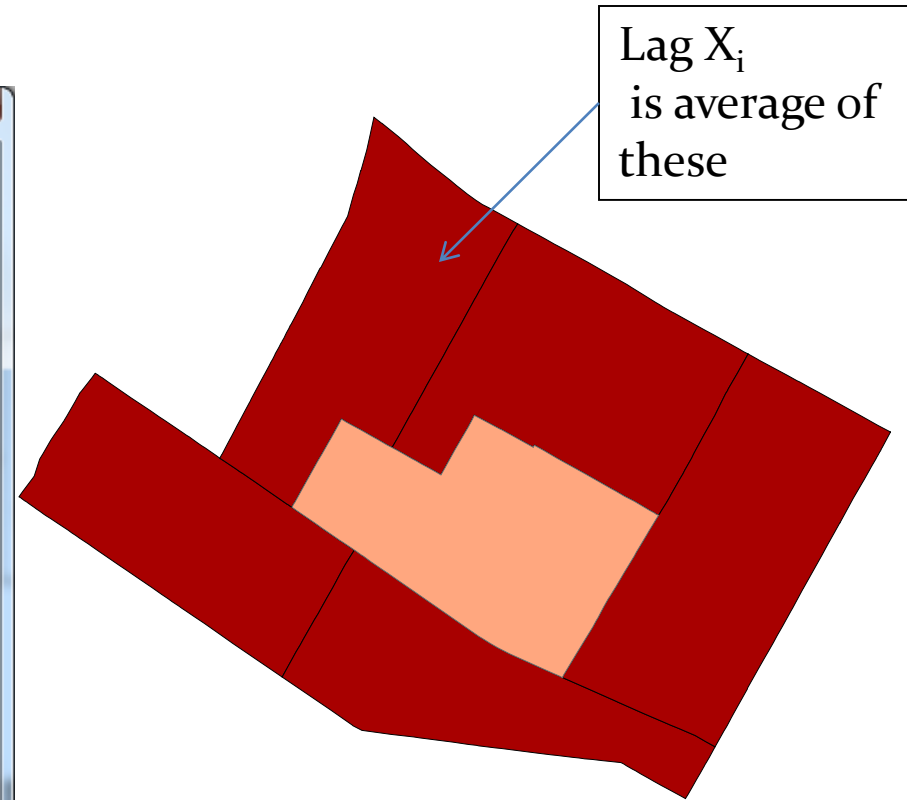
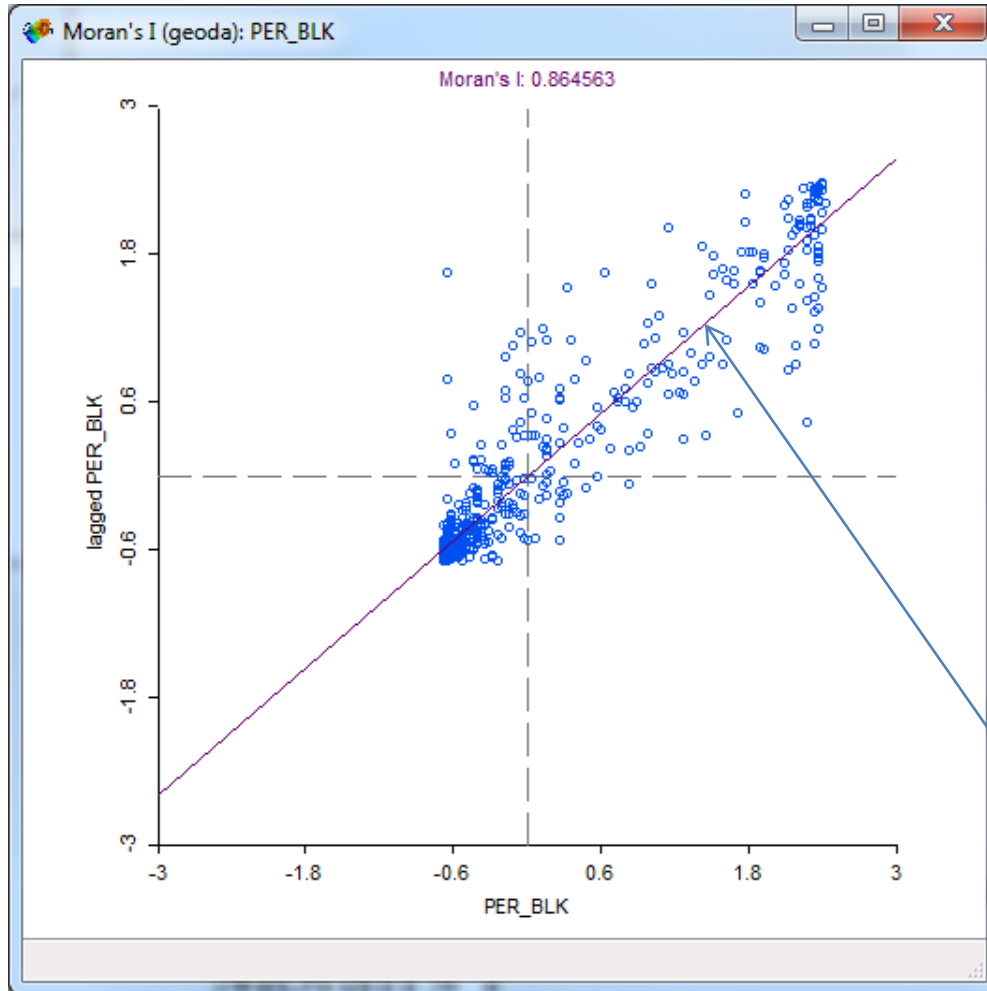
$w_{ij}$  = spatial matrix

# Moran Scatter Plots

Moran's  $I$  can be interpreted as the correlation between variable,  $X$ , and the “spatial lag” of  $X$  formed by averaging all the values of  $X$  for the neighboring polygons

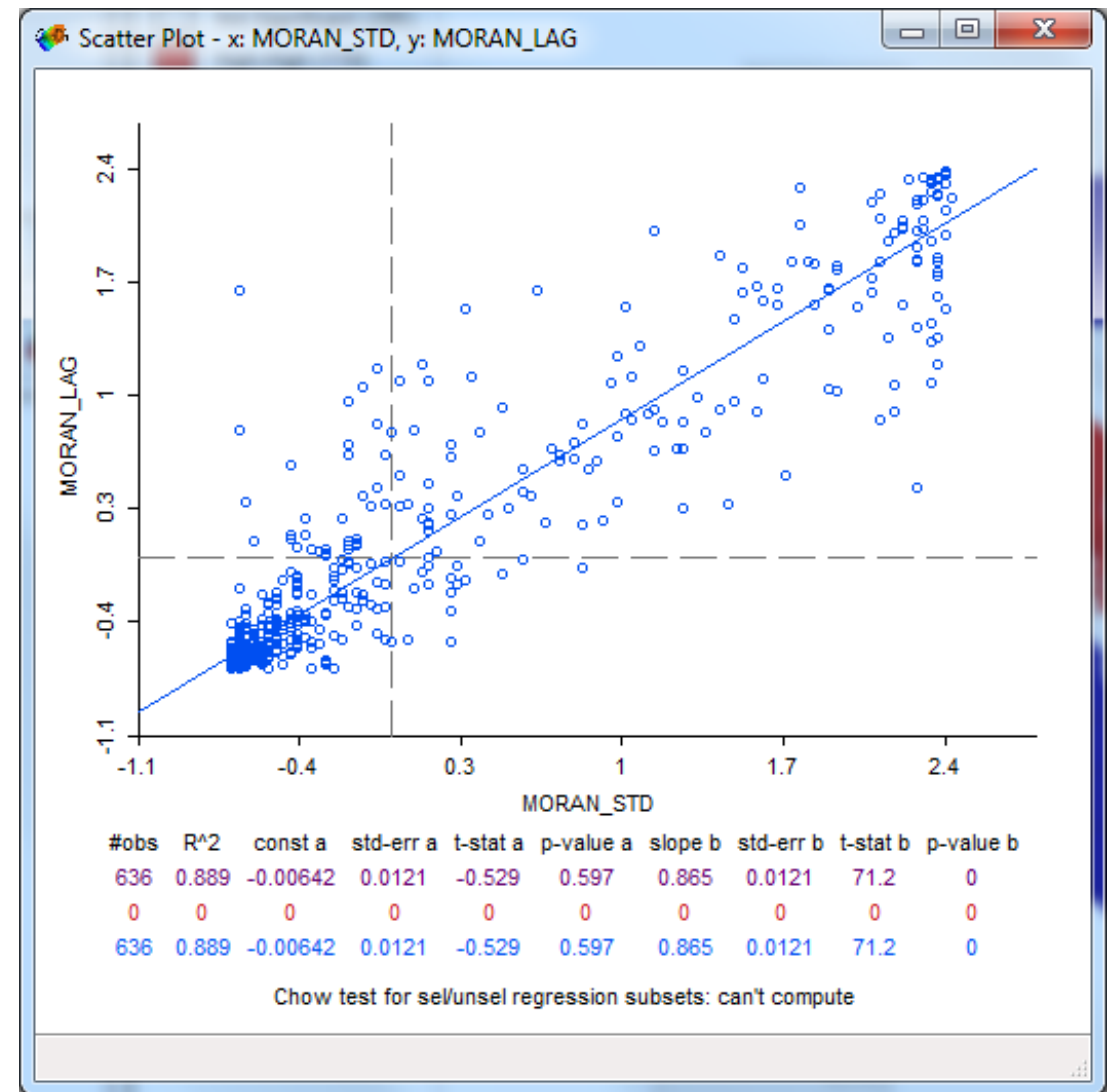
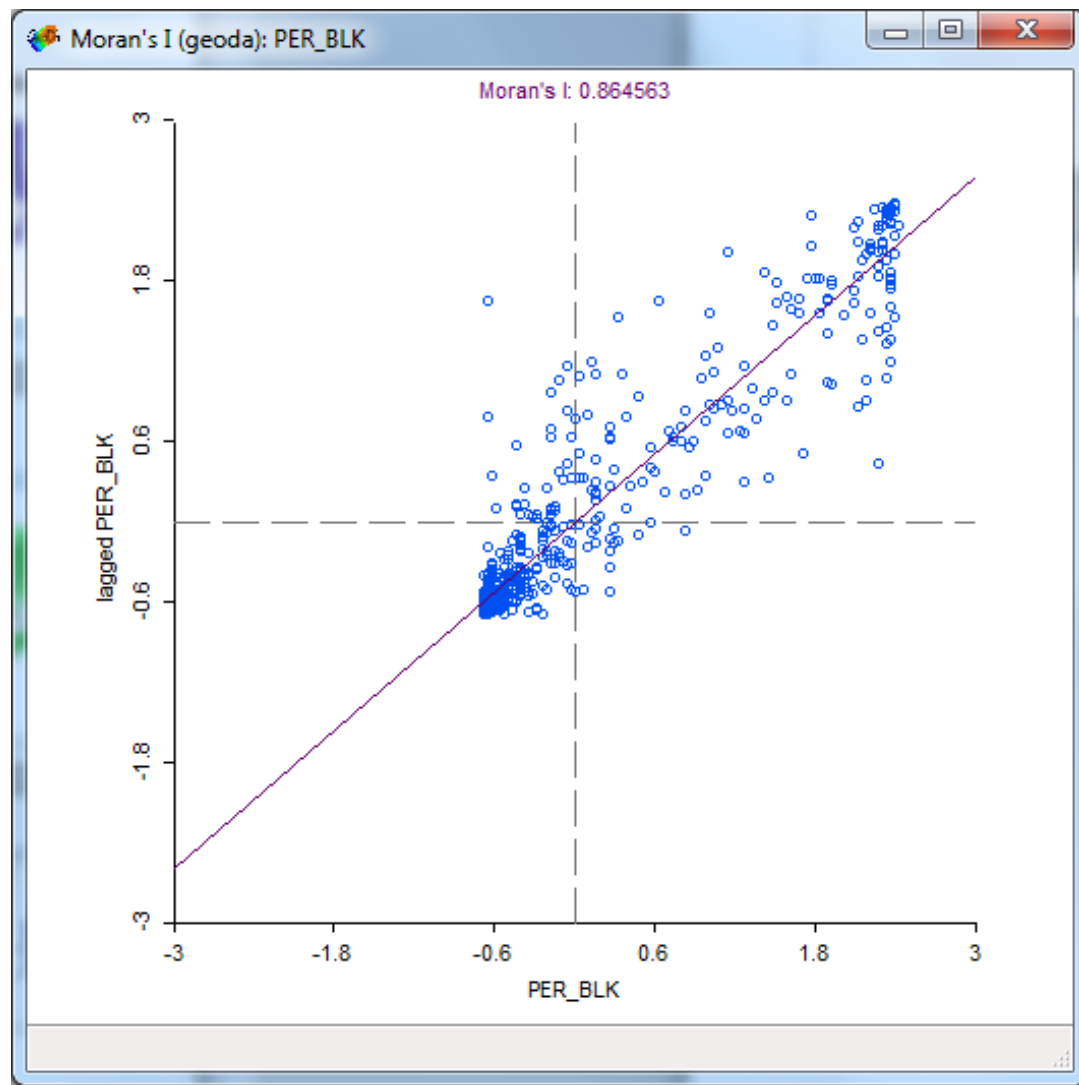
We can then draw a scatter diagram between these two variables (in standardized form):  $X$  and  $\text{lag-}X$  (or  $W_X$ )

# Example – Percent Black



Lag  $X_i$   
is average of  
these

Least squares "best fit" line to  
the points.  
The slope of this *regression line*  
is Moran's I



The Moran's I scatter plot regresses a spatially lagged transformation of a variable (y-axis) on the original standardized variable (x-axis). The values of X are standardized in standard deviation units with a mean of zero

# Moran's I for rate-based data

- Moran's I is often calculated for rates, such as crime rates (e.g. number of crimes per 1,000 population) or infant mortality rates (e.g. number of deaths per 1,000 births)
- An adjustment should be made, especially if the denominator in the rate (population or number of births) varies greatly (as it usually does)
- Adjustment is known as the *EB adjustment*:
  - see Assuncao-Reis *Empirical Bayes Standardization* [Statistics in Medicine](#), 1999
- *GeoDA* software includes an option for this adjustment



# Global vs. Local Spatial Autocorrelation

---

Global	Local
Summarize Data for the whole region	Local disaggregation of global statistics
Single-valued statistic	Multi-valued statistics – different values can occur in different locations.
Non-mappable	Mappable
GIS-unfriendly	GIS-friendly – show how the relationship vary over space. – our goal is to map these relationships
Aspatial or spatially limited	Spatial
Emphasize similarities across space	Emphasize differences across space
Search for regularities or “laws” – can be represented by one statistics – testing hypothesis – positivist school of thought	Search for exceptions or local “hot spots” – developing hypothesis from the data – grounded theory – exploratory spatial analyses.
Classic Regression	Spatial Regression or Geographically Weighted Regression

# Calculating Anselin's LISA

- The Local Moran statistic for areal unit  $i$  is:

$$I_i = z_i \sum_{j=1}^{J_i} W_{ij} Z_j$$

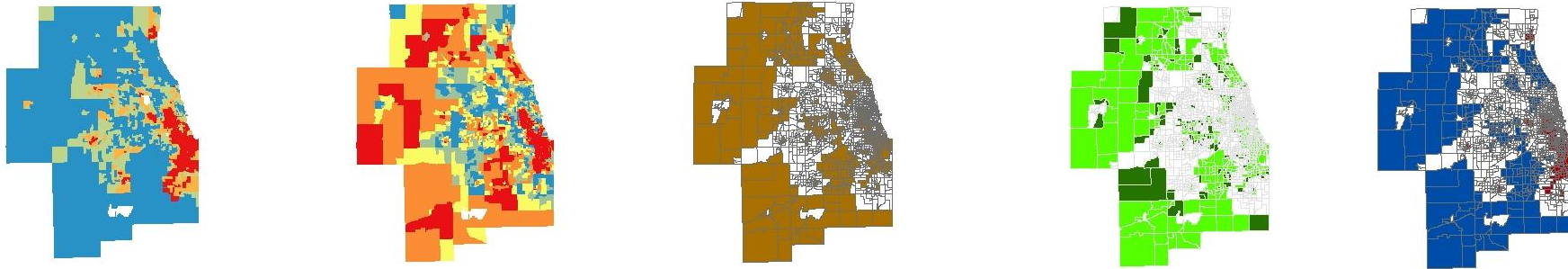
Where:

$z_i, z_j$  are standardized values

$w_{ij}$  is generally the row-standardized spatial weight matrix

The summation  $\sum_j$  is across each row  $i$  of the spatial weights matrix.

# Local Indicator of Spatial Association (LISA)



# Lab - ArcMap

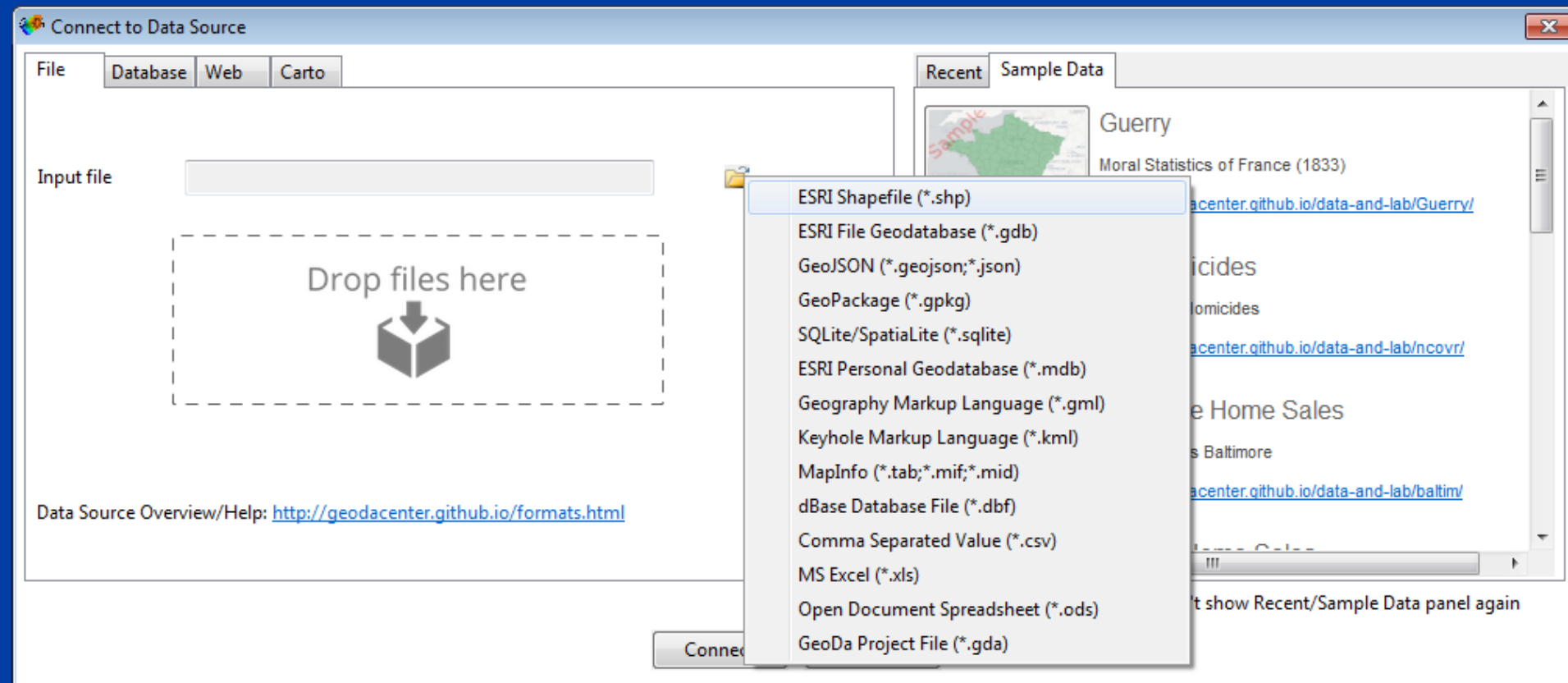
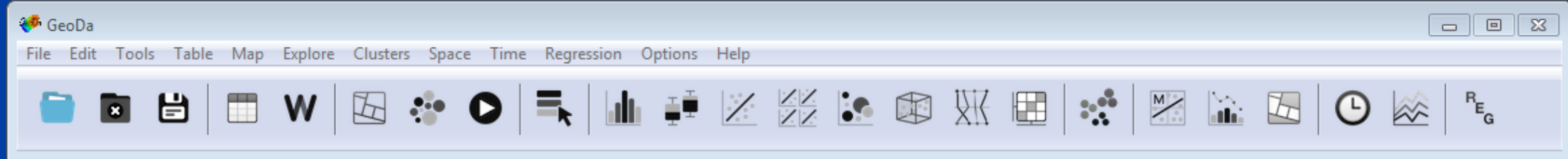
---

# Global Spatial Autocorrelation Lab

## GeoDa

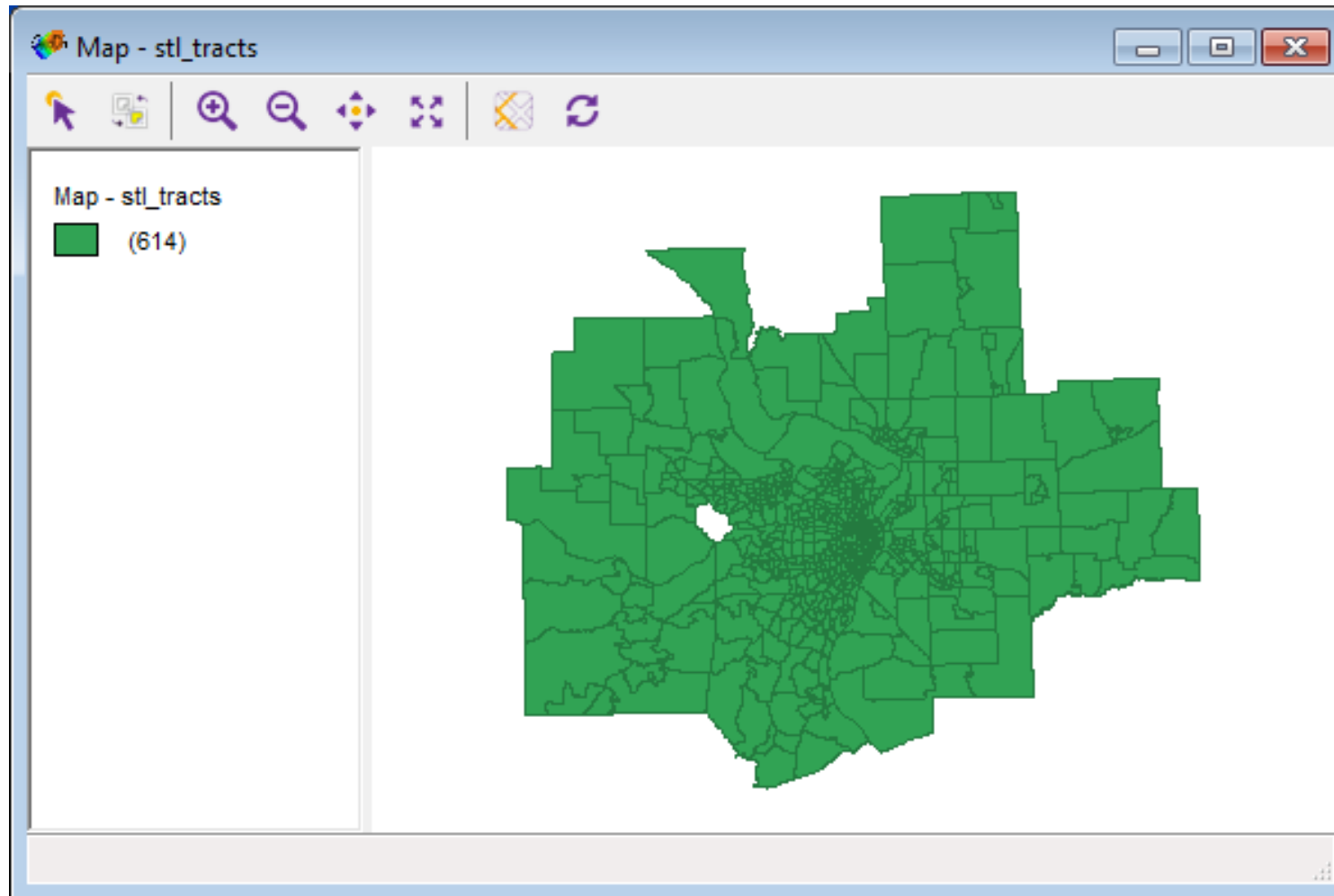
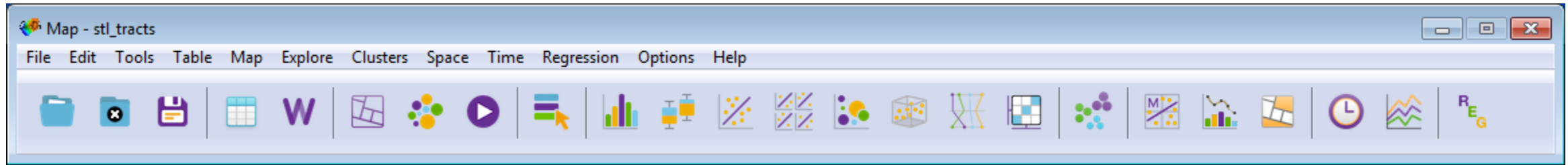
---

# (1) Create a new shapefile and open it in GeoDa



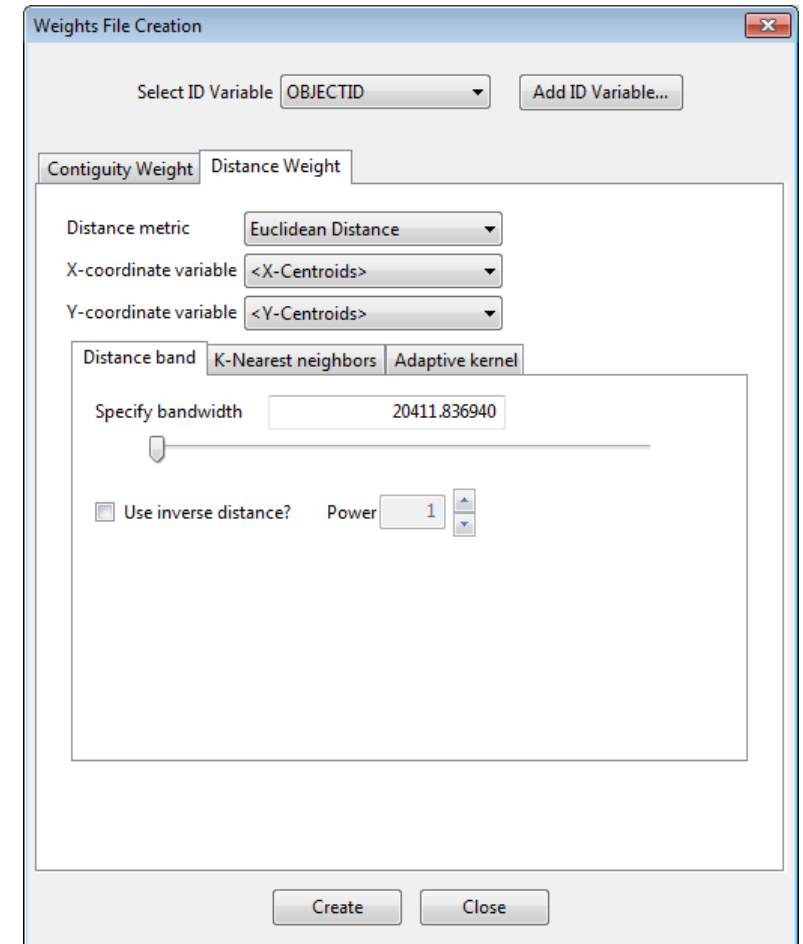
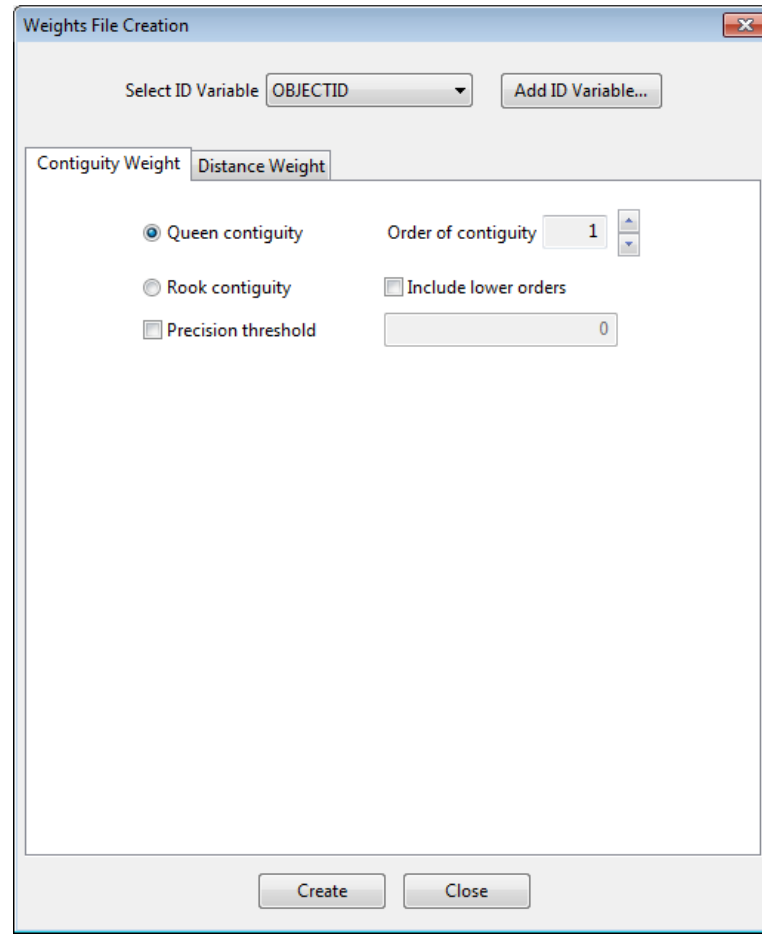
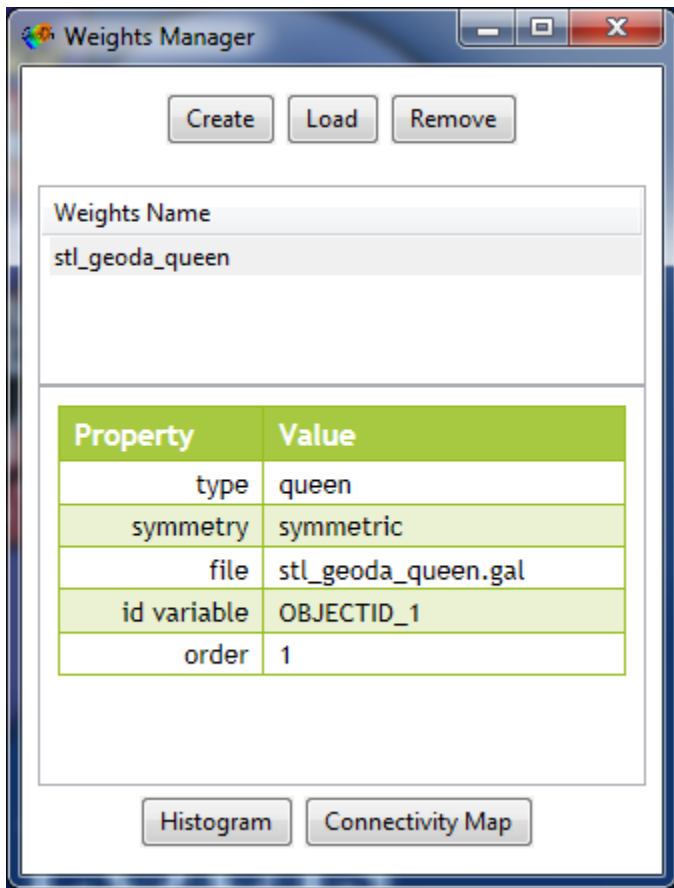
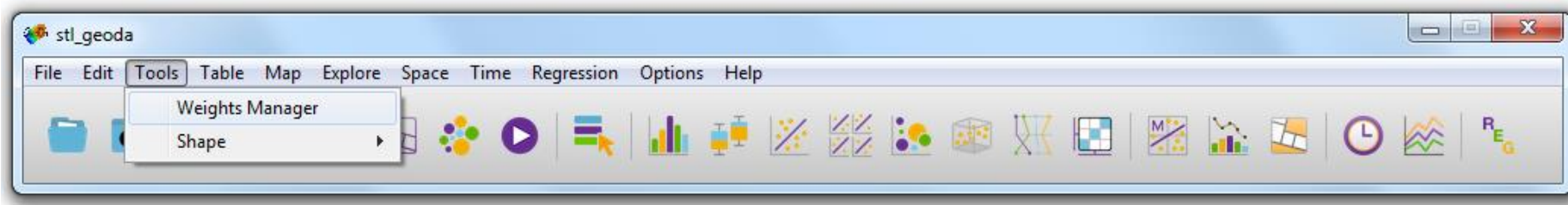
SAINT LOUIS

(2) You should have something like this....





### (3) Create your weights and view results→ Tools → Weights → Connectivity Histogram



#### (4) Review the connectivity histogram→ Tools → Weights → Connectivity Histogram

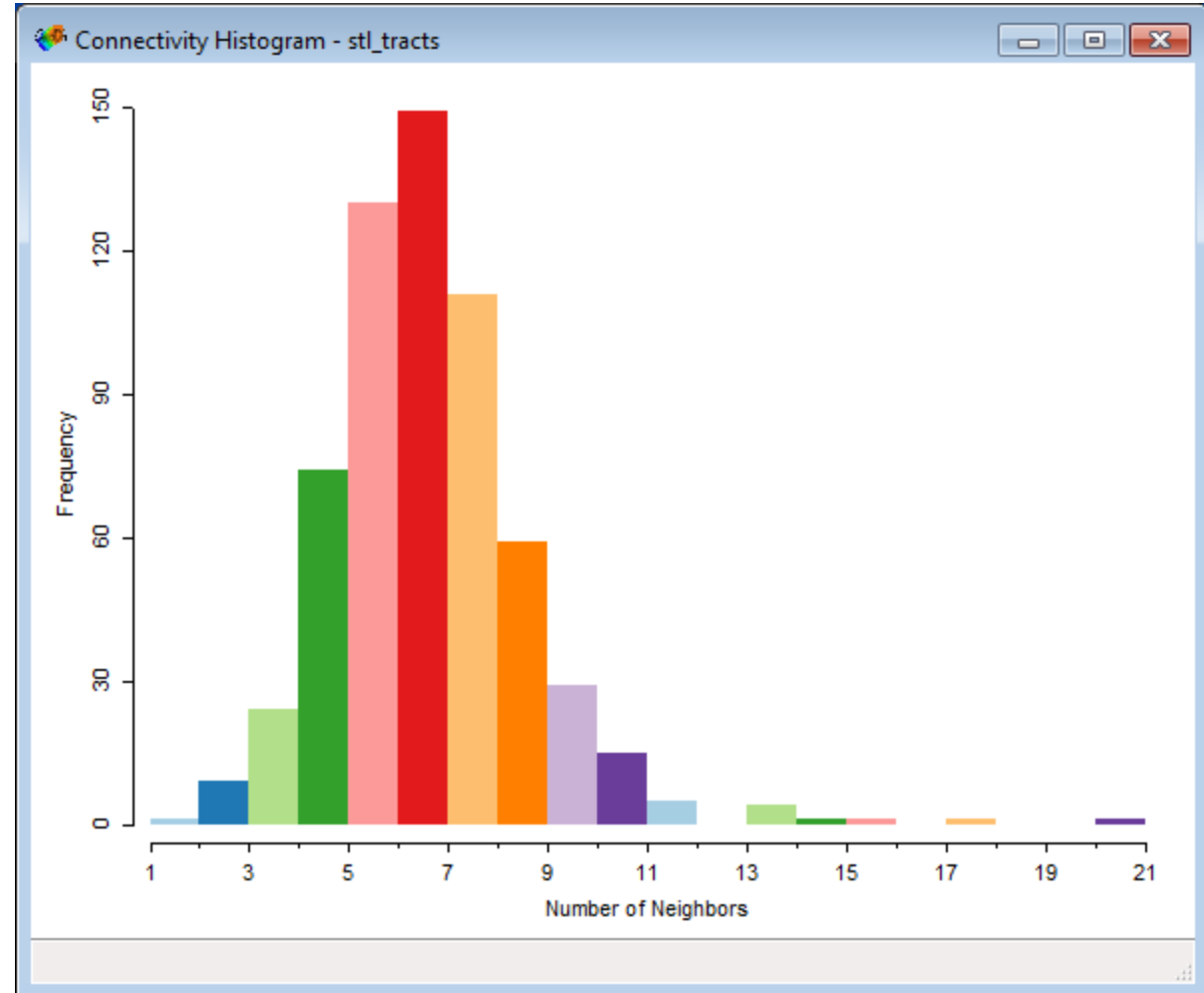
Weights Manager

Create Load Remove

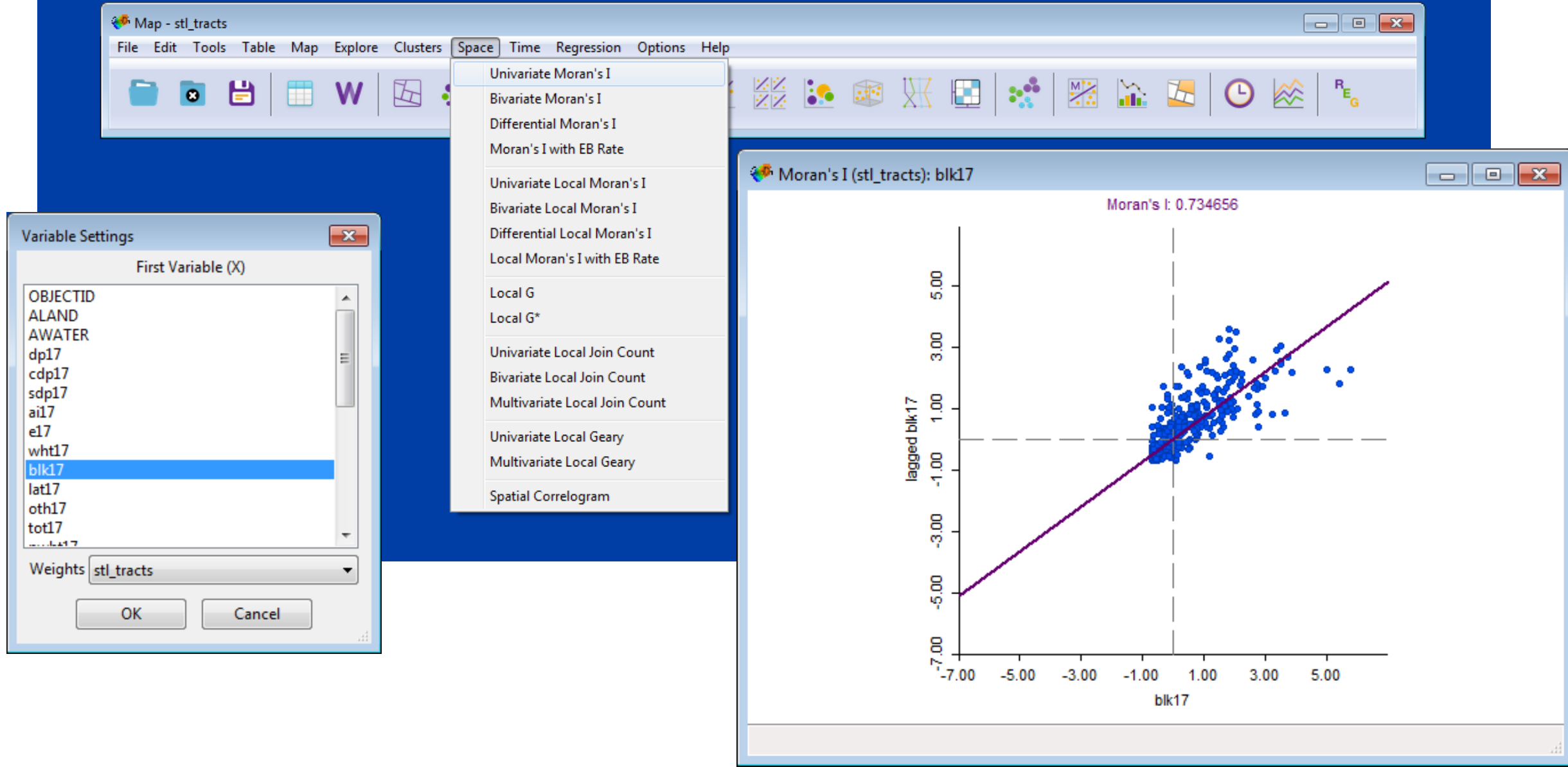
Weights Name  
stl\_tracts

Property	Value
type	queen
symmetry	symmetric
file	stl_tracts.gal
id variable	OBJECTID
order	1
# observations	614
min neighbors	1
max neighbors	20
mean neighbors	6.13
median neighbors	6.00
% non-zero	1.00%

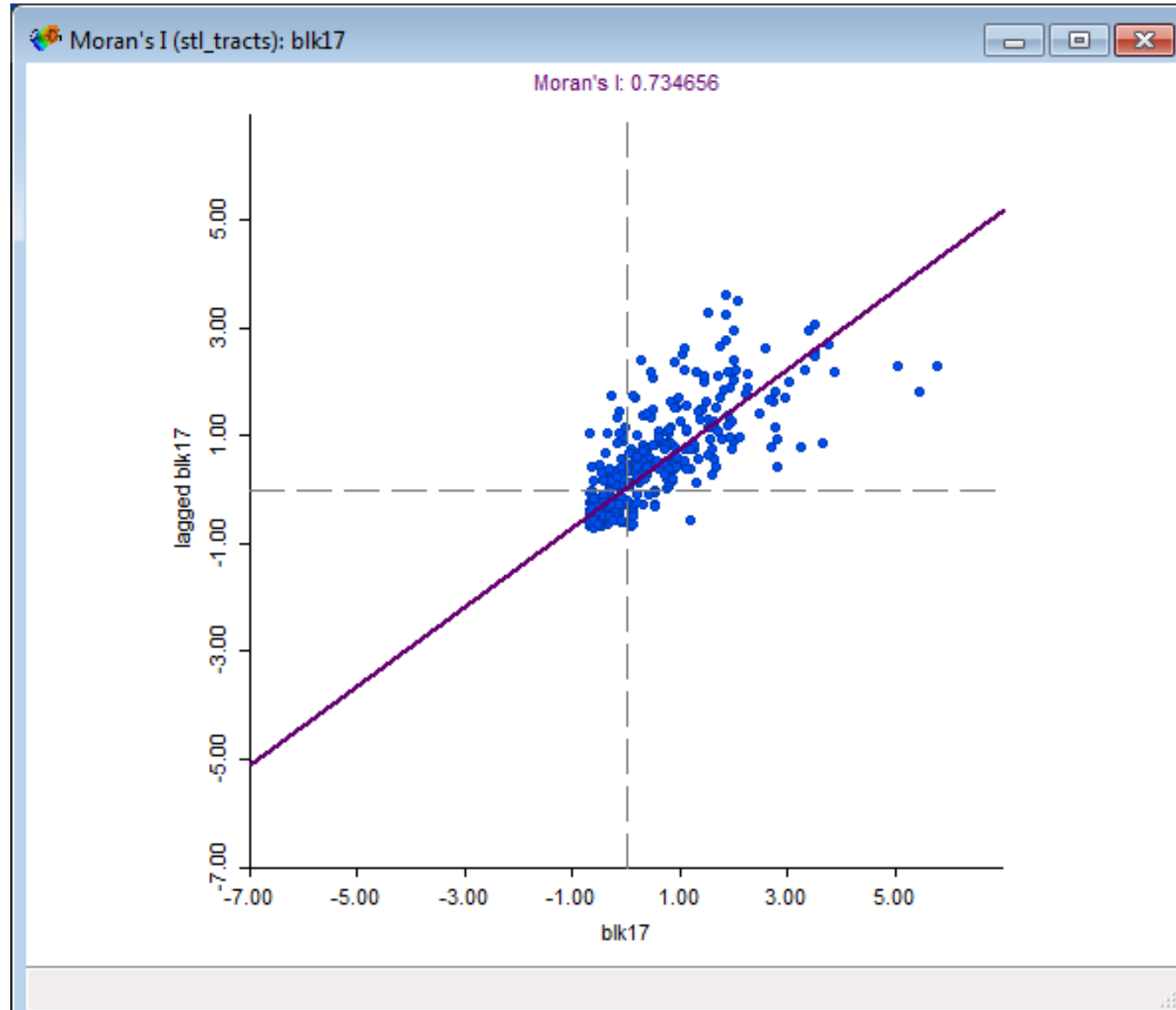
Histogram Connectivity Map Connectivity Graph



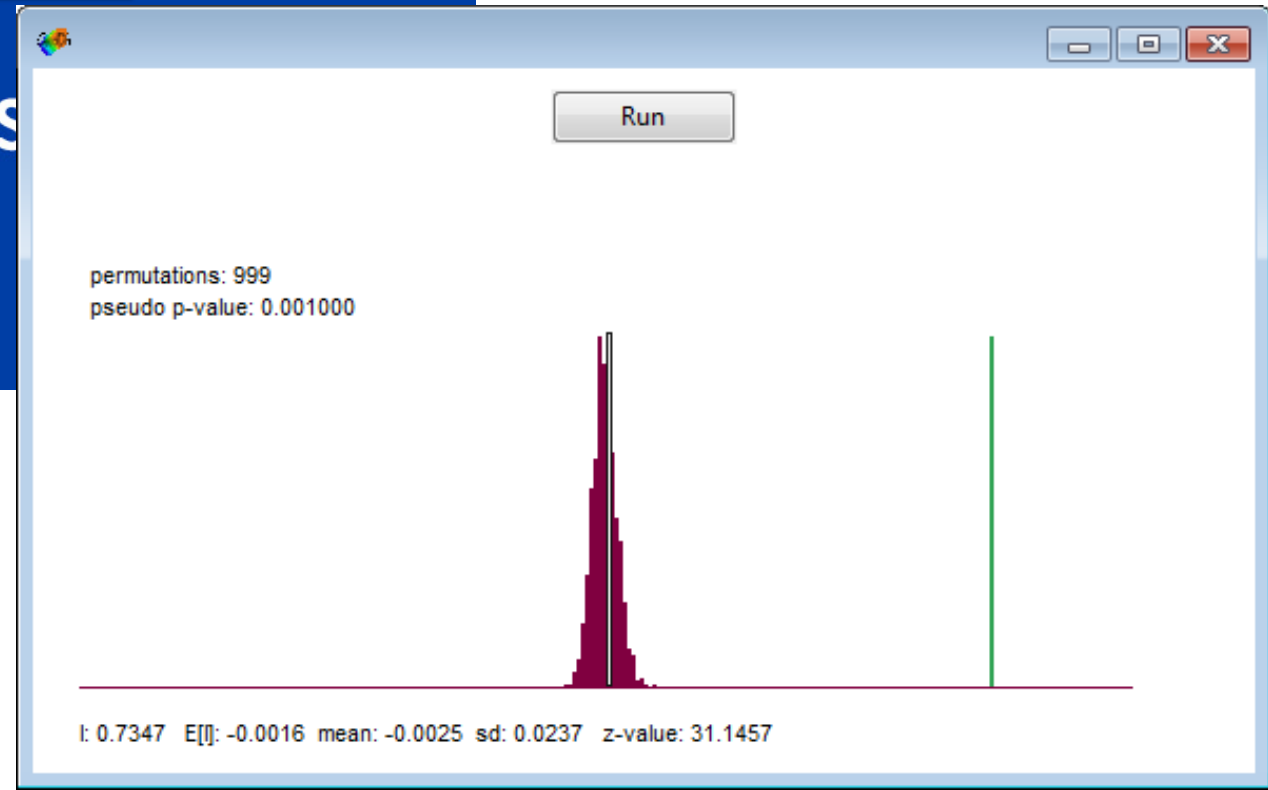
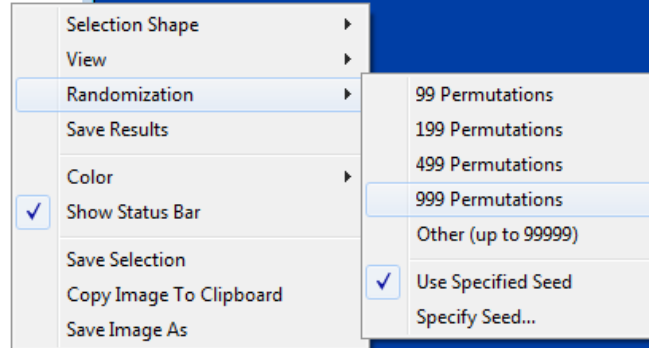
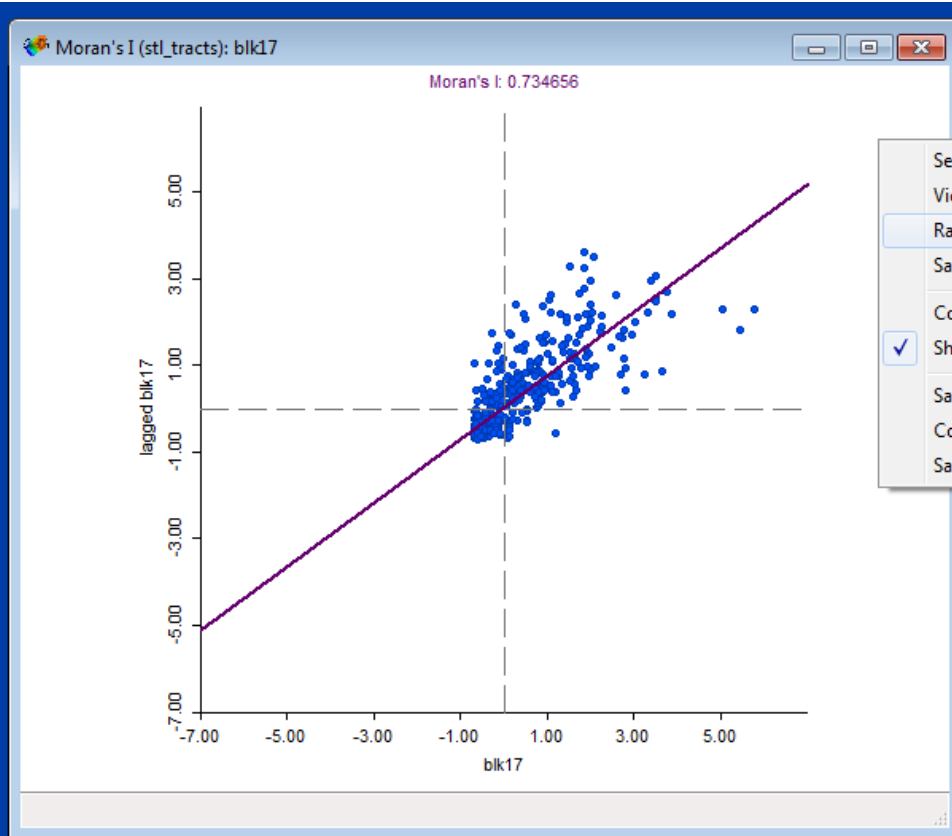
(4) Compute a global Moran's I statistic → Space → Univariate Moran's I → Select your variable → Interpret your output



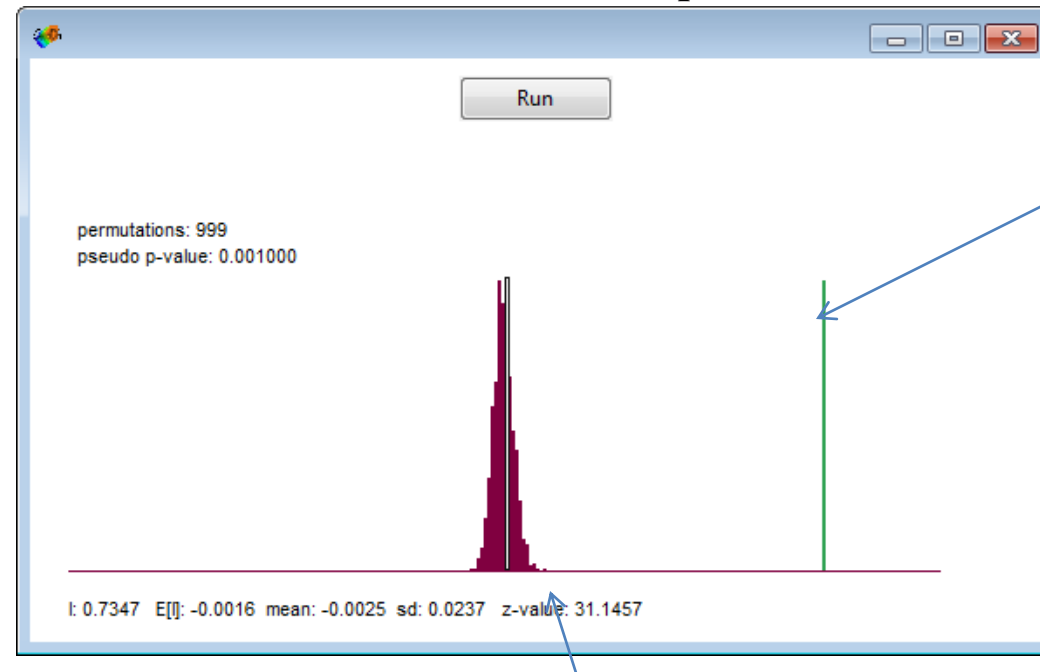
The four quadrants in the graph provide a classification of four types of spatial autocorrelation: *high-high* (upper right), *low-low* (lower left), for positive spatial autocorrelation; *high-low* (lower right) and *low-high* (upper left), for negative spatial autocorrelation. The slope of the regression line is Moran's I, listed at the top of the graph.



(5) Compute a pseudo p-value to determine if the Moran's I is = 0 → Right click in the scatterplot → select Randomization → select 999 Permutations



**(6) The sixth step is to test for significance.** Test for significance for Moran's I is based on a permutation approach, in which a reference distribution is calculated for spatially random layouts with the same data (values) as observed. The randomization uses an algorithm to generate spatially random simulated data sets outlined in Anselin (1986). The result is a window depicting a histogram for the reference distribution, with the observed Moran's I shown as a yellow bar. In addition to the histogram, a small number of summary statistics are listed as well. In the top left corner, the number of permutations used (999) and the pseudosignificance level (0.001) are given. Hence, the value of 0.001 indicates that none of the simulated values were larger than the observed 0.65. In the bottom left of the graph, the value for Moran's I is listed, its theoretical mean and the mean and standard deviation of the empirical distribution.



Observed  
Moran's I

Reference Distribution

Estimated (or predicted) y value

Estimate of the regression intercept

Estimate of the regression slope

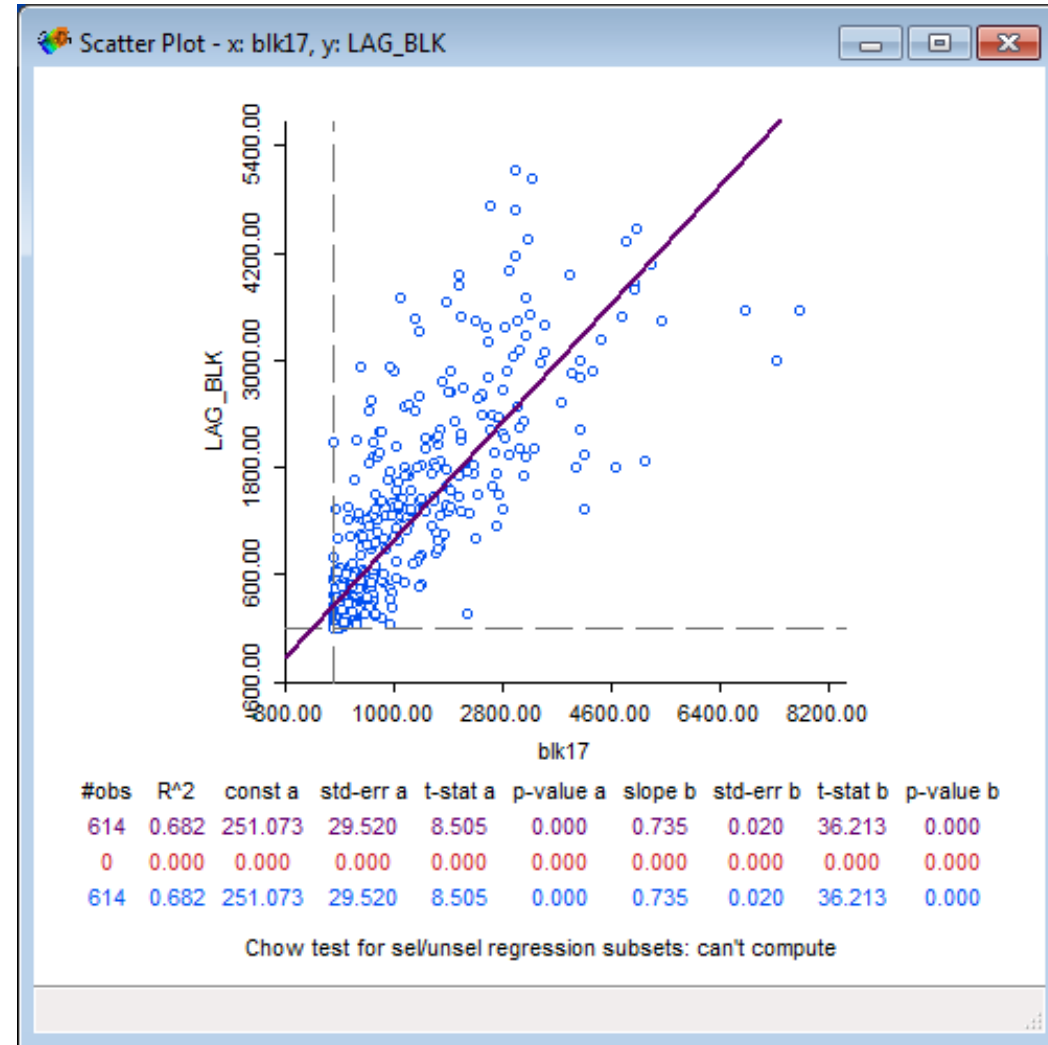
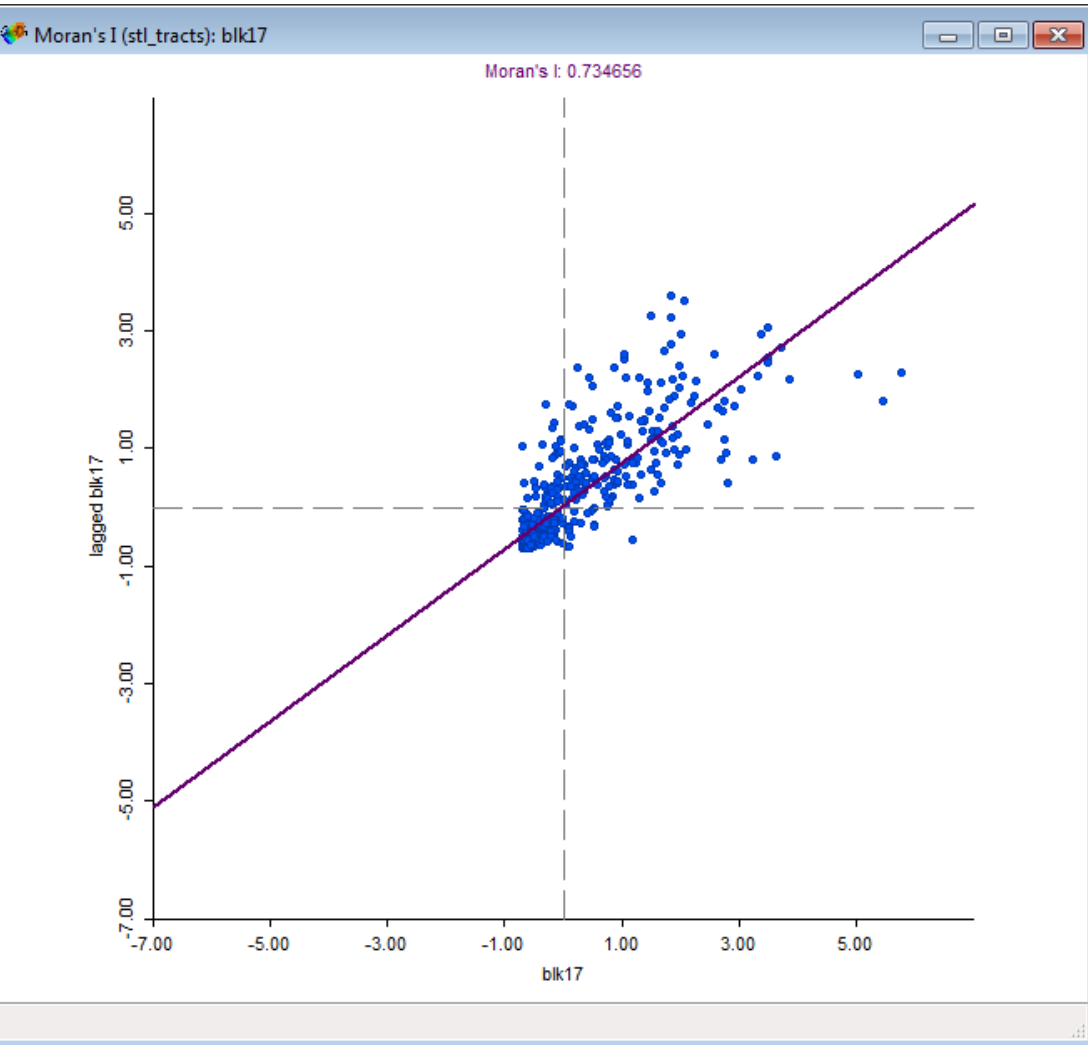
Independent variable

Error term

$$y_i = b_0 + b_1x + e$$

$$y_i = b_0 + b_1x + e$$

$$y_i = .251 + .735(X) + e$$



Let's look at the results from an OLS perspective and start to build a basic model.

Regression Report

X

>>02/27/19 09:22:48

REGRESSION

-----

SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION

Data set : stl\_tracts

Dependent Variable : LAG\_BLK

Mean dependent var : 862.874

S.D. dependent var : 1061.62

Number of Observations: 614

Number of Variables : 2

Degrees of Freedom : 612

R-squared : 0.681806

Adjusted R-squared : 0.681286

Sum squared residual:2.20192e+008

Sigma-square : 359792

S.E. of regression : 599.826

Sigma-square ML : 358620

S.E of regression ML: 598.849

F-statistic : 1311.35

Prob(F-statistic) : 0

Log likelihood : -4797.76

Akaike info criterion : 9599.53

Schwarz criterion : 9608.37

-----

Variable	Coefficient	Std. Error	t-Statistic	Probability
CONSTANT	251.073	29.5197	8.50526	0.00000
blk17	0.734656	0.0202873	36.2126	0.00000

-----

REGRESSION DIAGNOSTICS

MULTICOLLINEARITY CONDITION NUMBER 1.917393

TEST ON NORMALITY OF ERRORS

TEST	DF	VALUE	PROB
Jarque-Bera	2	824.6386	0.00000

DIAGNOSTICS FOR HETEROSKEDASTICITY

RANDOM COEFFICIENTS

TEST	DF	VALUE	PROB
Breusch-Pagan test	1	420.7078	0.00000
Koenker-Bassett test	1	117.2208	0.00000

===== END OF REPORT =====

	Coefficient	Standard Error
Diversity (E15)	.734	.0202***
Constant	251.07	.29.51***
Model Summary		
n	614	
F-statistics	1311	
R-squared	.6818	

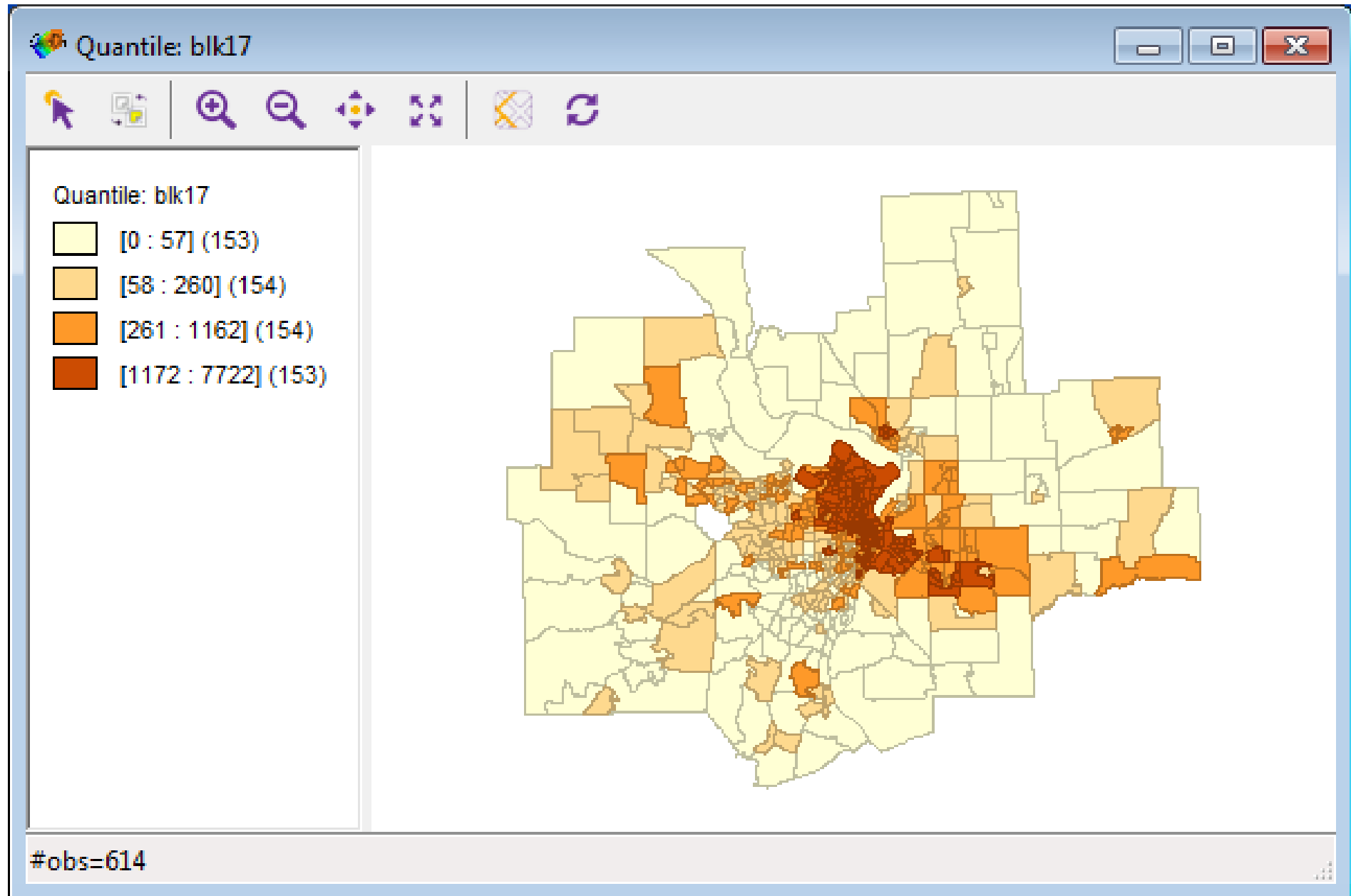
\* ≤ .05, \*\* ≤ .01, \*\*\* ≤ .001



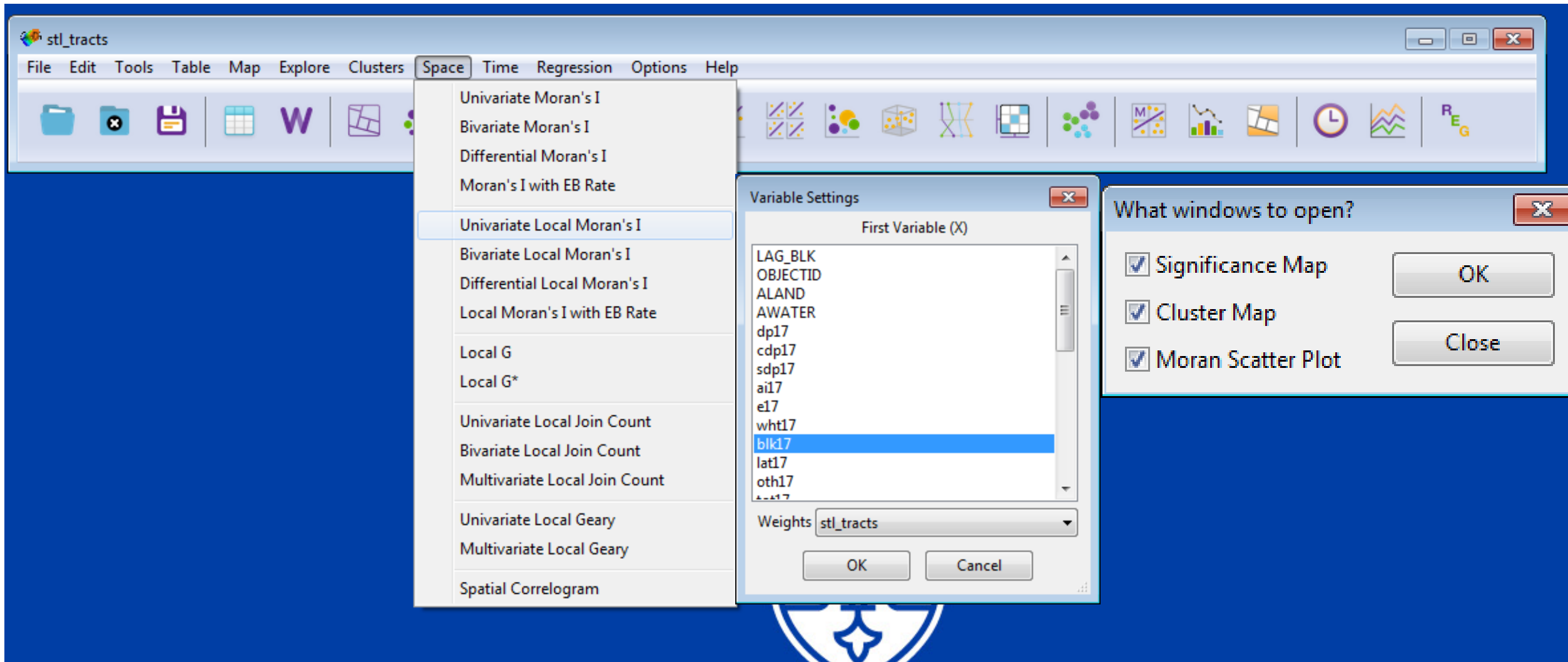
# Local Indicator of Spatial Association (LISA)

---

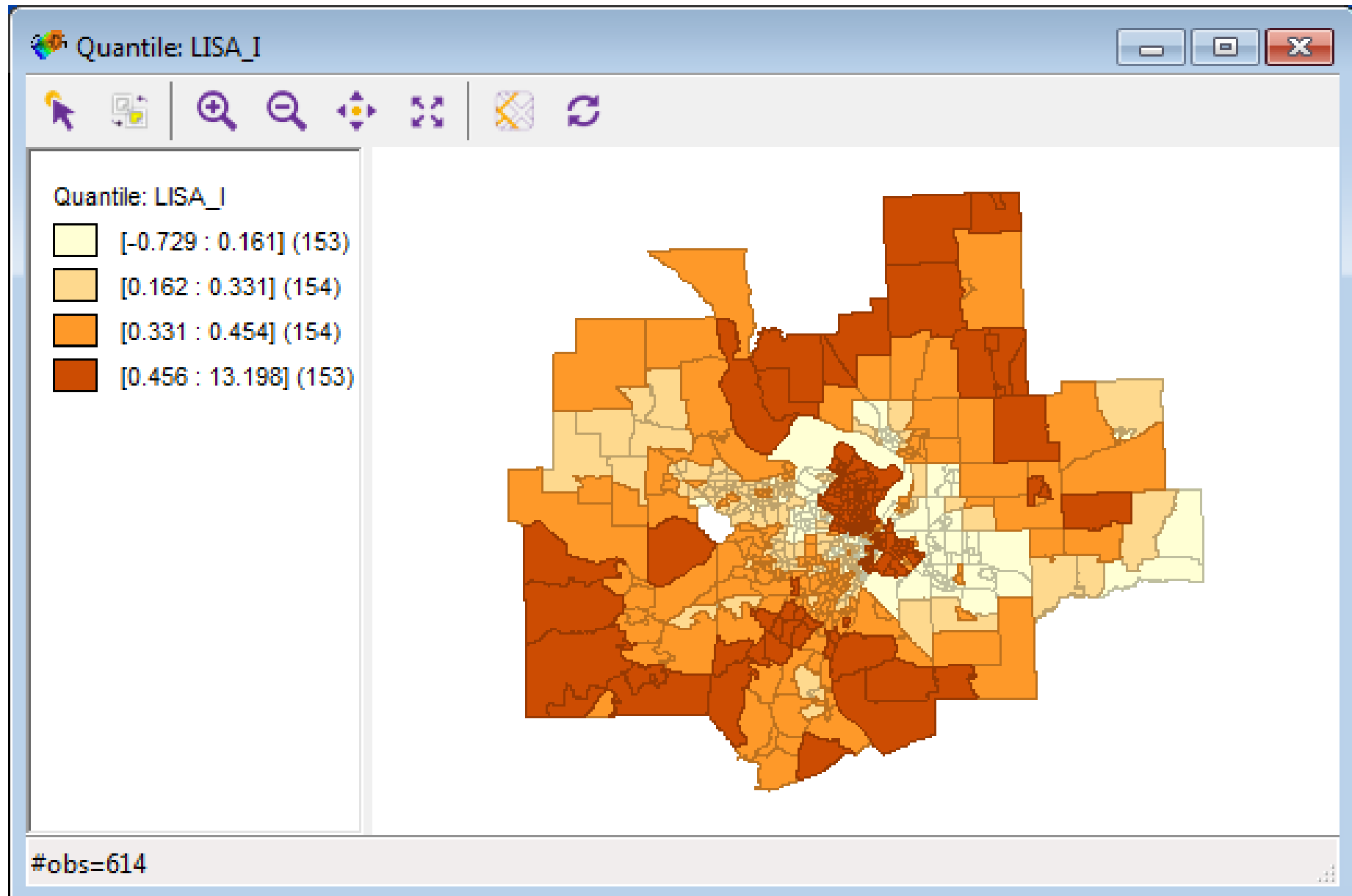
# Percent Black



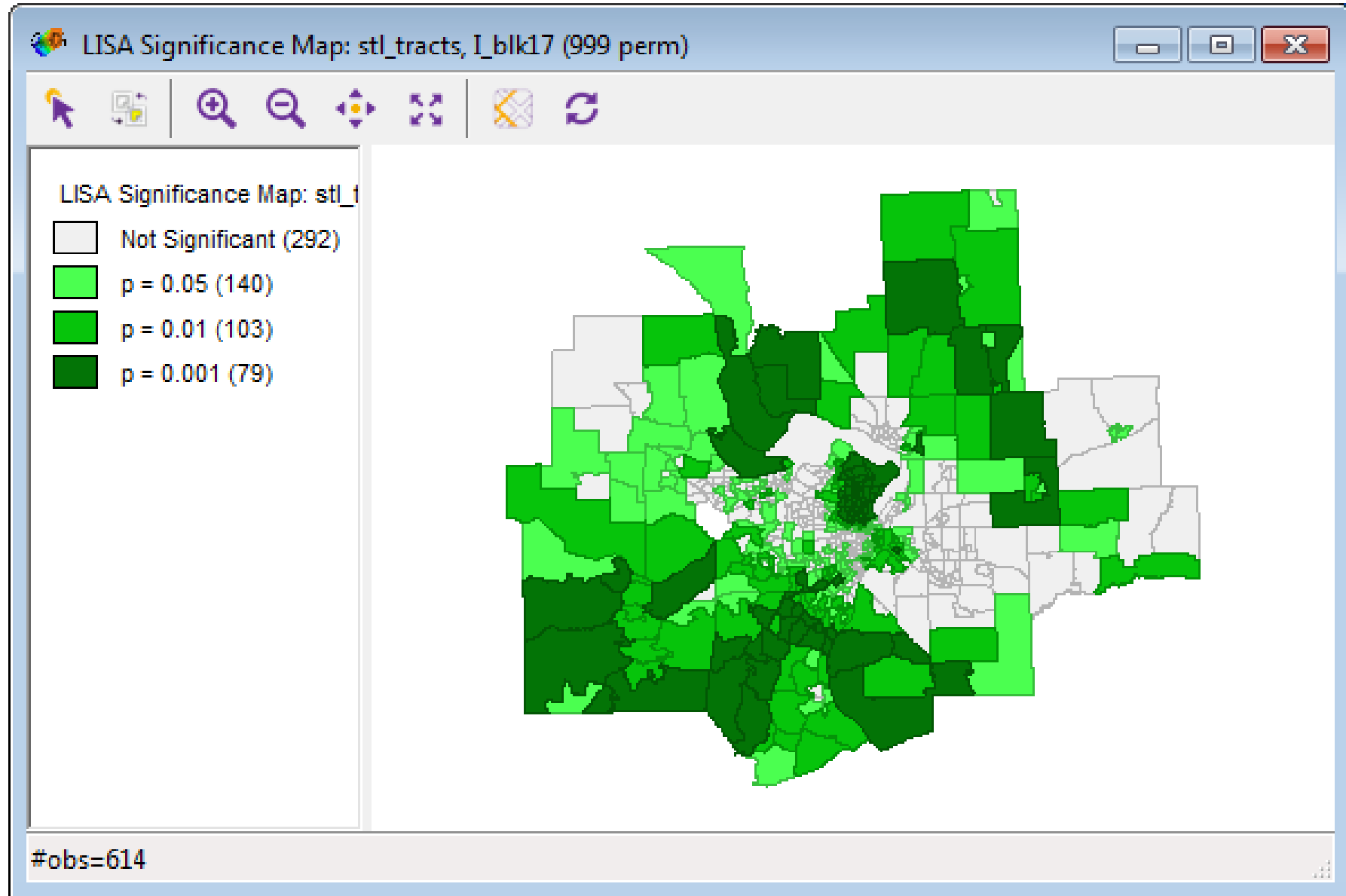
# How to produce LISA Statistics and Maps



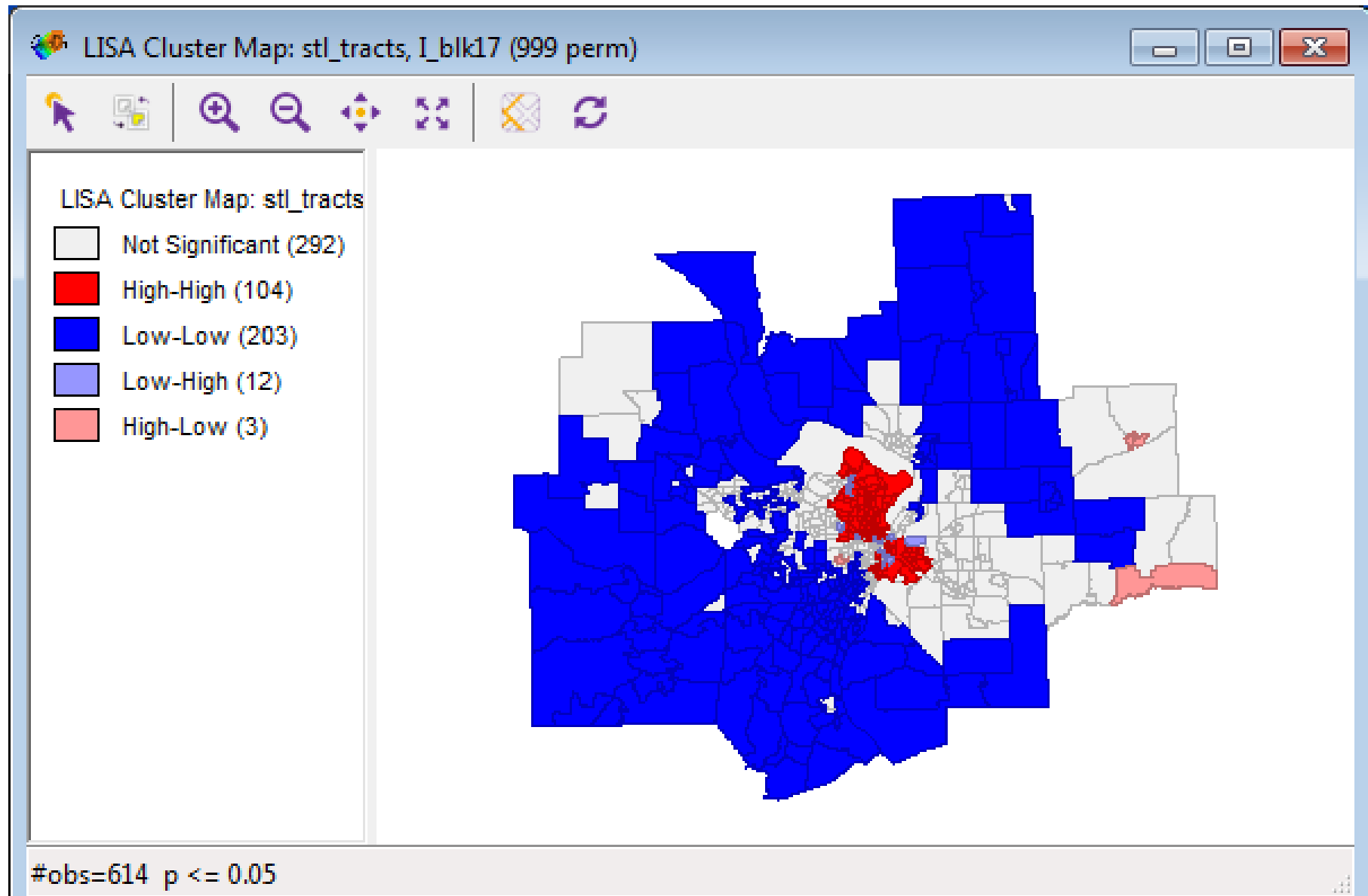
# LISA Statistic Map for Percent Black



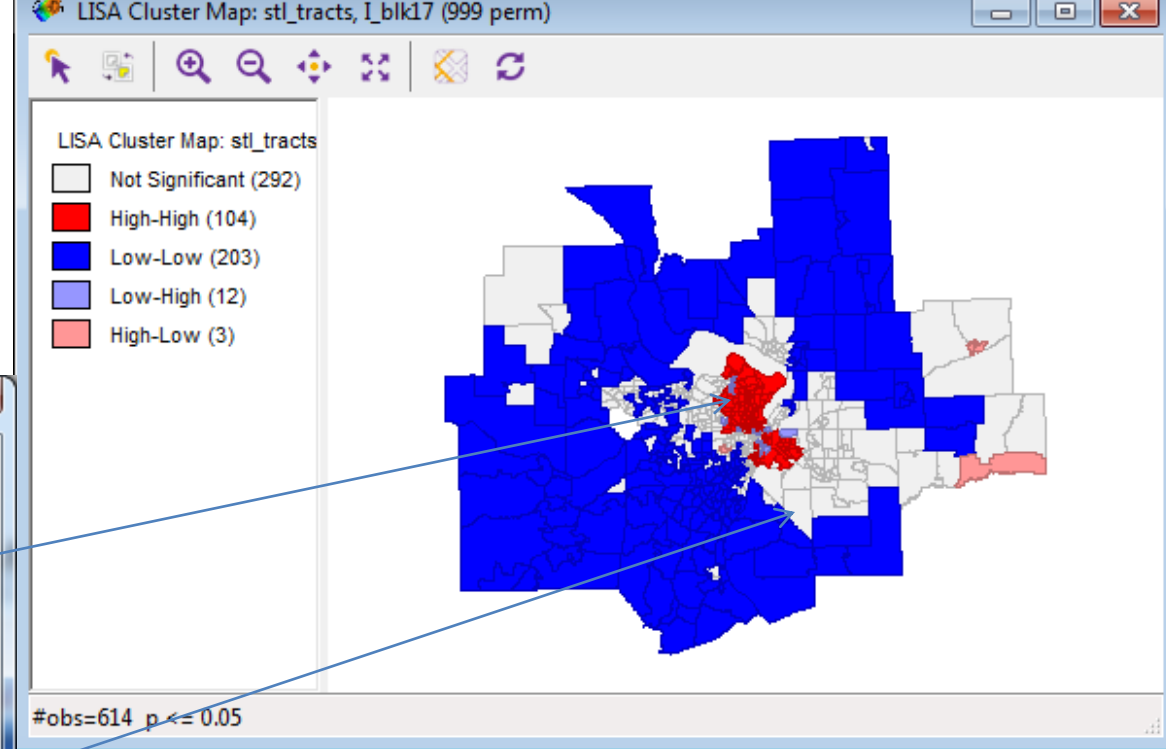
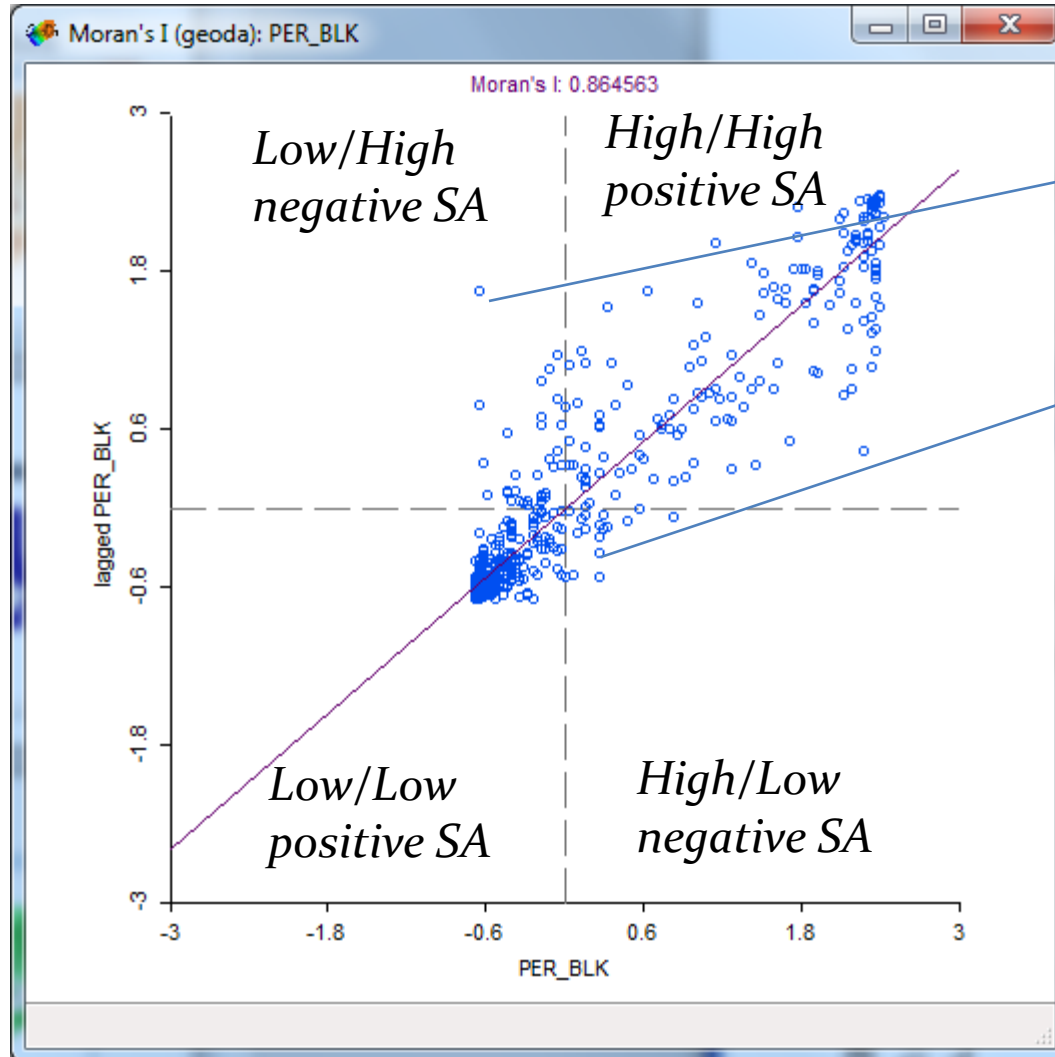
# LISA Significance Map for Percent Black



# LISA Cluster Map for Percent Black

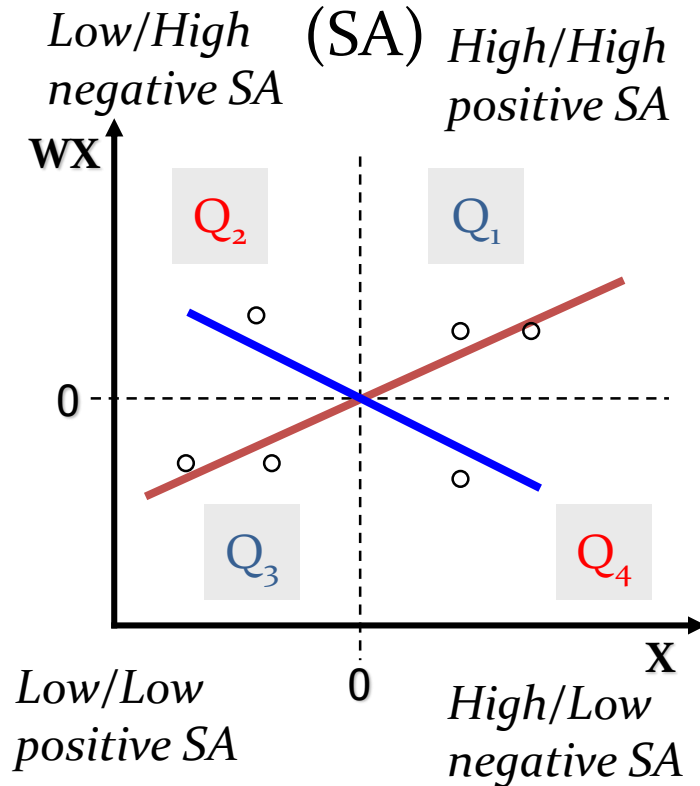


# LISA Cluster Map



# Quadrants of Moran Scatterplot

Each quadrant corresponds to one of the four different types of spatial association



*Locations of positive spatial association*  
(*"I'm similar to my neighbors"*).

Q<sub>1</sub> (values [+], nearby values [+]): **H-H**

Q<sub>3</sub> (values [-], nearby values [-]): **L-L**



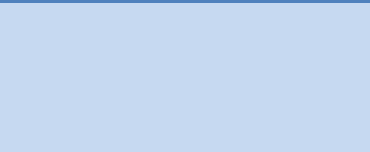
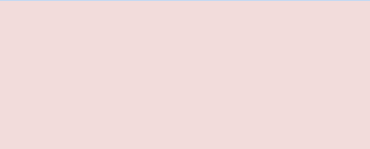
*Locations of negative spatial association*  
(*"I'm different from my neighbors"*).

Q<sub>2</sub> (values [-], nearby values [+]): **L-H**

Q<sub>4</sub> (values [+], nearby values [-]): **H-L**



# Interpreting the Legend

Color	X	Y	Correlation	Quadrant
	High	High	Positive	Q1
	Low	Low	Positive	Q3
	Low	High	Negative	Q2
	High	Low	Negative	Q4