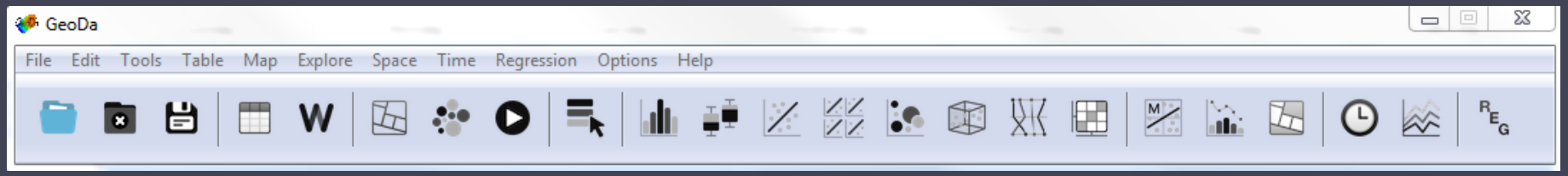
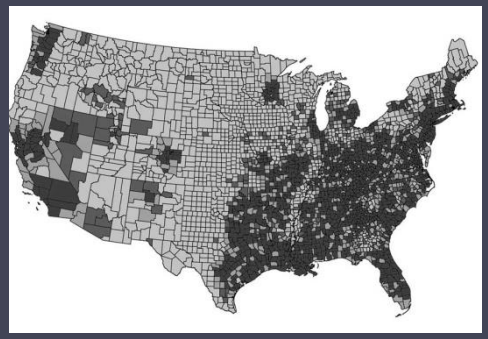




Spatial Econometrics Workshop

WAEA Annual Meeting

July 11, 2017

Brian Whitacre
Oklahoma State University

Agenda

1. Introduction to Spatial Data (1:30)
2. GeoDa Basics: Constructing Weights / Spatial Statistics (2:10)
3. Exploratory Spatial Data Analysis (ESDA) (3:00)
4. Spatial Regression (OLS Diagnostics, Lag, Error)(4:00)
5. Recent Advancements / Upcoming Trends (5:15)

1. Introduction to Spatial Data

- The concept of spatial dependence
- Basics of “neighbors” / spatial weight matrices
- Why OLS fails with spatial data

The Concept of Spatial Dependence

- Spatial analysis basically assumes that "space matters"
  - What happens in one region is related to what happens in neighboring regions
- Tobler's (1979) First Law of Geography
  - "Everything is related to everything else, but closer things more so"

Some Examples…

- Where Americans get enough exercise
  - http://www.theatlanticcities.com/arts-and-lifestyle/2014/01/where-americans-get-enough-exercise/5874/
- Pockets of persistent poverty
  - http://www.ers.usda.gov/topics/rural-economy-population/rural-poverty-well-being/geography-of-poverty.aspx#.UtQRRvRDthE
- Obesity
  - http://www.cdc.gov/obesity/data/adult.html
- Unemployment rates
  - http://www.latoyaegwuekwe.com/geographyofarecession.html
- A whole host of interesting maps!
  - https://www.washingtonpost.com/blogs/govbeat/wp/2014/02/24/25-maps-and-charts-that-explain-america-today/?hpid=z4

Examples of Spatial Processes

- Patterns of interdependence
  - "Spatial dependence"
  - Increased police presence in one neighborhood may alter crime levels in nearby neighborhoods
- Broad patterns of similarity based on history / climate
  - "Spatial heterogeneity"
  - People living in northern areas are more likely to play hockey than those in the south
  - Poverty rates differ dramatically across counties, but "pockets" exist
- Diffusion
  - Technically spatial dependence – but – potentially different outcomes
  - Language / dialect drift
  - Knowledge centers

More Technically…

- Typical OLS model:

$$y_i = X_i\beta + \varepsilon_i$$
$$\varepsilon_i \sim N(0, \sigma^2) \qquad i = 1, \ldots, n$$

- Each observation has an underlying mean of $X_i\beta$ and a random component ε_i
- If i represents regions or points in space, OLS assumes that observed values at one location are independent of observations at other locations.
  - Statistically independent observations:

$$E(\varepsilon_i\varepsilon_j) = E(\varepsilon_i)E(\varepsilon_j) = 0$$

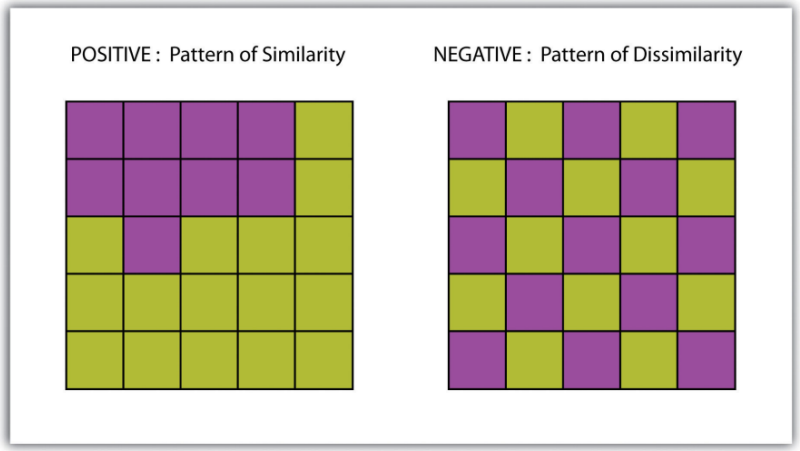
Spatial Dependence

- In spatial contexts, the assumption of statistically independent observations is unlikely
- Instead, we have a situation where the values at one location depend on values of neighboring observations
  - This is spatial dependence!

$$
\begin{aligned}
y_i &= \alpha_i y_j + X_i\beta + \varepsilon_i \\
y_j &= \alpha_j y_i + X_j\beta + \varepsilon_j \\
\varepsilon_i &\sim N(0, \sigma^2) \qquad i = 1 \\
\varepsilon_j &\sim N(0, \sigma^2) \qquad j = 2
\end{aligned}
$$

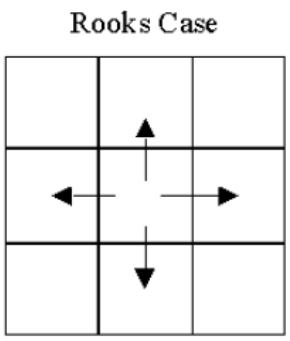
Spatial Autocorrelation

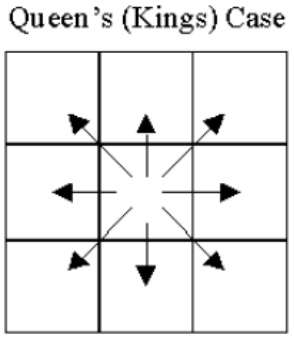
- Positive spatial autocorrelation
  - High or low values tend to cluster in space
- Negative spatial autocorrelation
  - Locations tend to be surrounded by neighbors with very *dissimilar* values (this is rare)



Spatial Neighbors and Weights

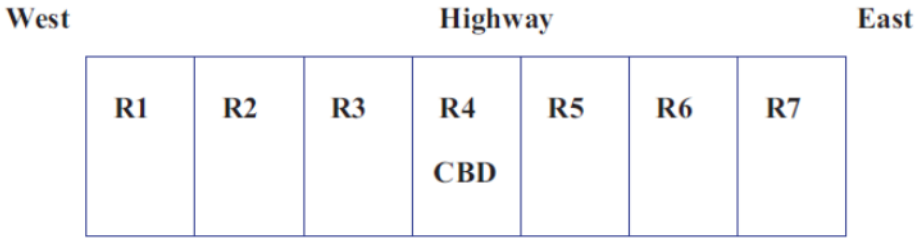
- The Spatial Weight matrix provides information on which regions are neighbors (or are 'contiguous' – share a common boundary)
  - Elements that are neighbors get a '1', non-neighbors get a '0'
  - Diagonals are set to 0 by convention
  - But we can define "neighbors" very differently!
- Some examples:
  - Rook contiguity
  - Queen contiguity
  - Distance-based
  - K-nearest neighbors





Spatial Weight Matrix Example

- Example of spatial weight matrix for 7 regions:



- Note that regions are NOT considered neighbors to themselves

$$C = \begin{pmatrix} & R1 & R2 & R3 & R4 & R5 & R6 & R7 \\ R1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ R2 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ R3 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ R4 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ R5 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ R6 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ R7 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Spatial Weight Matrix

- We then normalize the matrix C so that each row sums to 1 …this is a "row-stochastic matrix" (W)

$$W = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

- This 7 x 7 matrix can then be multiplied by a 7 x 1 vector containing y values from each region:

$$Wy = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \end{pmatrix} = \begin{pmatrix} y_2 \\ (y_1 + y_3)/2 \\ (y_2 + y_4)/2 \\ (y_3 + y_5)/2 \\ (y_4 + y_6)/2 \\ (y_5 + y_7)/2 \\ y_6 \end{pmatrix}$$

Spatial Weight Matrix

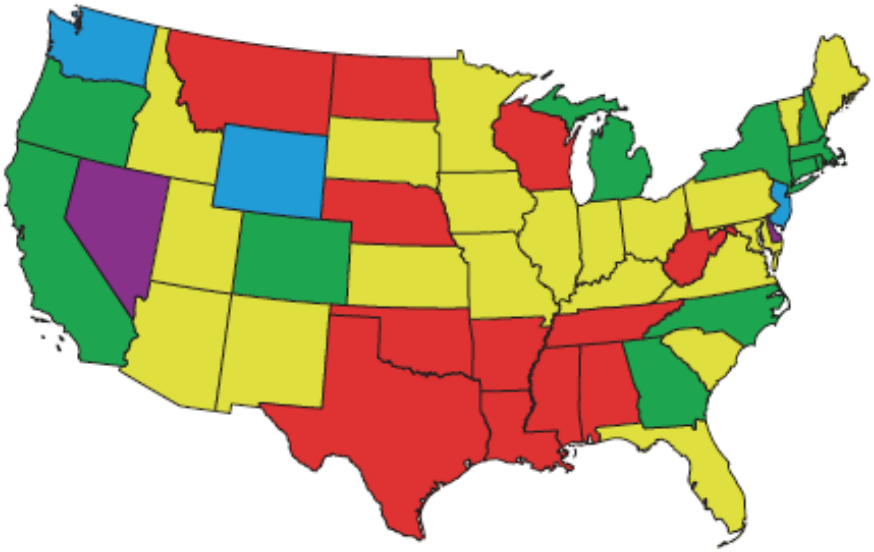
- There are many ways of formulating spatial weight matrices (rook, queen, distance, etc.)
  - Alternative ways of weighting neighboring observations
  - 1st order vs. 2nd order…
- This can get messy with 3,000+ counties, 30,000 ZIP codes, or 8.2M census blocks!
- We use software to calculate these weights
  - Arc GIS
  - GeoDa

Illustration: Non-spatial regressors ignore spatial dependence

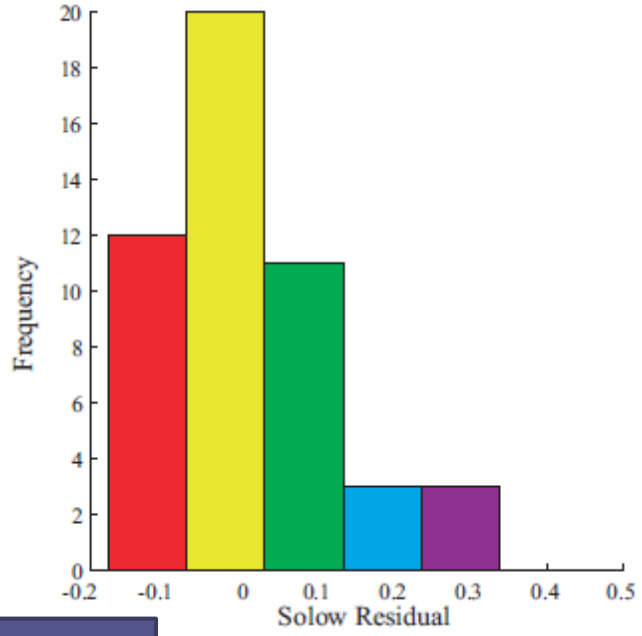
- Traditional production function (state-level):
  - $\ln(Q) = \alpha i_n + \beta \ln(K) + \gamma \ln(L) + \varepsilon$
    - Output = f(capital, labor)
    - Q = Gross State Product in 2001
    - K = Capital estimates in 2001 (Garofalo & Yamarik)
    - L = Total non-farm employment in 2001
  - Residuals from OLS = Solow Residuals
    - Reflect economic growth not explained by K,L

Map of Solow Residuals

Solow Residuals, 2001 U.S. States



Solow Residuals Map Legend



What Patterns Do You See?
What Types of Residuals are 'Clustered'?
What Does That Mean?

Spatial Dependence

- Clustering represents a visual depiction of spatial dependence in the residuals
  - What leads to this spatial dependence?
    - One answer: spillovers in technological innovation
    - Or: cultural, infrastructure, or recreation variables that may not be measurable
  - How can we model this?
    - One way: include a *spatially lagged dependent variable*
    - Uses average of DV values (Q) from neighbors on RHS

Spatial Lag Model

- Sometimes referred to as Spatial Autoregressive (SAR) Model or Process
- Includes a SPATIAL LAG of the DV (i.e. average of neighboring values of the DV) on the RHS

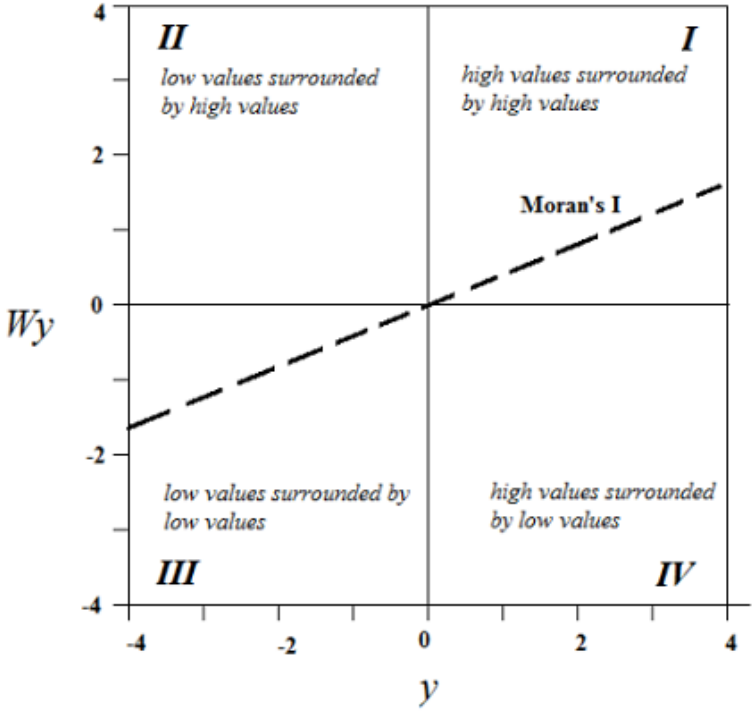
$$y_i = \rho \sum_{j=1}^{n} W_{ij} y_j + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma^2) \qquad i = 1, \ldots, n$$

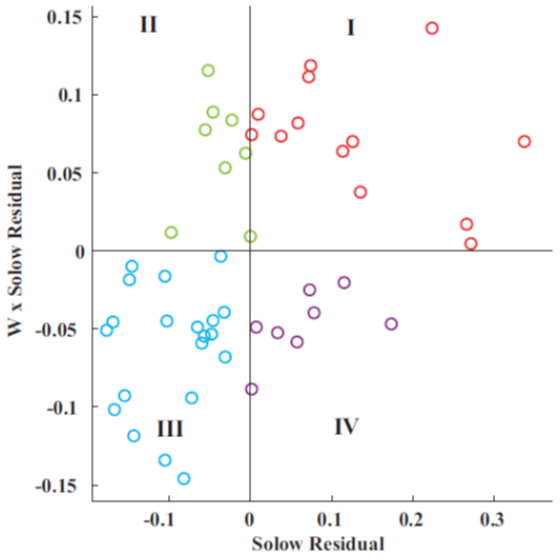
- Key concept here:
  - The spatial weights matrix W_{ij} (n x n)
  - Associated spatial weight parameter ρ

Moran Scatter Plot

- Relationship between y and the average values of neighboring observations in Wy

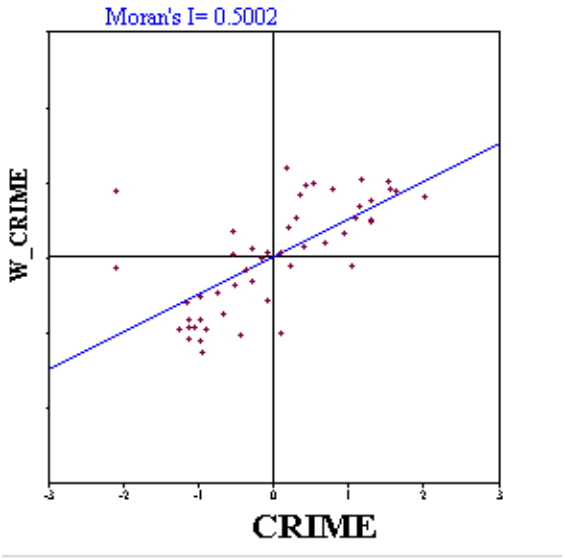


In our example:



Moran Scatter Plot

- This type of plot suggests that the spatial weight parameter, ρ, is >0
  - There is positive association between y and Wy
  - This is evidence of positive spatial dependence



What would the plot look like if there were negative spatial dependence?

Later, we will formally test for spatial dependence using a "Moran's I statistic"

Returning to the Spatial Lag Model

- Also referred to as the Spatial Autoregressive Model
  - $y = \rho W y + X\beta + \varepsilon \qquad \varepsilon \sim N(0, \sigma^2 I_n)$
- Note that if $\rho = 0$, there is no spatial dependence and the model is essentially OLS
- Real-world example:

Baller, Anselin, Messner, Deane and Hawkings: Homicide rate

$$H_i = \rho W H_i + \beta_0 + \beta_1 D_i + \beta_2 P_i + \beta_3 V_i + \beta_5 U_i + \beta_6 E_i + \varepsilon_i$$
$$\varepsilon \approx N(0, \sigma^2 I)$$

H: homicide rate
i: U.S. county
D: resource deprivation index
P: population structure
V: percent divorced
U: unemployment rate
E: median age

W: spatial weight matrix
ρ: spatial autoregressive parameter
β_0, β_1, β_2, β_3, β_5, β_6: exogenous vars. Parameters
ε: i.i.d. stochastic error term

. BALLER, R., L. ANSELIN, S. MESSNER and D. HAWKINS (2001), *"Structural covariates of U.S. county homicide rates: incorporating spatial effects"*. Criminology.

Another Example of a Spatial Lag Model

Mobley, Root, Anselin, Lozano and Koshinsky: Admissions for Ambulatory Care Sensitive Conditions (ACSCs)

$$ACSC_i = \rho \cdot W \cdot ACSC_i + \beta_1 + \beta_2 phy_i + \beta_3 vis_i + \\ + \beta_4 pov_i + \beta_5 hosp_i + \beta_6 insur_i + \beta_7 sprawl_i + u_i$$

ACSC:	Admissions for Ambulatory Care Sensitive Conditions rates
i:	6,475 U.S. Primary Care Service Areas (PCSAs)
W:	spatial weight matrix
phy:	physicians per capita
vis:	elderly visits (per capita) to doctors
hos:	hospitals (per capita)
insur:	availability of supplemental coverage (instead or in addition to Medicare)
sprawl:	urban sprawls (long commutes for the local workforce)
W:	distance-based spatial weight matrix
ρ, β_j:	parameters to be estimated
u:	stochastic error term

. MOBLEY L., E. ROOT, L. ANSELIN, N. LOZANO and J. KOSHINSKY (2006), "Spatial analysis of elderly access to primary care services", *International Journal of Health Geographics* 5.

Spatial Lags of OTHER Variables

- Spatial Durbin: Captures spatial dependency of DV but also accounts for spillover effects of *independent* variables (x)

$$y = \rho Wy + x\beta + Wx\gamma + \varepsilon$$

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

- Allows for measurement of "spillover" effects from neighboring locations

Spatial Durbin Example

- Deller and Watson (2016) explore whether economic diversity affected employment stability during the Great Recession
  - y = stability in unemployment rates, wages
  - x = Herfindahl (diversity) index

DIRECT effect of diversification

INDIRECT effect of diversification (from neighboring counties)

$$y = \rho W y + x\beta + W x\gamma + \varepsilon$$

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

Coming Up…

- Intro to GeoDa
- Opening Simple Maps
- Constructing / Understanding Various Spatial Weight Matrices
- Basic Spatial Statistics
  - Moran's I
  - Local Moran's I (Local Indicator of Spatial Association)

Assignment

- Write down a hypothetical cross-section model (for a topic that interests you) where you suspect spatial dependence will be observed.
  - What is your main hypothesis?
  - What is your unit of analysis?
  - What are your control variables?

BREAK

2. GeoDa Basics: Constructing Weights / Spatial Statistics

Luc Anselin

- Introduction to GeoDa
- Constructing Spatial Weights in GeoDa
- Basic Spatial Statistics
  - Moran's I (global)
  - Local Moran's I (LISA)

Introduction to GeoDa

Luc Anselin

- Free open source software tool that facilitates exploration and analysis of geospatial data
- Result of decades-long work on software development for spatial analysis (Anselin, 1991; 2006; 2012)
- First released in 2002
- Current version 1.8 (2017)
- ~200,000 users

The GeoDa Toolbar

- Data Entry
- Data Manipulation
  - Table functionality (joining tables)
  - New variable creation / transformation
- Weights Manager
  - Create spatial weights
  - Connectivity histogram

The GeoDa Toolbar

- Mapping and Geovisualization
  - Choropleth maps (quantile, s.d., box map)
  - Cartogram
  - Map movies
- Exploratory Spatial Data Analysis (ESDA)
  - Histogram
  - Box Plot
  - Scatter Plot
  - Bubble Chart

The GeoDa Toolbar

- Spatial Autocorrelation Analysis
  - Global spatial autocorrelation (Moran's scatterplot)
  - Local spatial autocorrelation (LISA)

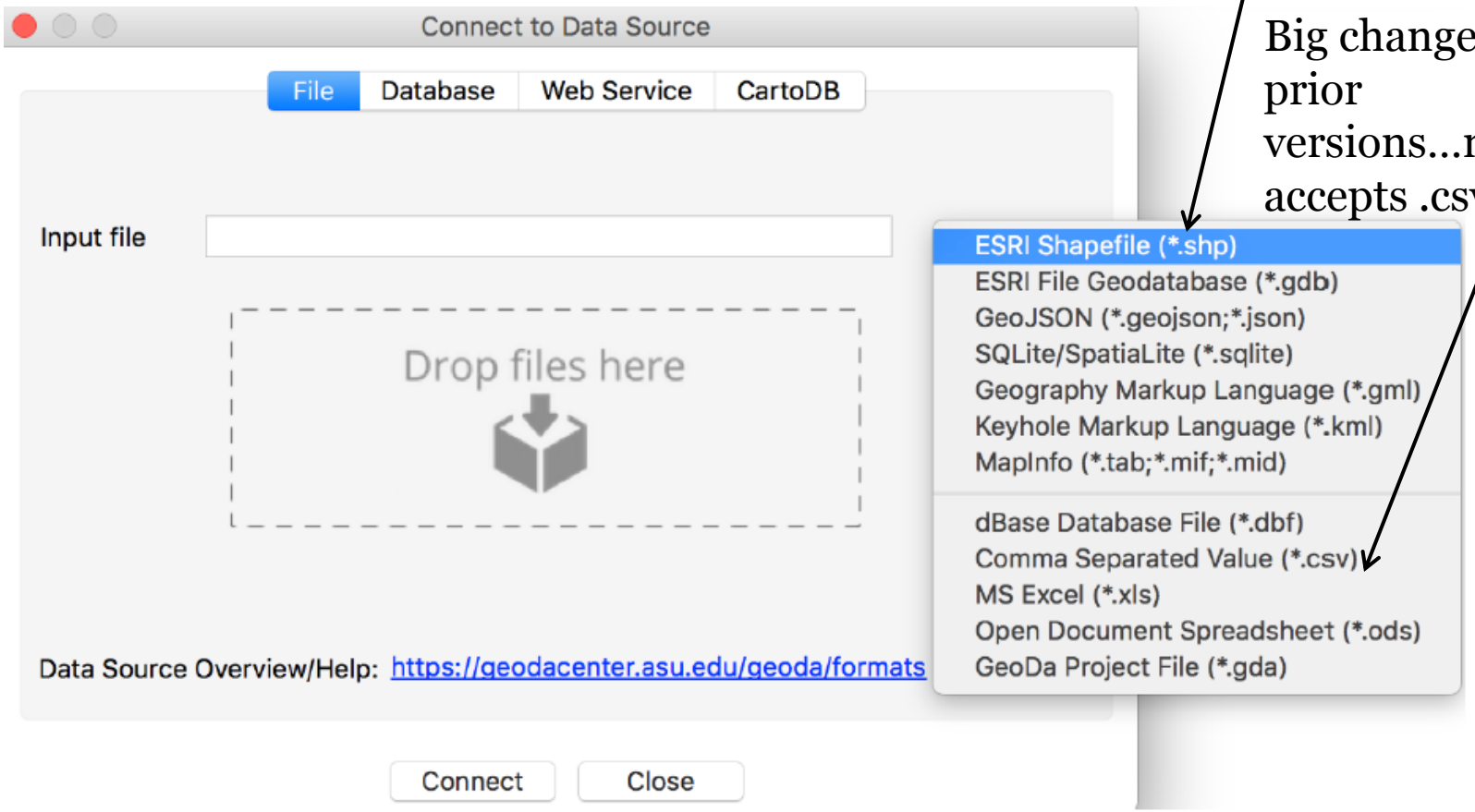
- Spatial Regression
  - OLS with spatial diagnostics
  - ML estiamtation of spatial lag model
  - ML estimation of spatial error model
  - Residuals / predicted values

Data Input

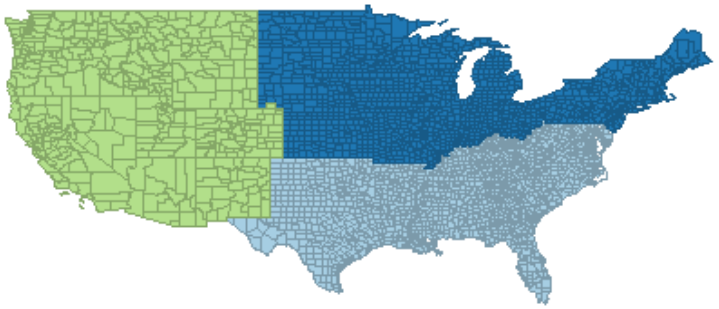
- Loading different file types

Most will start by constructing .shp in ArcGIS

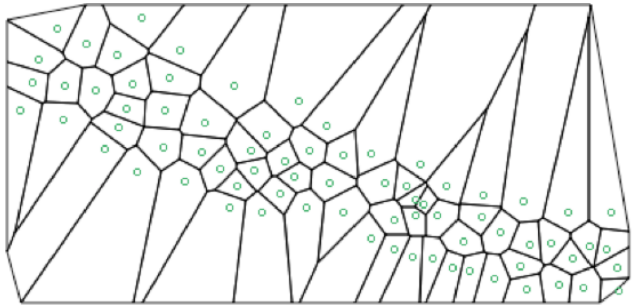
Big change from prior versions…now accepts .csv, .xls



About Spatial Data…



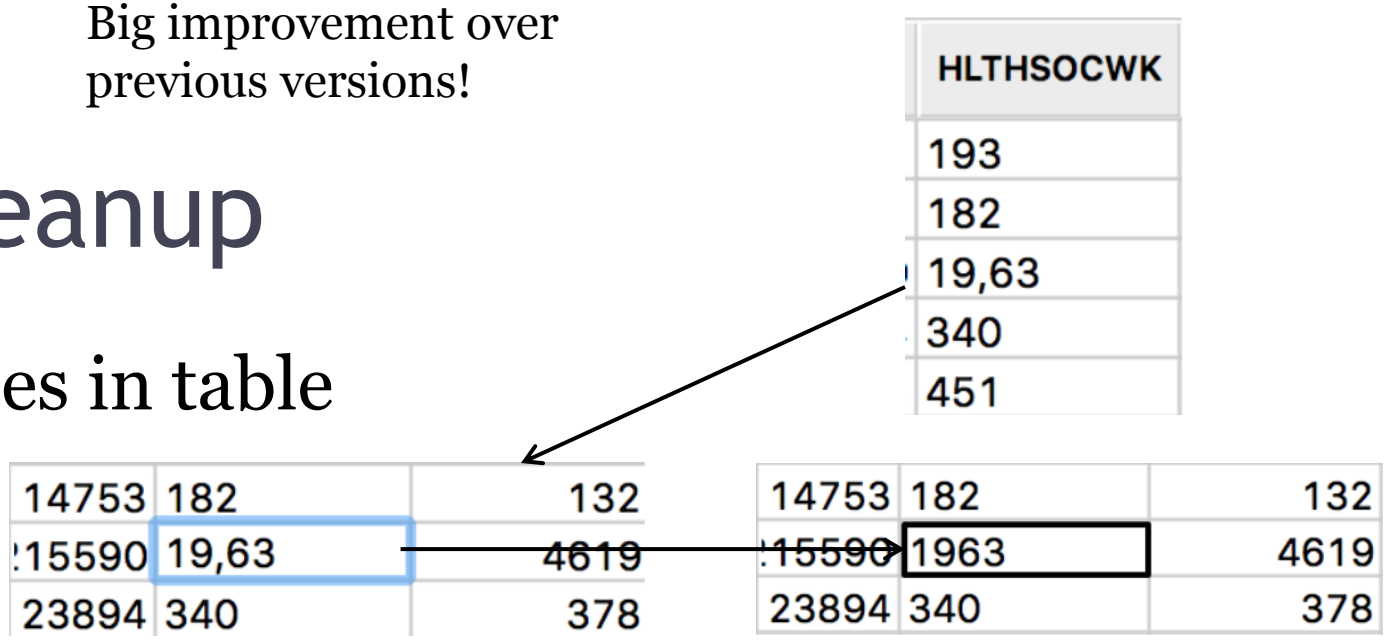
- GeoDa can handle:
  - Shapefiles (.shp) – popular data format for GIS
    - Census TIGER files (counties, ZIP codes, Census tracts)
    - USDA
  - Creating point layers from x,y coordinates (lat / long)
  - Creating polygons from points



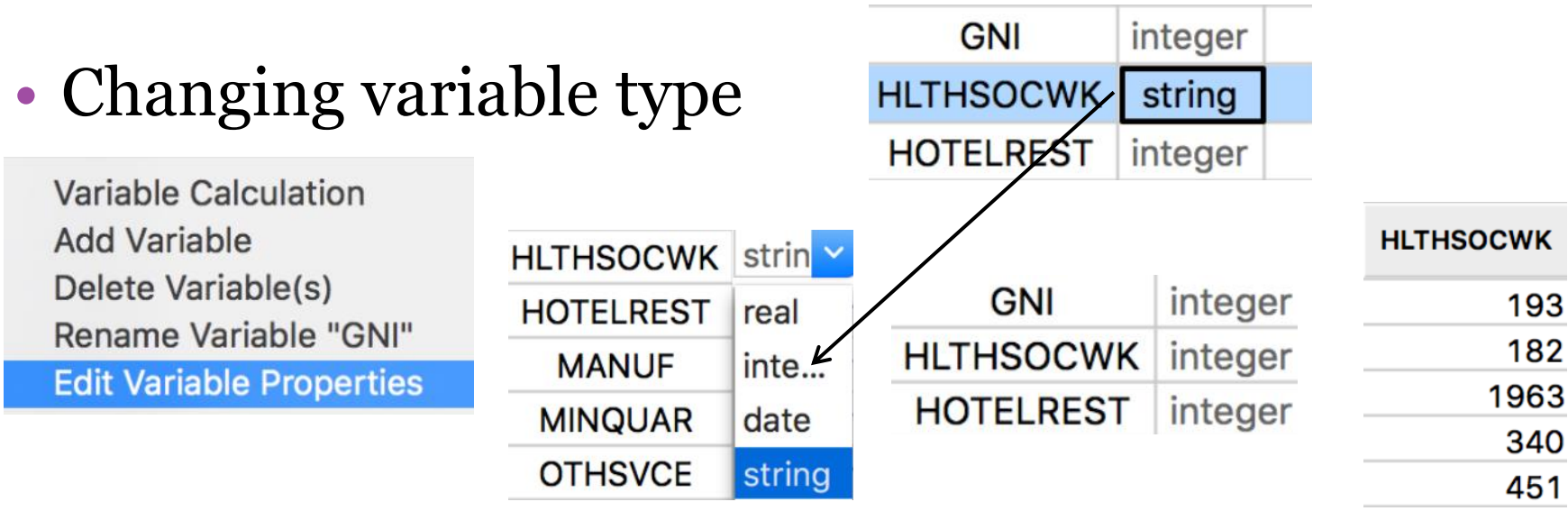
Big improvement over previous versions!

Data Cleanup

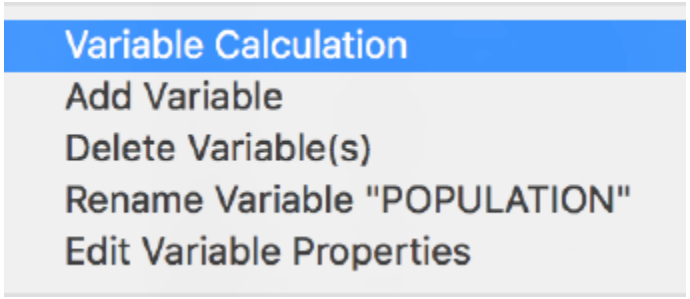
- Edit values in table



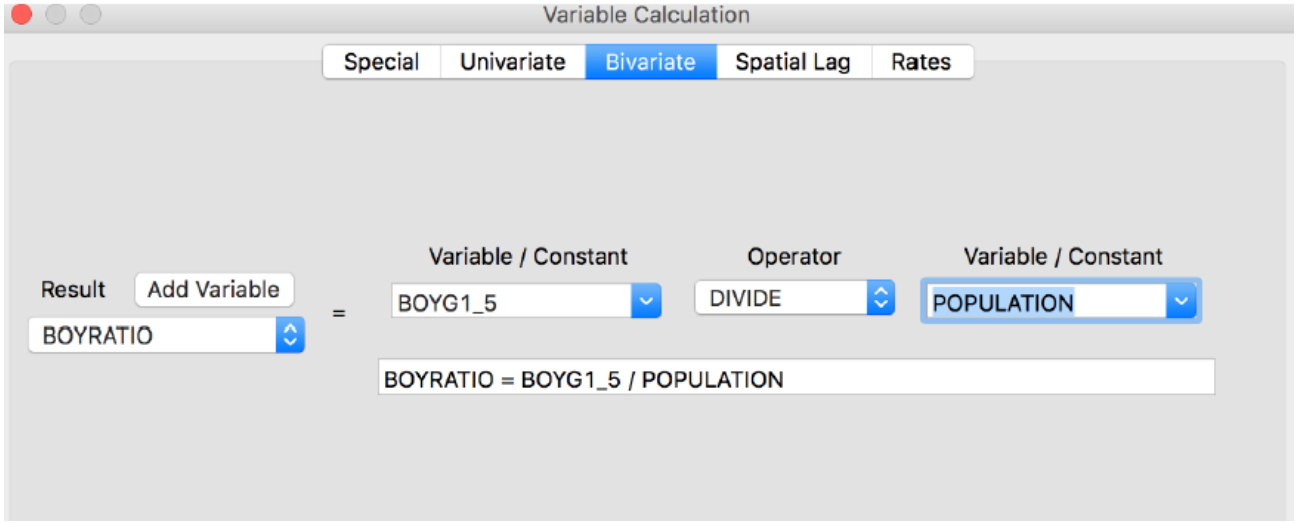
- Changing variable type



Variable Calculation

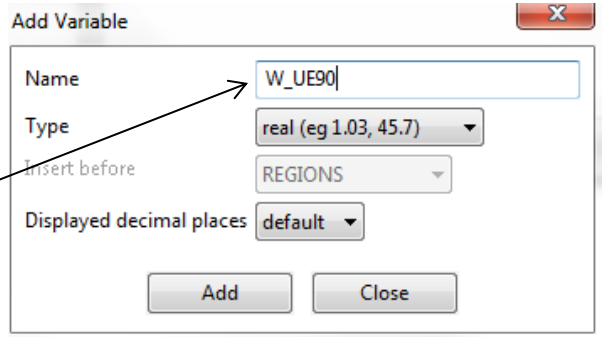


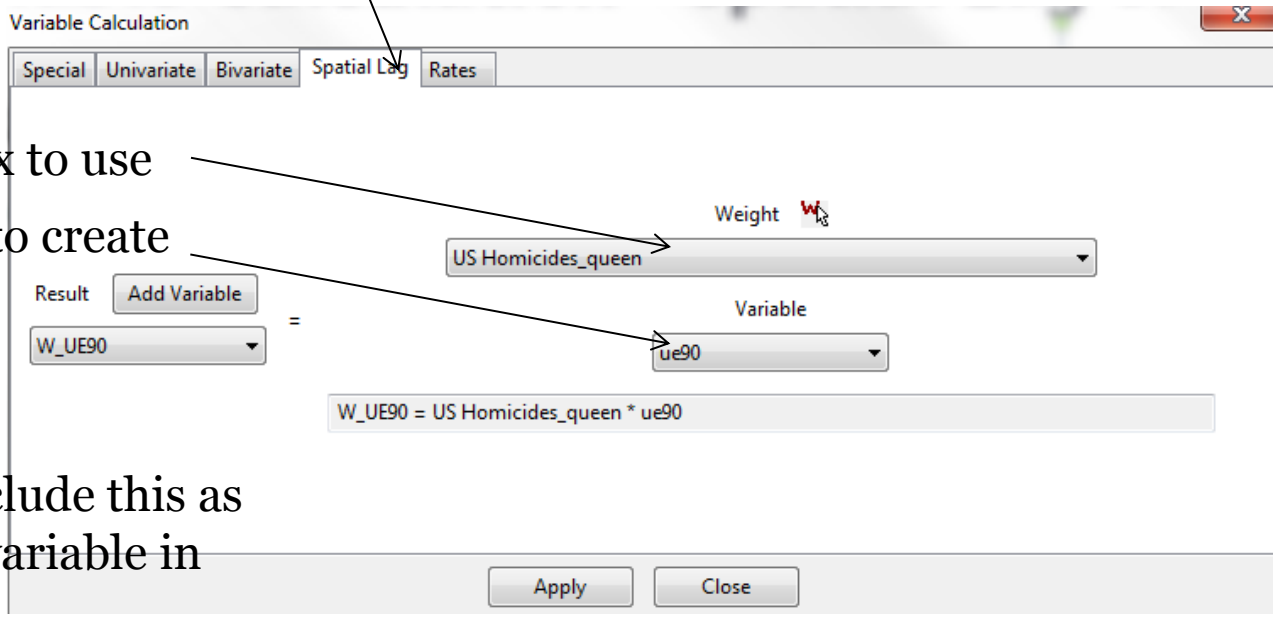
	BOYRATIO
1	0.116546
2	0.067674
3	0.082438
4	0.076907
5	0.055722



Variable Calculation – Spatial Lag of X

- Table – Variable Calculation
  - Add variable: name it W_*Var*
  - Choose 'Spatial Lag'



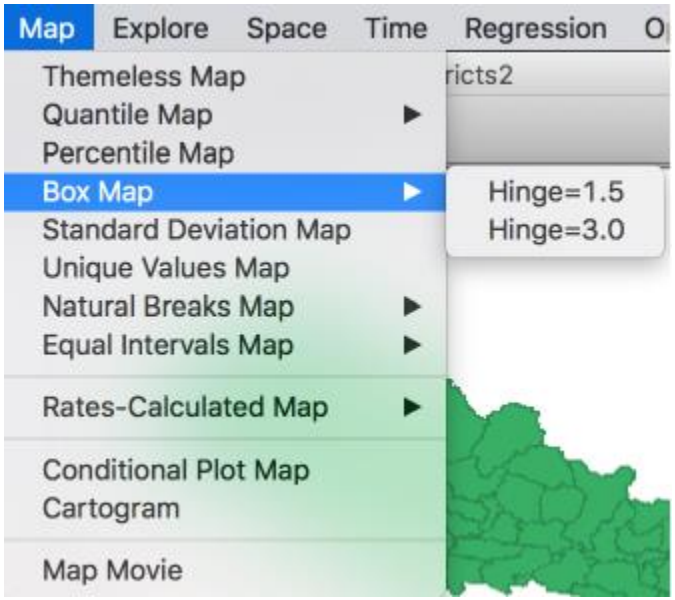


Choose weight matrix to use

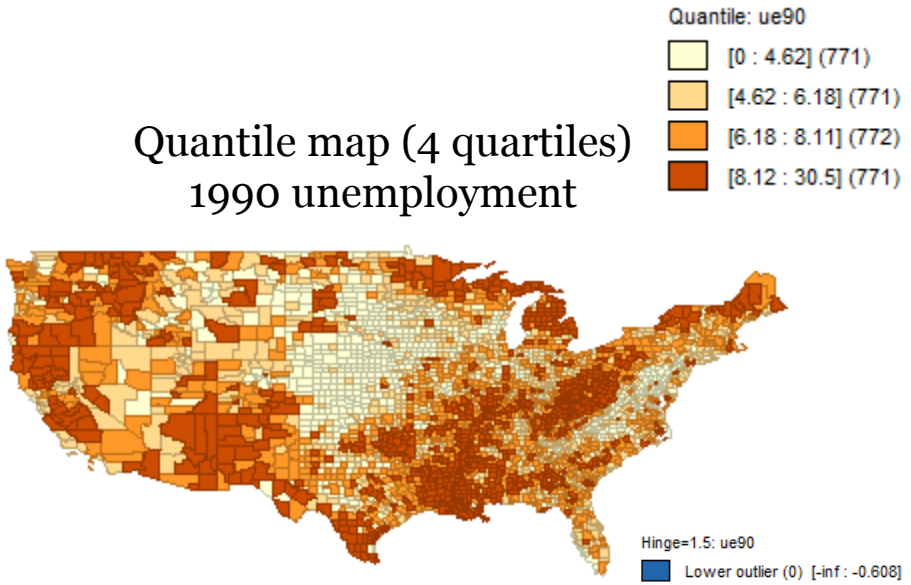
& variable you want to create a spatial lag of

You can then include this as an explanatory variable in regressions!

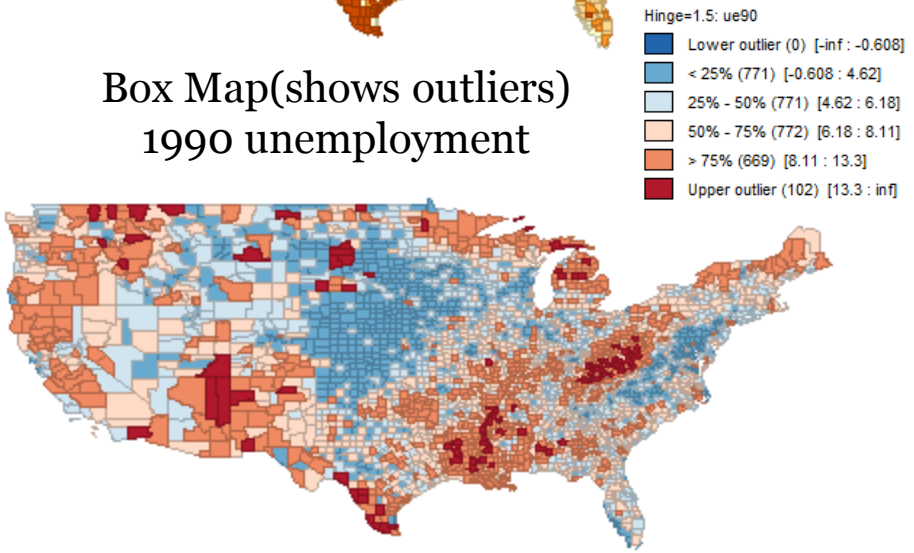
Opening a simple map



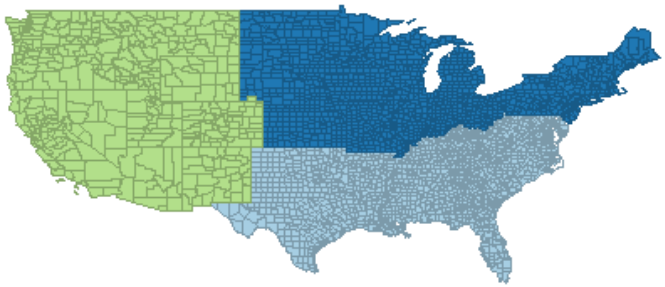
Quantile map (4 quartiles)
1990 unemployment



Box Map(shows outliers)
1990 unemployment

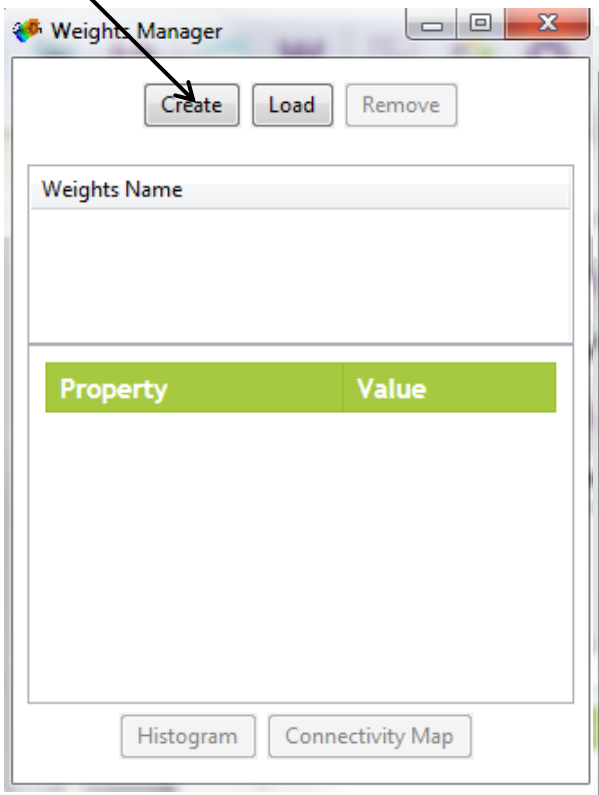


Unique Values Map
Regions (0,1,2)

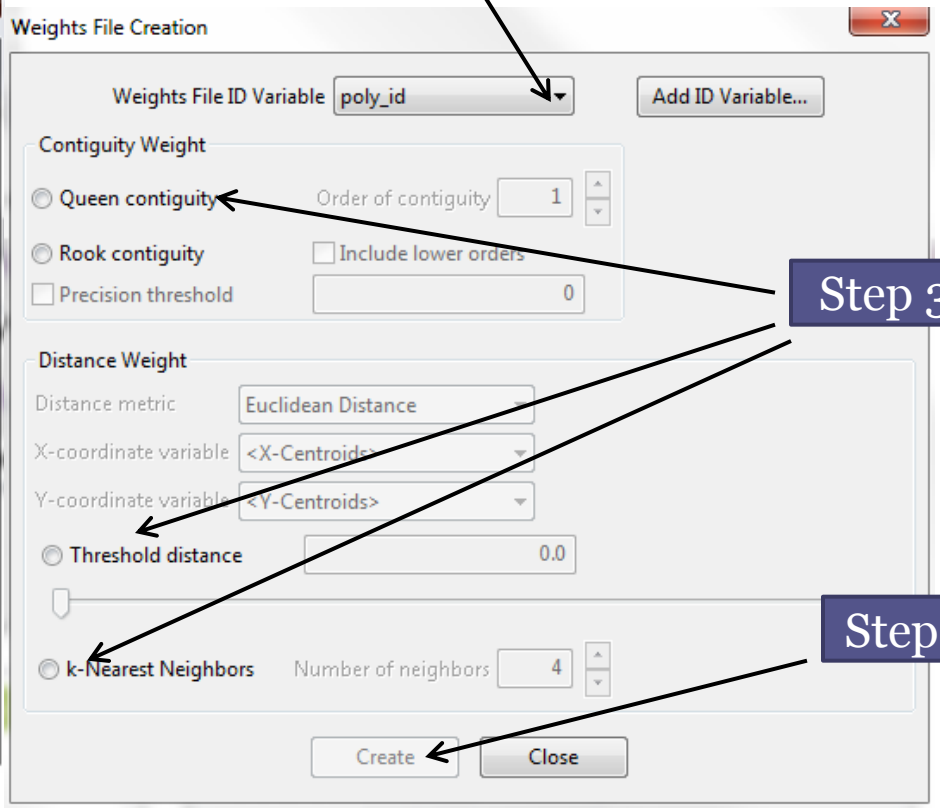


Constructing Spatial Weights

Step 1



Step 2

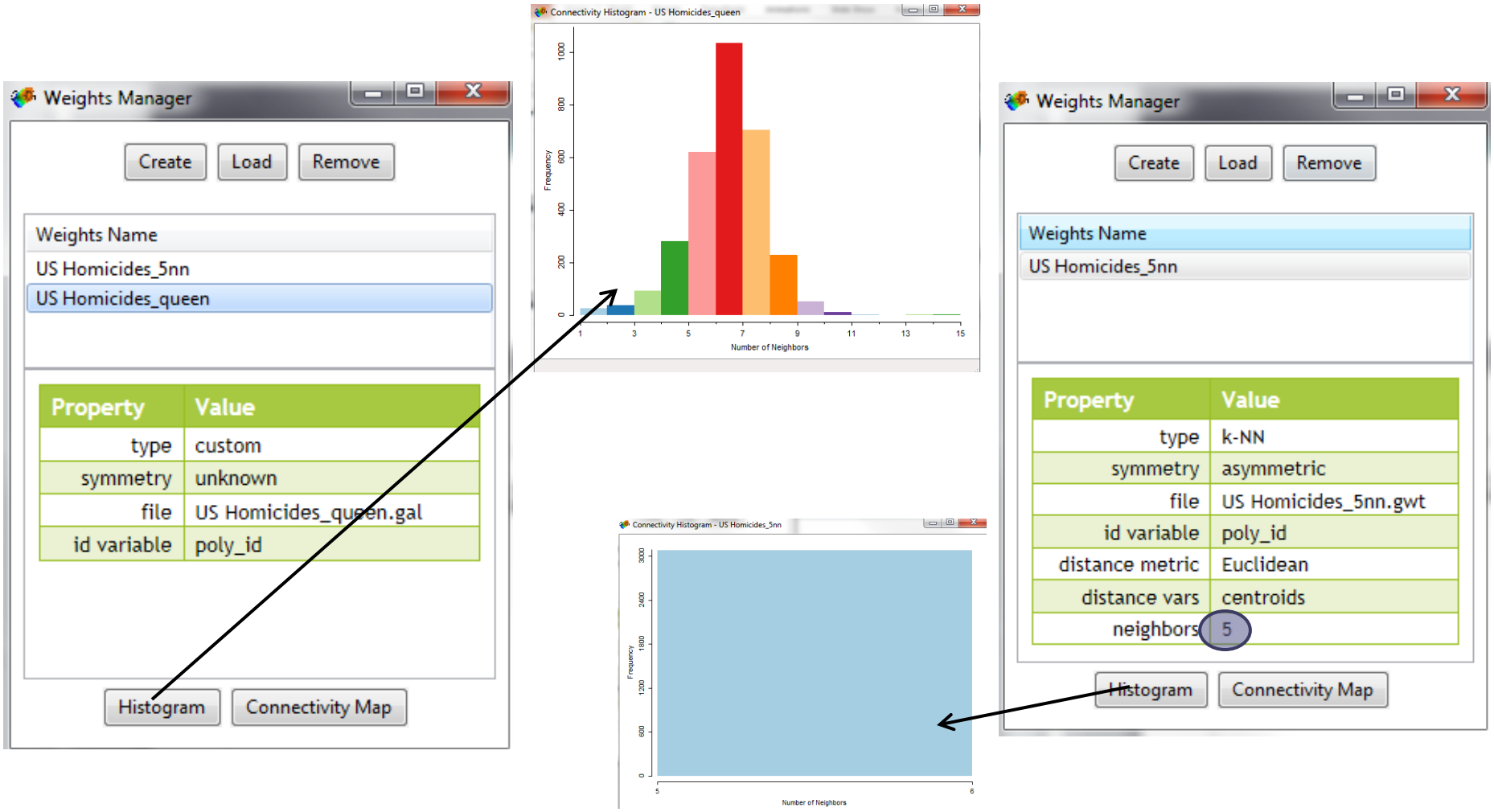


Step 3: Select 1

Step 4

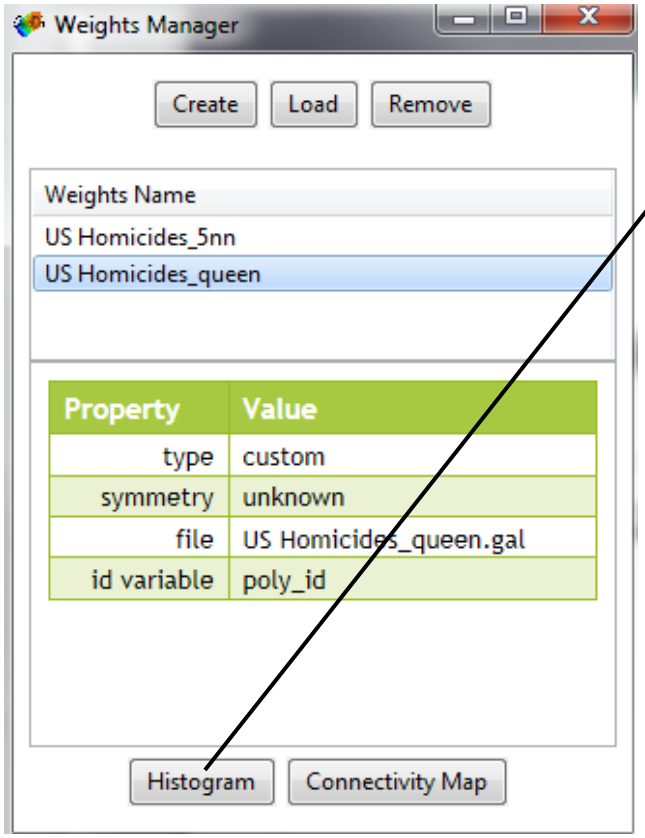
(saved as .gwt)

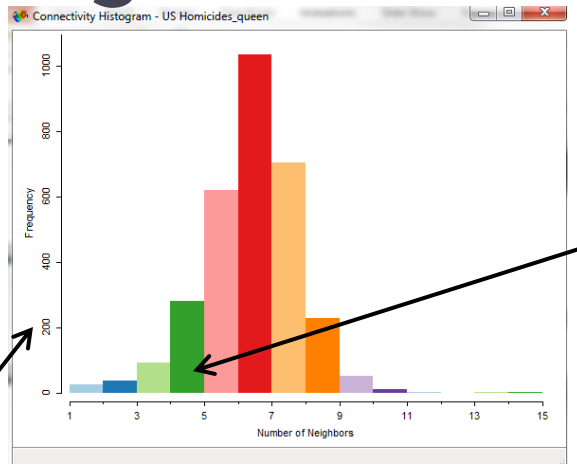
Histograms of neighbors



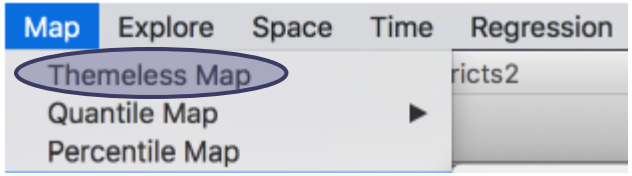
A Preview of the Coolest GeoDa Feature…Linking and Brushing

1. Open up your weights histogram

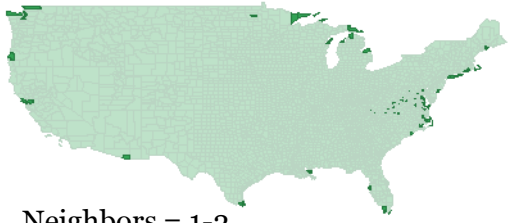




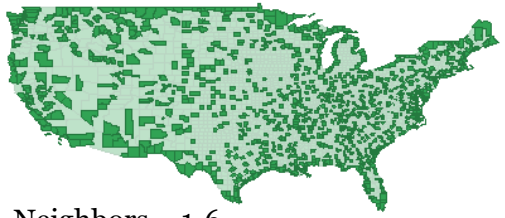
2. Select "Themeless Map" from main menu



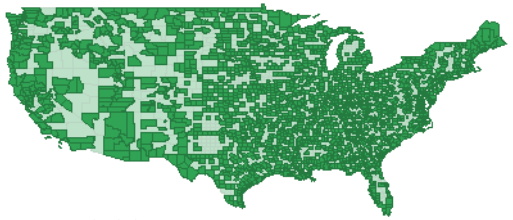
3. Watch the map change as you select points from your histogram!



Neighbors = 1-2



Neighbors = 1-6



Neighbors = 1-7

Assignment:

- Create the following weights for the US Homicide data:
  - Queen (1st order)
  - 5 nearest-neighbor
  - Distance-based (90 miles)

- Explore how your histograms interact with your maps

Moran's I

Measures GLOBAL spatial autocorrelation

- One of the oldest indicators of spatial autocorrelation (Moran, 1950)
- Compares the value of the variable at any one location with the value at all other locations

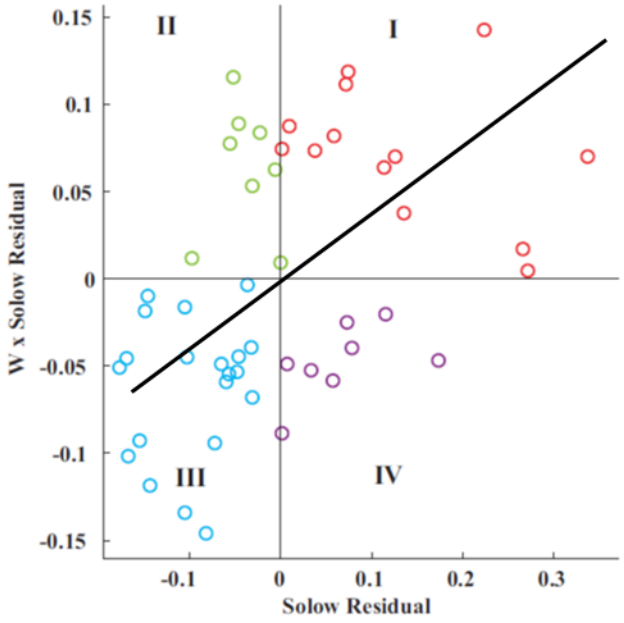
$$I = \frac{n}{\left(\sum_i \sum_j W_{i,j}\right)} \frac{\sum_i \sum_j W_{i,j}(X_i - \bar{X})(X_j - \bar{X})}{\sum_i (X_i - \overline{X})^2}$$

n = Number of cases
X_i is variable value at location i
X_j is variable value at location j
$\bar{X}$ is the mean of the variable
$W_{i,j}$ is the spatial weight applied to the relationship between i, j

Moran's I

- Recall the Moran Scatterplot:
  - X-axis: DV
  - Y-axis: Weighted neighboring values of DVs

The slope of this fitted line is the Moran's I statistic!

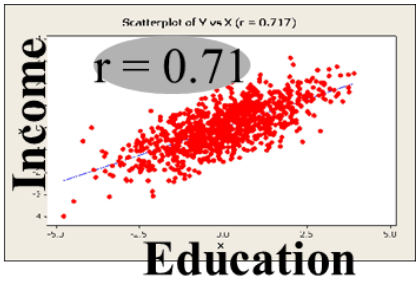


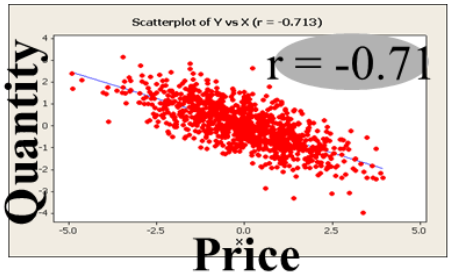
Moran's I

- Values Range from -1 (perfect dispersion) to +1 (perfect correlation)
  - Similar to correlation coefficient
  - High (and significant) I value indicates positive autocorrelation
- Value of 0 indicates a random spatial pattern
- Hypothesis testing performed via random permutation
  - Randomly arrange the values spatially 999 times, and calculate Moran's I (should be close to 0)
  - Compare actual I value to the 999 randomly generated I's
  - If the actual I falls into a statistically significant area (5%, 95%), it is significant at the 5% level

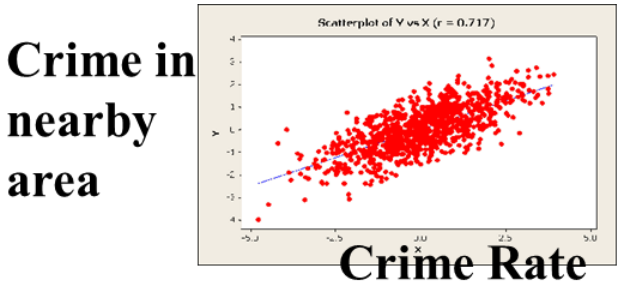
Moran's I and Correlation Coefficient *r*

- Correlation Coefficient *r*
  - Relationship between two variables





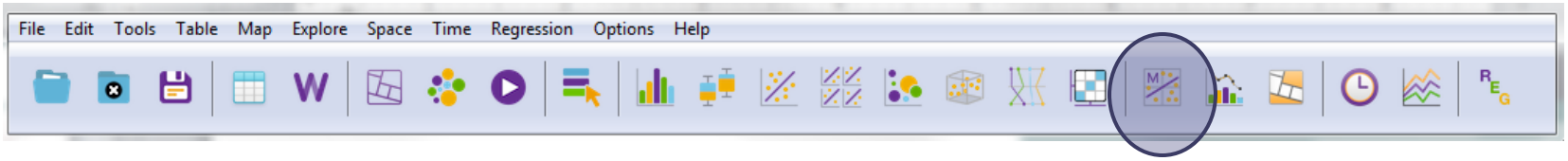
- Moran's I
  - Involves only one variable
  - Correlation between X and the spatial lag of X formed by averaging all the neighboring X values



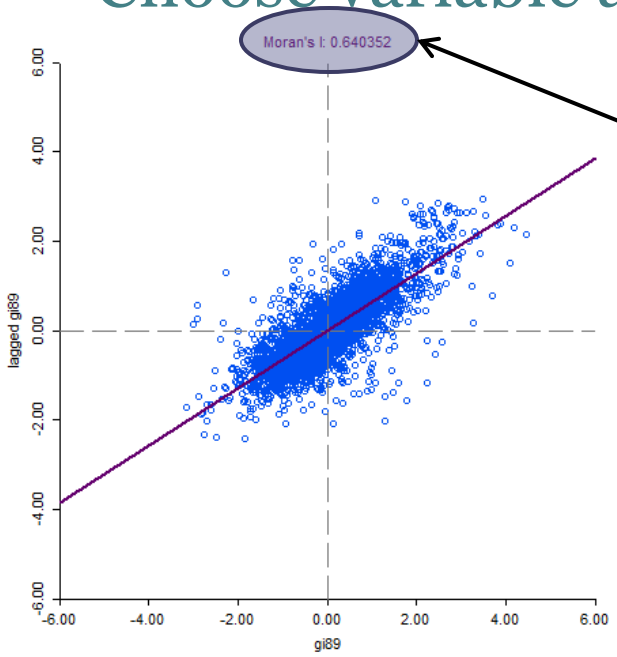
Moran's I

- Typically transformed to Z-scores, evaluated with p-values
  - The usual levels of significance:
    - 0.10*
    - 0.05**
    - 0.01***
- Indicator of GLOBAL spatial autocorrelation
  - LISA is an indicator of local spatial autocorrelation

Moran's I in GeoDa



- Or, Space-> Univariate Moran's I
  - Choose variable and weight matrix



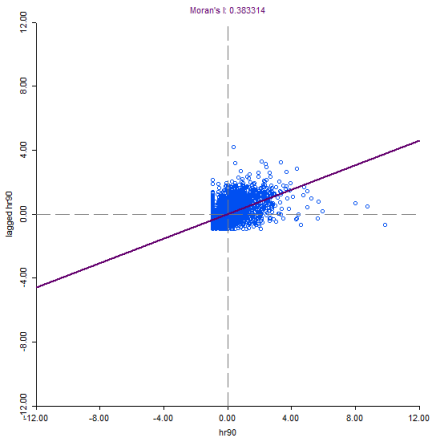
Moran's I value

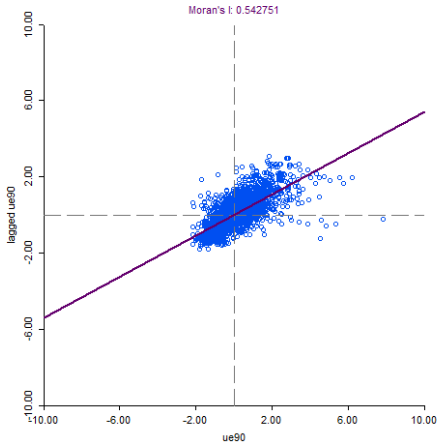
To find p-value:
Right-click
Select "randomization"
and # of permutations

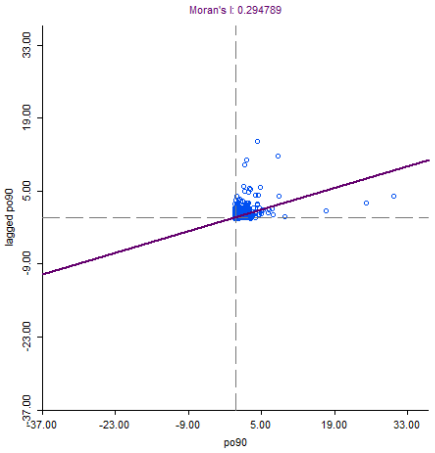
Resulting p-value:

Assignment

- Explore Moran's I for variables in your US Homicide data
  - Which has the biggest value? Lowest?
  - What happens when you change the weight matrix being used?







Local Moran's I

- Local Indicator of Spatial Association (LISA)
  - The LISA for each observation gives an indication of significant spatial clustering of similar values around that observation.
  - The sum of LISAs for all observations is proportional to a global indicator of spatial association (Moran's I)
- $I_i = z_i \sum_j W_{ij} Z_j$
- z_i is the standardized version of the original x_i
- The summation $\sum_j$ is across each row i of the spatial weights matrix
- $I = \sum_i \frac{I_i}{N}$ (Global Measure)
- Can still find local clusters even if there is no global clustering!

Local Moran's I

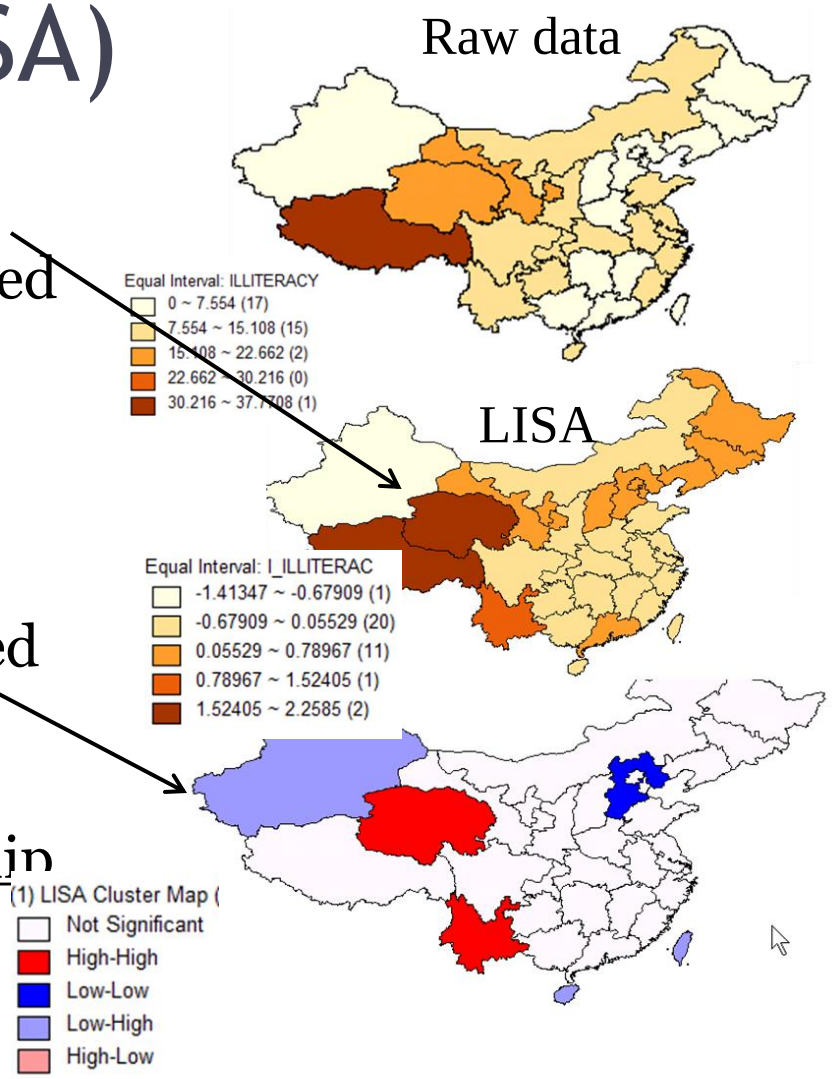
- The statistic is calculated for each area in the data (i.e. county, ZIP code, country)

- For each area, the index is calculated <u>based on neighboring areas with which it shares a border</u> (according to the Weight being used)

Local Moran's I (LISA)

- Since a measure is available for each polygon, these can be mapped to indicate how spatial autocorrelation varies over the study region

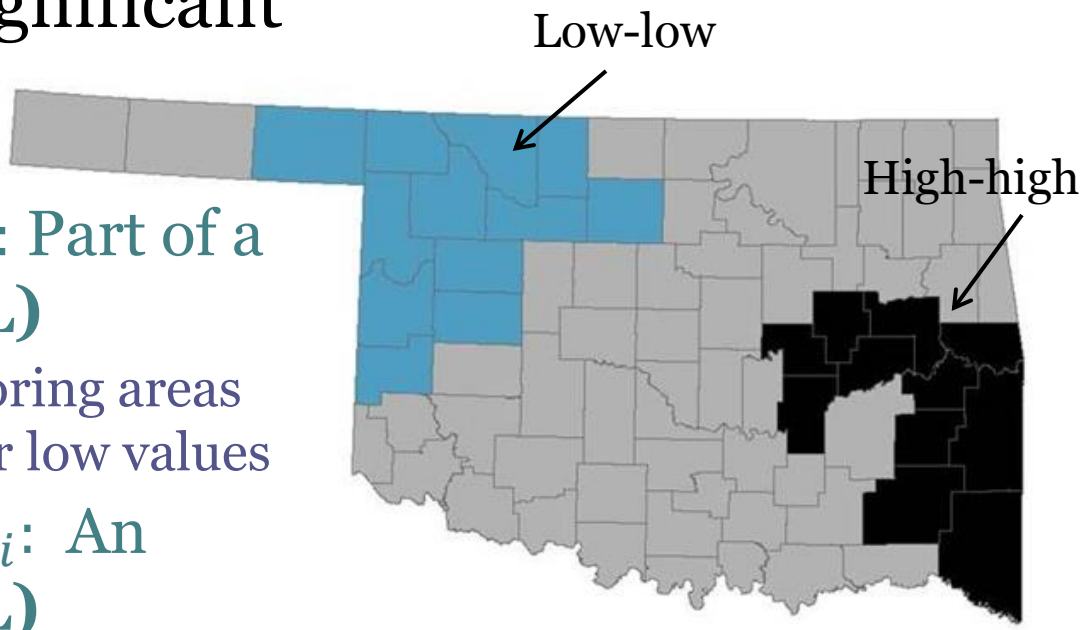
- Since each index has an associated test statistic, we can also map which of the polygons has a statistically significant relationship with its neighbors, and show the type of relationship



Interpreting LISAs

- Resulting map shows where H-H, L-L, L-H, and H-L clusters are significant

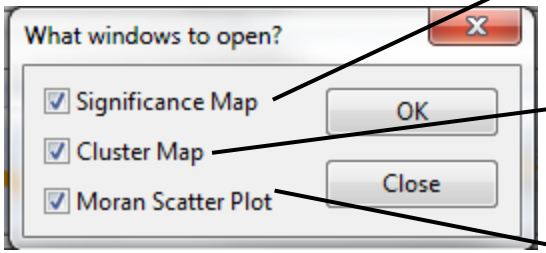
- Interpretation:
  - Positive value for I_i: Part of a **cluster (H-H, L-L)**
    - The area has neighboring areas with similarly high or low values
  - Negative value for I_i: An **outlier (L-H, H-L)**
    - The area has neighboring areas with dissimilar values

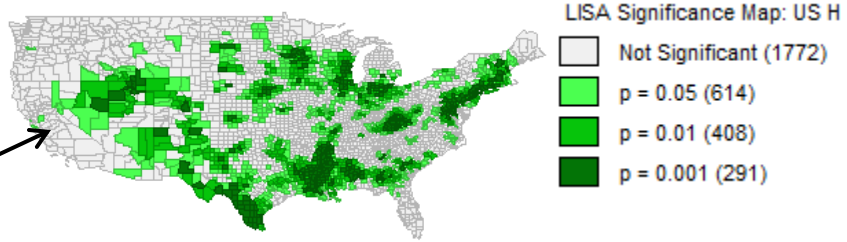


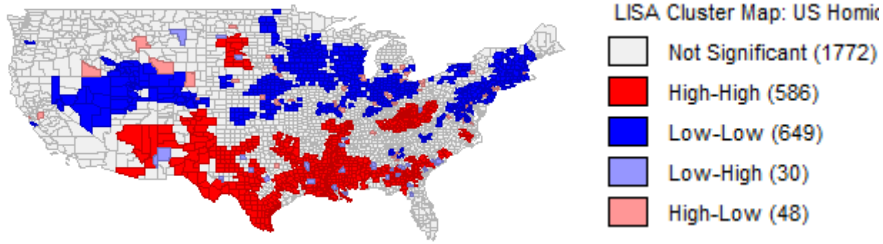
2015 Oklahoma unemployment rates

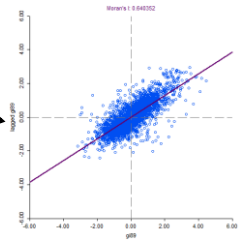
LISAs in GeoDa

- Space – Univariate Local Moran's I (or)
  - Select variable & weight
  - Select all 3 types of plots:



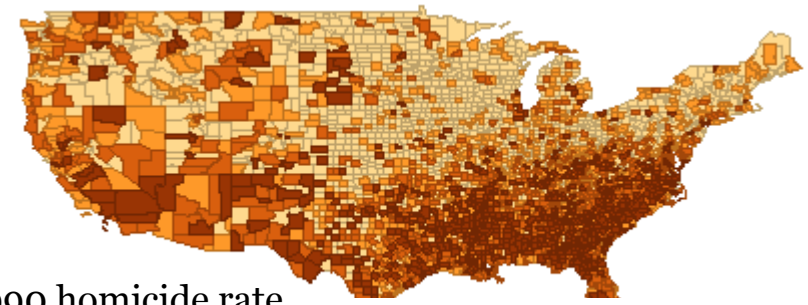




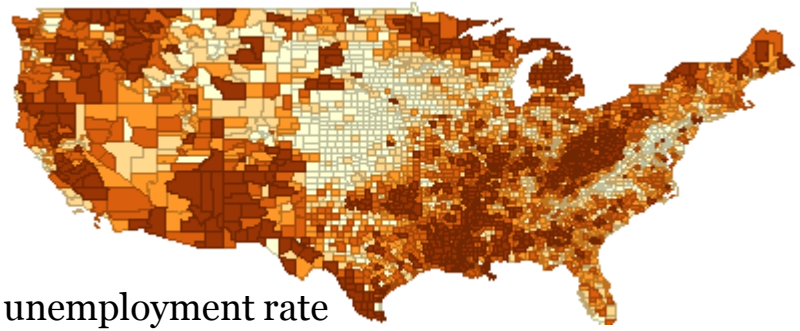


Assignment

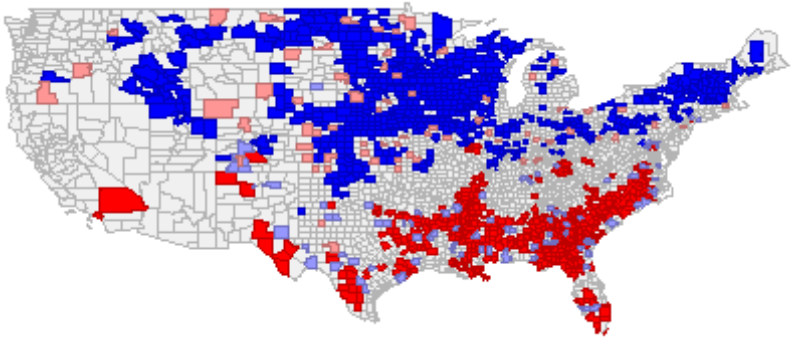
- Explore LISAs for variables in your US Homicide data (& compare to original maps)
  - Identify hotspots for homicide, unemployment, poverty,…
  - Impact of spatial weight matrix used?

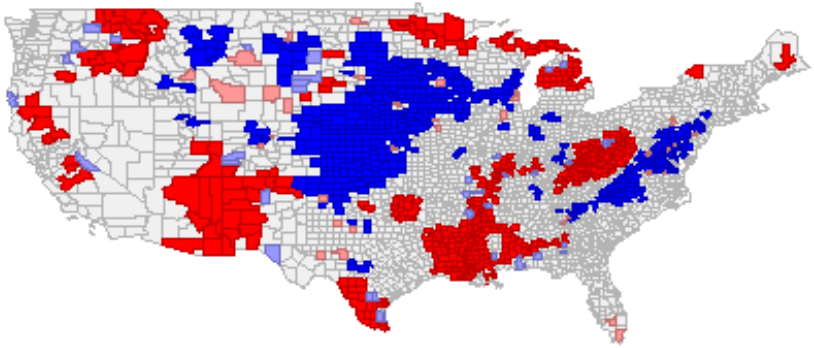


1990 homicide rate



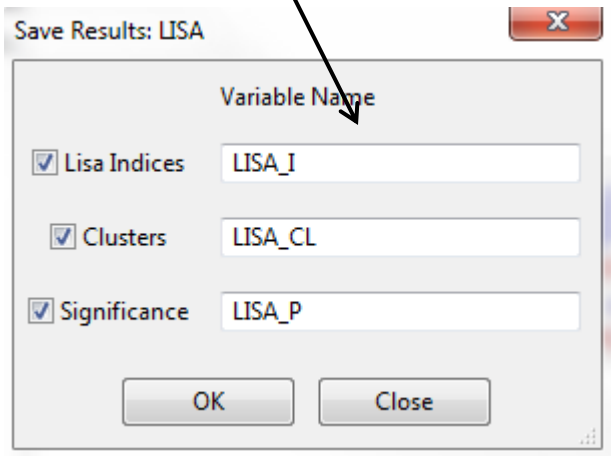
1990 unemployment rate





Assignment Hints

Right click on Cluster Map:
- Randomization
- Significance Filter
  - p = 0.05 vs. 0.01
- Save Results



Then view in Data Table

Table - US Homicides

	fh70	fh80	fh90	west	LISA_I_FP	LISA_CL_FP	LISA_P_FP
2425	15.700000	19.624381	23.053218	0	2.3305328	1	0.0020000
2426	8.900000	10.186263	12.851527	0	4.2129123	1	0.0010000
2427	20.200000	28.348214	39.375542	0	13.3025263	1	0.0010000
2428	12.300000	16.591760	21.307331	0	-0.0107965	0	0.1500000
2429	7.500000	8.837614	11.651875	0	-0.0202563	0	0.4600000
2430	11.400000	17.701691	19.431664	0	-0.1025895	0	0.1800000
2431	5.500000	6.319558	8.948460	0	0.2573644	0	0.1670000
2432	9.300000	12.645503	14.688926	0	0.4246301	2	0.0340000
2433	11.300000	15.762218	18.508026	0	0.3161432	2	0.0050000
2434	7.200000	7.193541	7.658341	0	0.5990382	0	0.1330000
2435	18.900000	24.253920	32.356656	0	5.8691698	1	0.0010000
2436	7.200000	6.366048	9.241095	0	-0.0252398	0	0.3810000
2437	9.200000	9.583936	12.994901	1	0.2446608	0	0.1050000
2438	4.600000	7.992895	9.684440	0	-0.0524331	0	0.1540000
2439	7.000000	8.315965	9.903804	0	0.0886941	0	0.3070000
2440	6.000000	5.011933	10.473458	0	0.1909003	0	0.2120000
2441	6.600000	8.370385	9.228223	0	0.0404265	0	0.1420000
2442	3.100000	2.723735	4.602511	0	-0.3155857	3	0.0200000
2443	7.800000	6.738980	12.069864	0	1.0930517	0	0.0590000
2444	9.000000	10.329289	11.686794	0	0.0887444	0	0.3180000

L-L Cluster

H-H Cluster

L-H Outlier

BREAK

3. Exploratory Spatial Data Analysis (ESDA)

- ESDA Basics – why we explore data

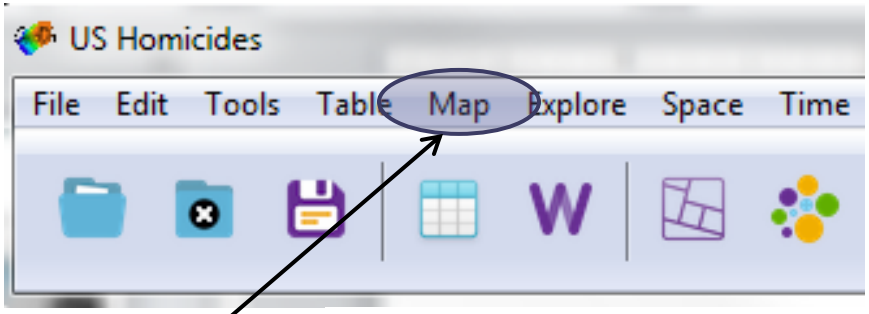
- GeoDa Tools
  - Basic Maps
  - Histograms
  - Box Plots
  - Scatterplots
  - Space / Time Mapping

ESDA Basics

- Why We Explore Our Data
  - General understanding
  - Hypothesis formulation
  - Suitability for inclusion in statistical analysis

- Why Does Clustering Matter?
  - Evidence of a spatial process at work
    - Evidence of clustering can support many hypotheses about what is happening in your data
  - Can indicate potential problems for statistical analysis
    - We will cover this when we discuss spatial regression

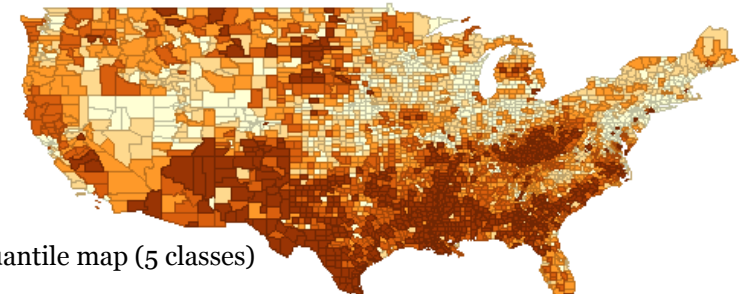
Simple Maps



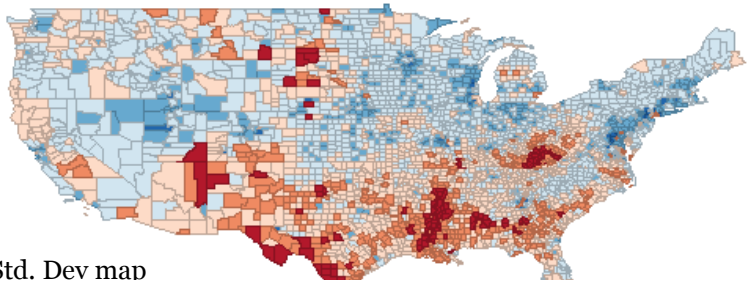
- Lots of choices!
- Map -> Quantile Map
  - Select variable
  - Select # of classes
- Map -> Std. Deviation Map
  - Select variable
  - Select # of classes
- Natural Breaks, Box,…

Right click on map:
"Change Current Map Type"

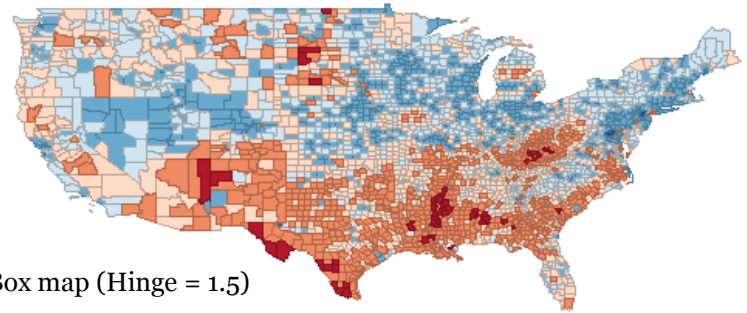
Maps of gi89



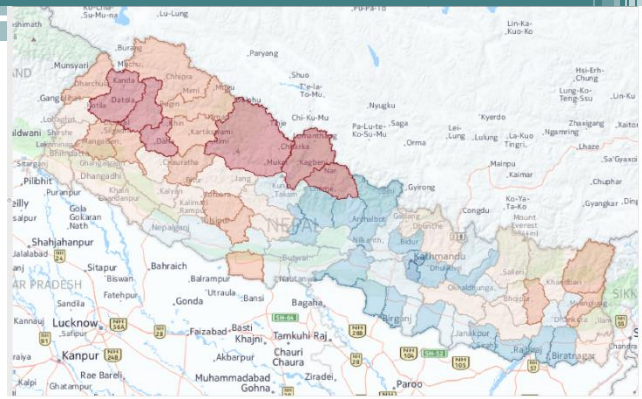
Quantile map (5 classes)



Std. Dev map

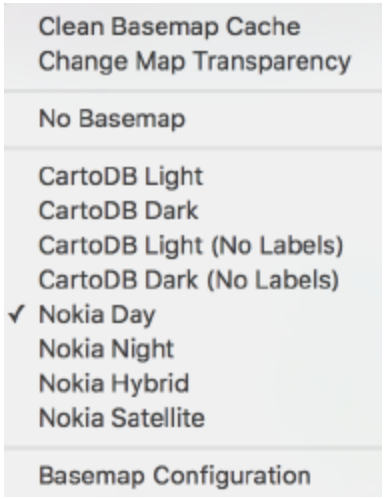


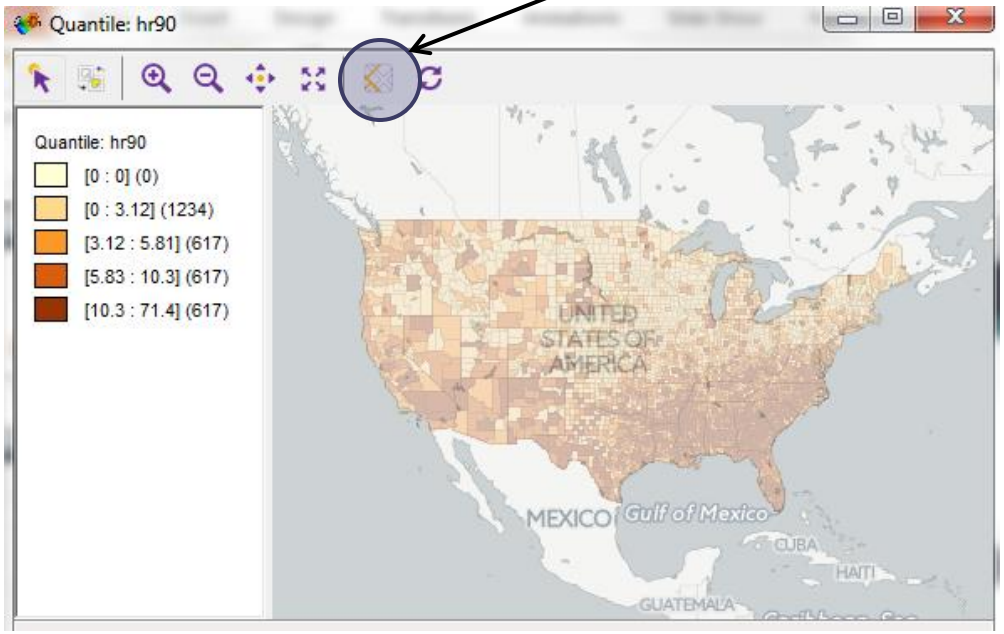
Box map (Hinge = 1.5)



A Pretty Cool Upgrade…

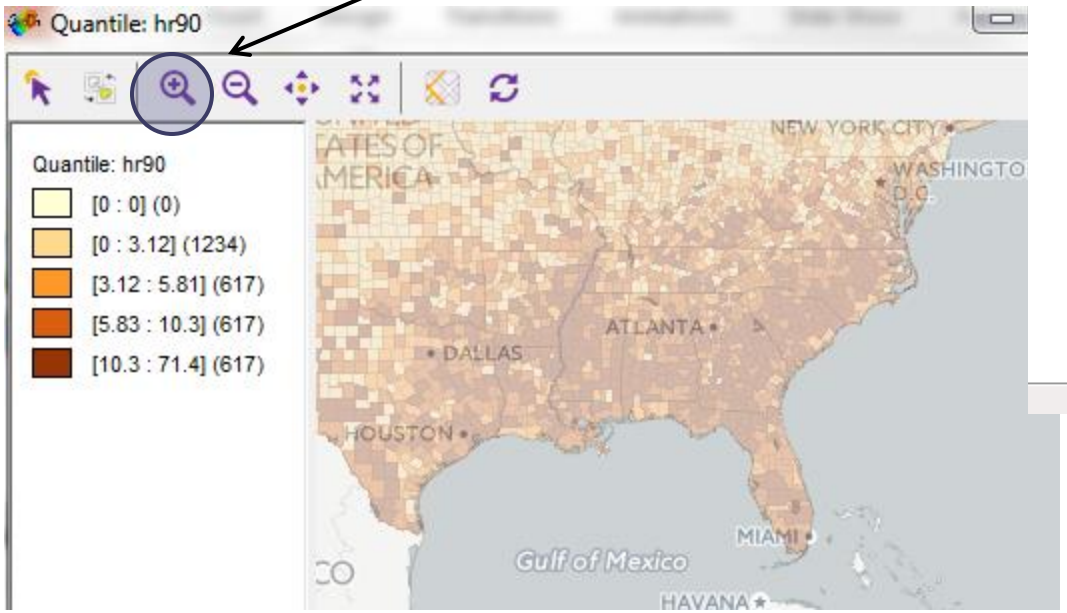
- Ability to add realistic "Basemap" behind the data (mapping tiles from Nokia)
  - Requires Internet connection

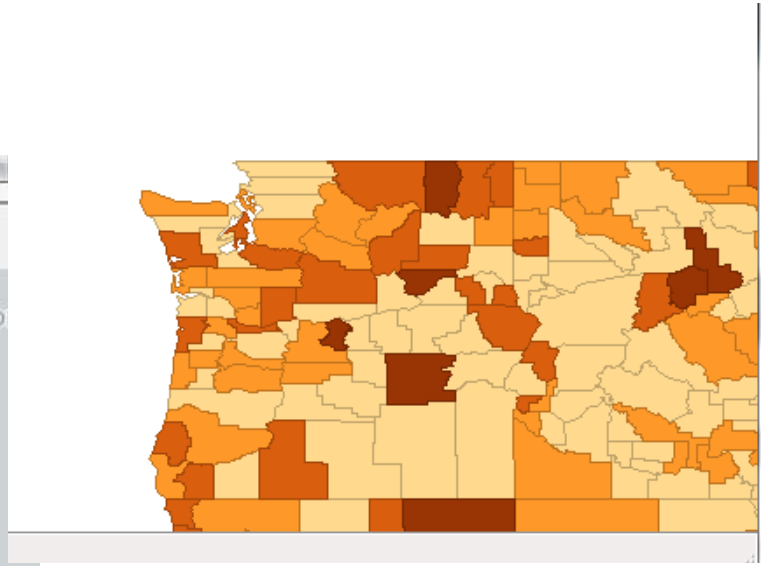




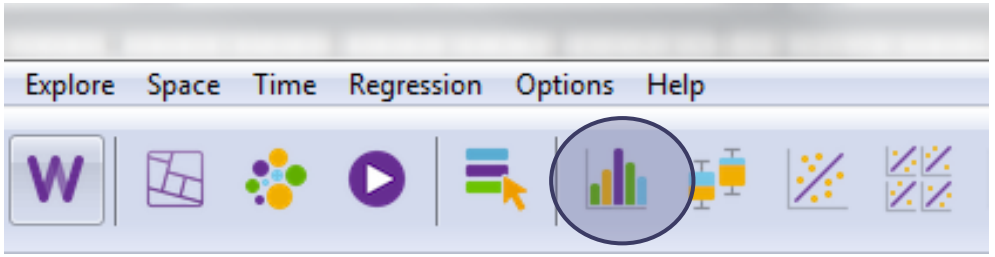
And more map upgrades

- Ability to zoom in / out

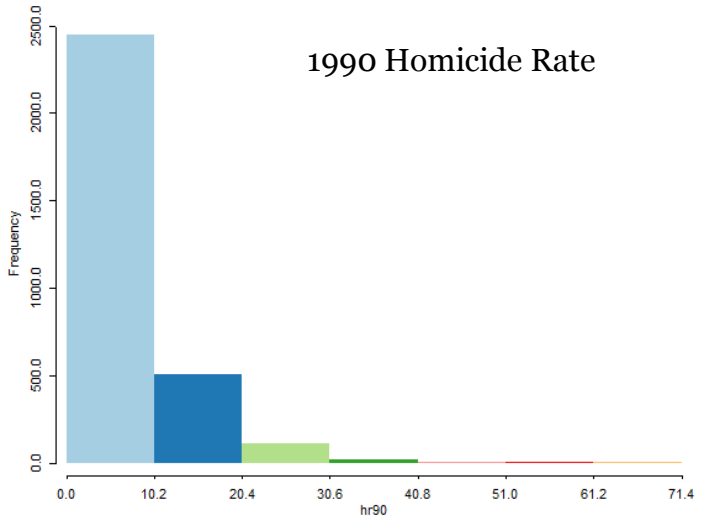


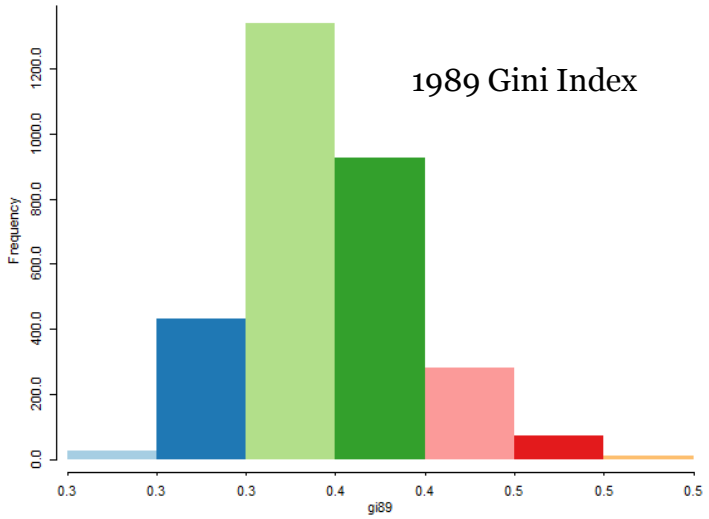


Histograms

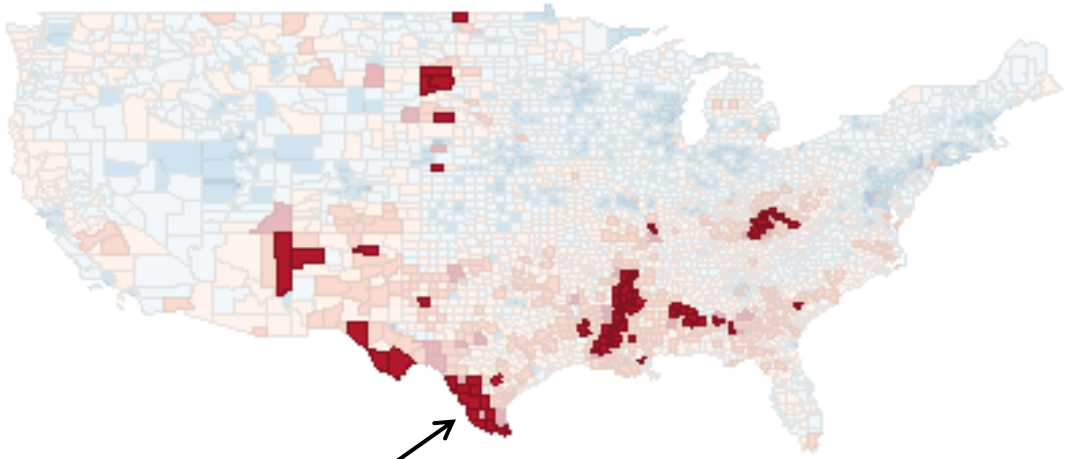


- Explore -> Histogram -> Choose variable
  - Frequency of outcomes is shown
  - Can choose # of intervals (right click)
  - Useful for comparing distributions across variables



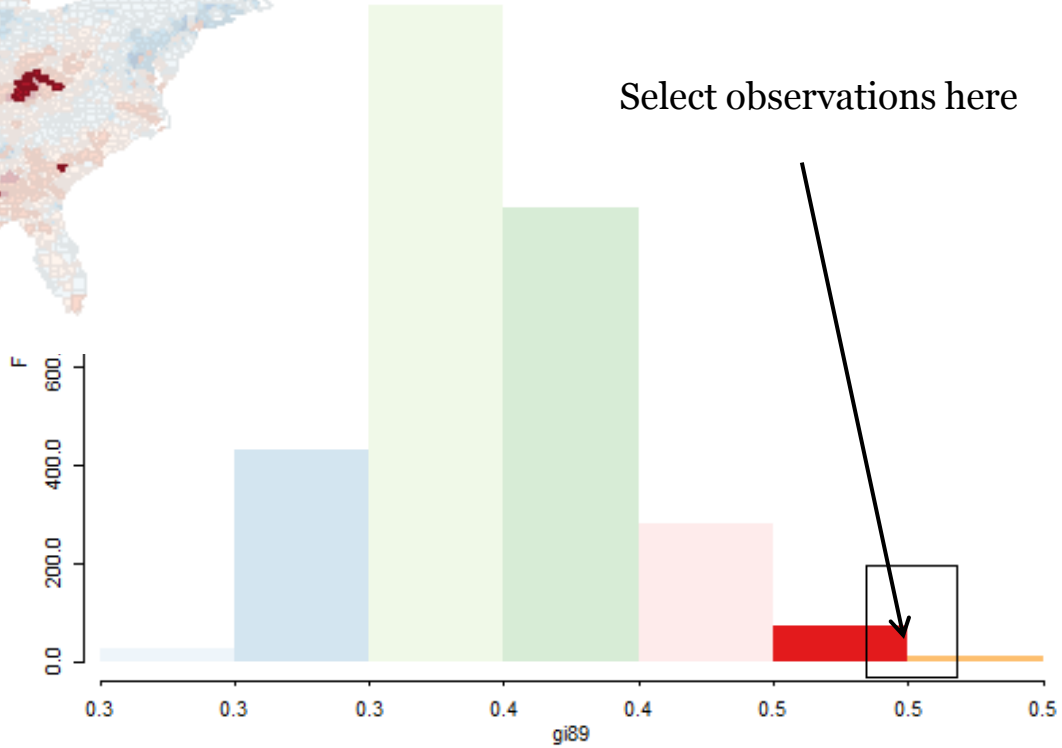


And remember…brushing / linking

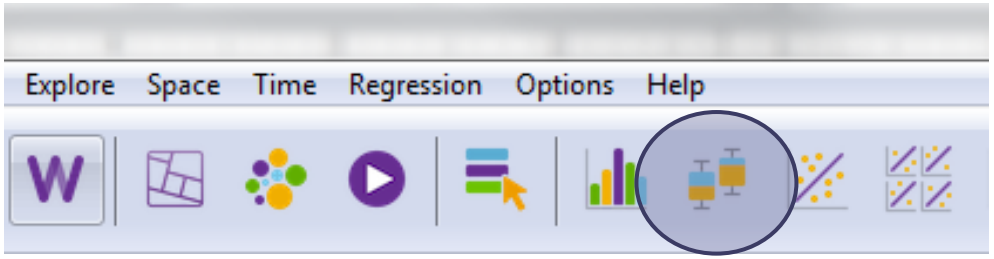


…and they show up here

Select observations here

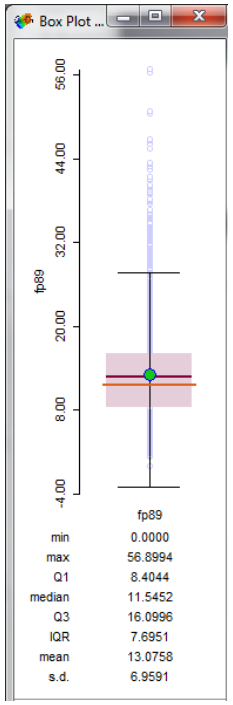


Box Plots

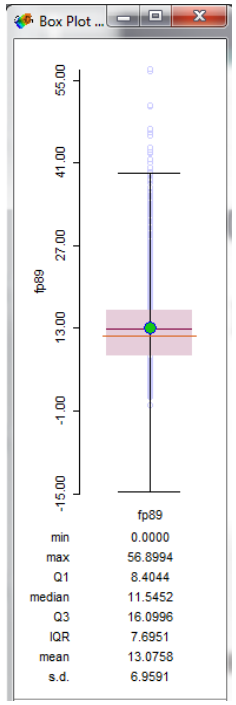


- Explore -> Box Plot -> Choose Variable
  - All observations displayed
  - Bounds shown are 1.5 (or 3.0) times the 25th and 75th quartile
  - Useful for detecting potential outliers

Hinge = 1.5

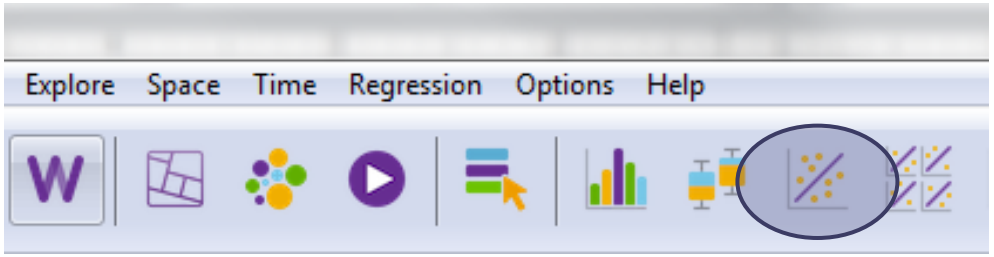


Hinge = 3.0

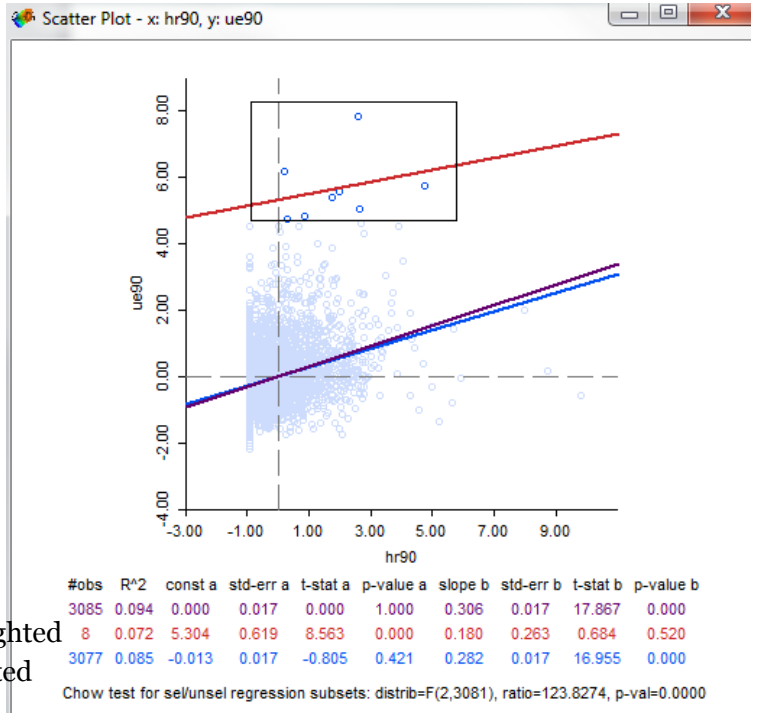


Again, explore linking / brushing!

Scatterplots



- Explore -> Scatterplots
  - Select IV (X-axis) and DV (Y-axis)
- Right-click on scatterplot
  - Data -> View standardized data
  - Axes are now standard deviations
- When you select observations, 3 sets of fit statistics are shown

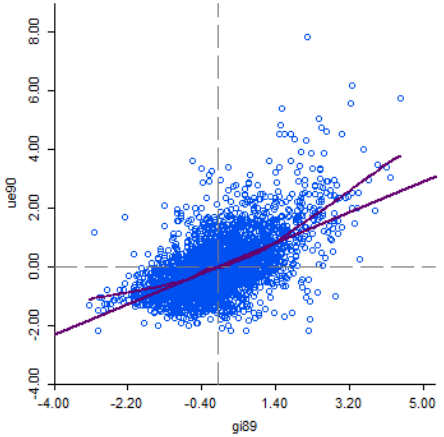


All
Only those highlighted
Without highlighted

Helps assess impacts of outliers

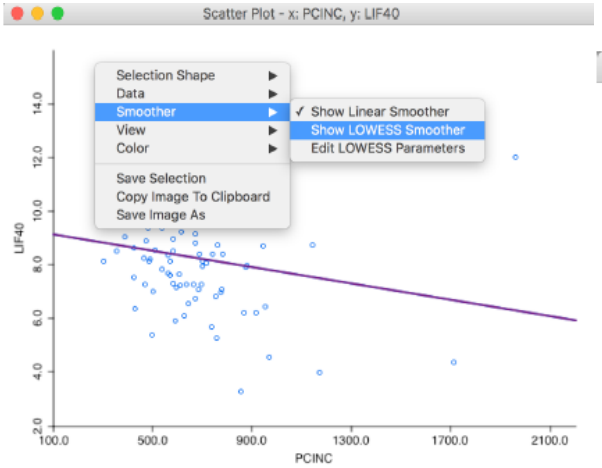
Scatterplots

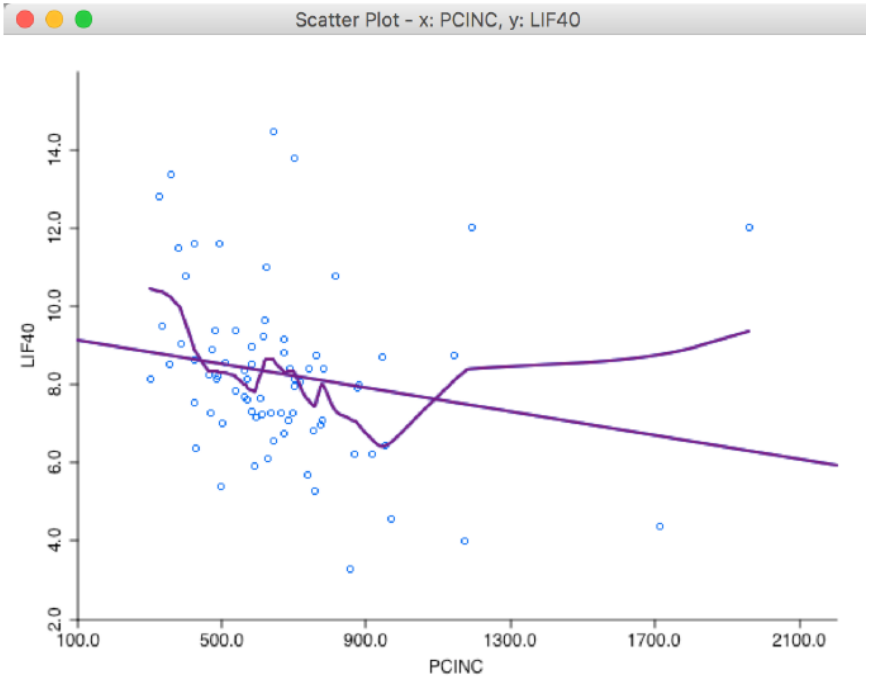
- Can also run "Lowess Smoother"



#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
3085	0.334	0.000	0.015	0.000	1.000	0.578	0.015	39.335	0.000
0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3085	0.334	0.000	0.015	0.000	1.000	0.578	0.015	39.335	0.000

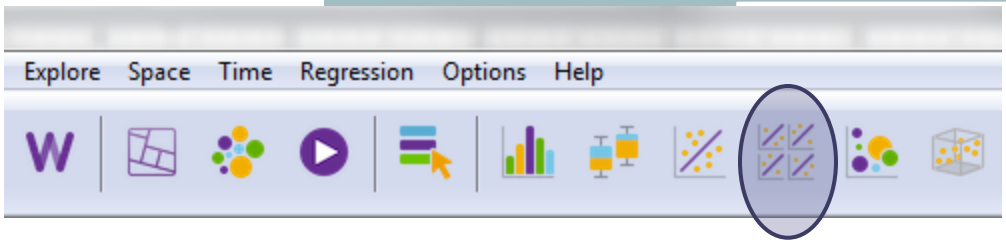
Chow test for sel/unsel regression subsets: need two valid regressions

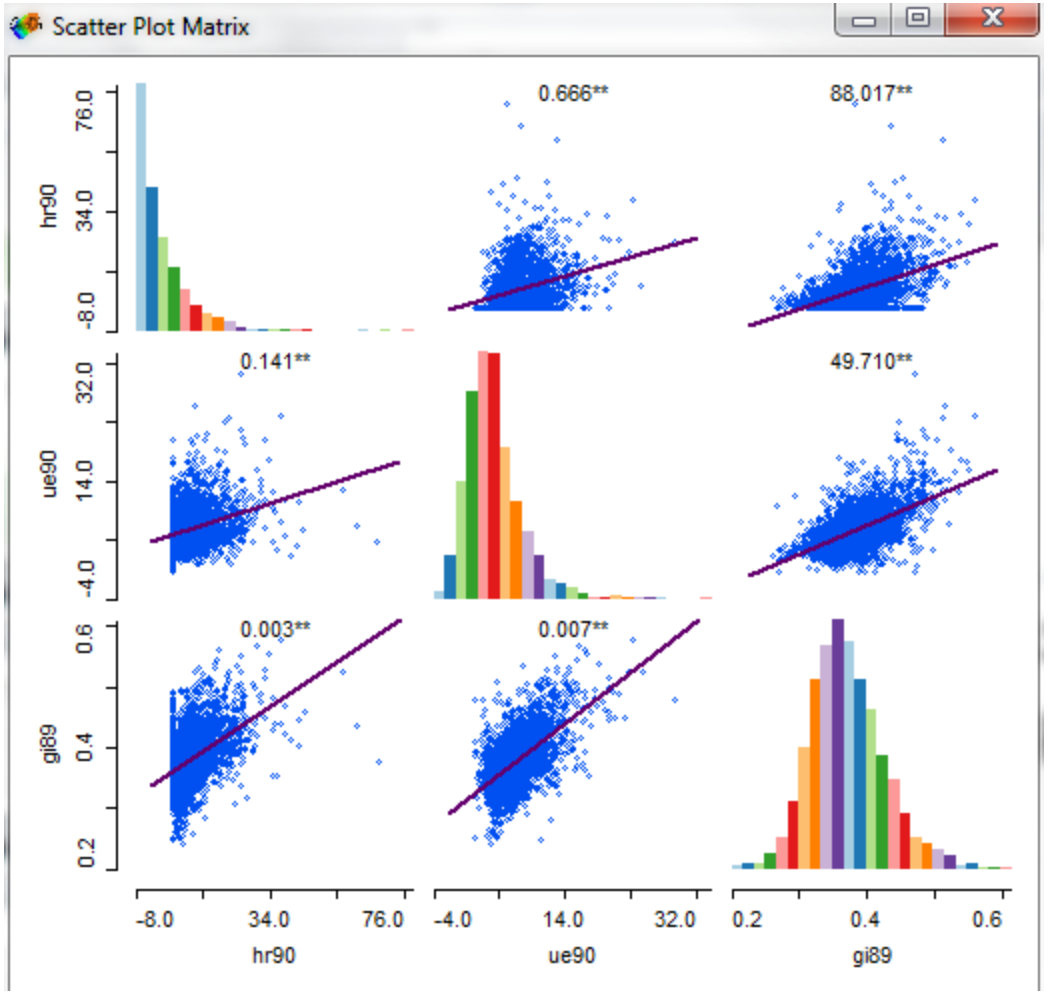




Scatterplots

- Can also create scatterplot matrix (with histogram on diagonal)
  - Can also standardize
  - Linking / brushing very useful here





Assignment

- Become familiar with ESDA tools by using them to:
  - 1) Explore the US Homicide data
    - Quantile, Std Dev, Box Maps
    - Histograms
    - Box Plots
    - Scatterplots
    - Brushing / linking!

Variables to explore:
hr90 – Homicide Rate (1990)
rd90 – Resource Deprivation
ue90 – Unemployment rate
dv90 – Divorce rate
gi89 – Gini index (income ineq)

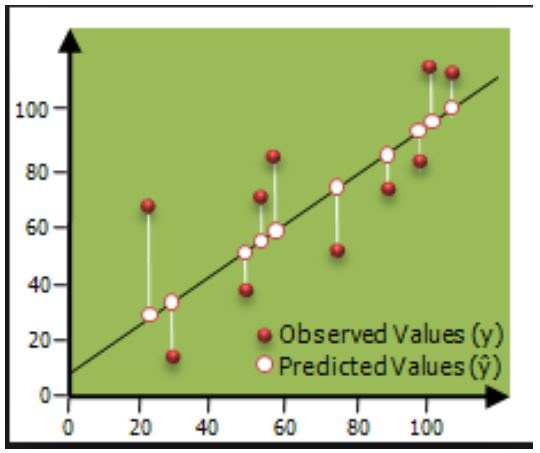
  - 2) Complete the worksheet on the Mississippi Police data (county-level)

BREAK

4. Spatial Regression

- Review of non-spatial linear regression (OLS)
- Spatial regression models
  - Spatial lag
  - Spatial error
- Using GeoDa
  - OLS
  - Spatial lag
  - Spatial error
  - How to decide which to use?

Basics of OLS



- $y = \beta X + \varepsilon$
- Essentially finding the line that minimizes the total squared distance from that line to the observed values
- Assumptions required:
  - Linear relationship
  - Variables are mean independent
  - Disturbances are normally distributed

Why Not OLS??

- Spatial dependence and heterogeneity often (if not always) violate the statistical assumptions used in traditional OLS (LeSage and Pace, 2009)
  - Independence
  - Constant Variance
- Presence of spatial autocorrelation or heteroskedasticity violates *iid* assumption in standard OLS
- Spatial regression is arguably the most common approach to spatial dependence and heterogeneity (to some extent)

How Do We Know?

- Is OLS appropriate?
- ESDA is key
  - Consider structural covariates of dependent variable
  - Visually inspect maps / plots for outliers / clusters
  - Global / local tests for spatial autocorrelation
  - Extent of spatial heterogeneity?

Types of Spatial Processes

- **Spatial dependence**
  - Similarity (or difference) of nearby observations – linked to proximity through an active process
  - Functional relationship between what happens at one point and what happens at another
- **Spatial heterogeneity**
  - Similarity of nearby observations acting on a region larger than a single observation
  - Mean / variance / covariance "drifts" over geographies
  - Undesirable in regression analysis (stationarity needed to reduce number of parameters)

Examples of Spatial Heterogeneity vs. Dependence

- Sociology
  - Heterogeneity: history of auto manufacturing links surrounding counties to Detroit
  - Dependence: auto plants closing impacts all employees, regardless of county where they live
- Ecology
  - Heterogeneity: dandelion prevalence resembles city-wide rate due to shared climate, season,…
  - Dependence: dandelions in one yard scatter seeds that increase dandelions in neighboring yards

What is Spatial Regression?

- Regression techniques that explicitly include spatial information related to our observations
- 3 main ways of introducing this info:
  - Spatially lagged independent variable
  - Spatially lagged dependent variable
  - Spatially lagged error term
- Do we even need to include a spatial term in our models?

Spatial Regression: Nuts and Bolts

- There are 2 main types of spatial regression:
  - Spatial lag
    - Estimates a 'spatial' coefficient similar to the other independent variables
    - Appropriate when spatial *dependence* is the issue
  - Spatial error
    - Estimates a 'spatial' coefficient within the error term
    - Appropriate when spatial *heterogeneity* is the issue

Spatial Lag Model

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

- Dependent variable is actively influenced by its neighbors
- Easily modified to incorporate a spatially lagged independent variable as well
  - "Spatial Durbin"
- Can easily interpret ρ

Spatial Error Model

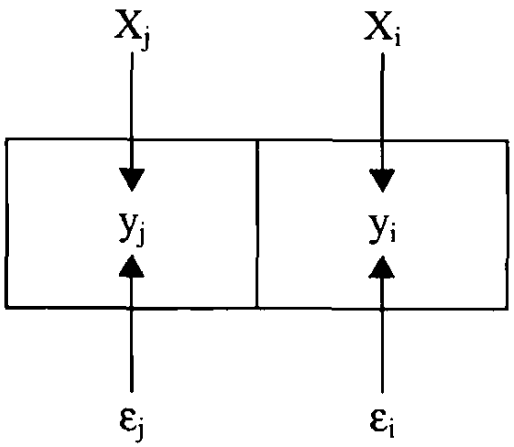
$$y = X\beta + u$$

$$u = \lambda W u + \epsilon$$

- Less compelling with respect to what this tells us about spatial processes
- Spatial lag is in the error term
  - Addresses missing variables with spatial effects
- Employed to counter heterogeneity in units of observation
- λ is not easily interpreted

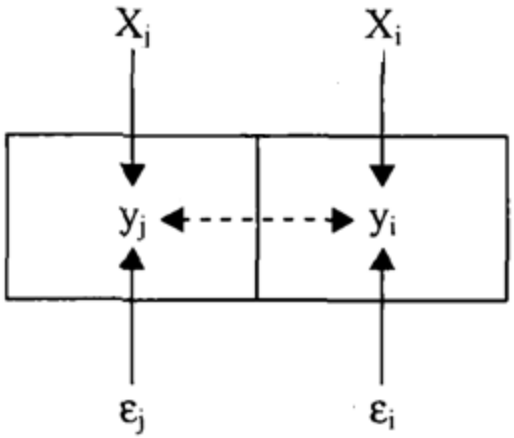
Conceptual Comparison

OLS



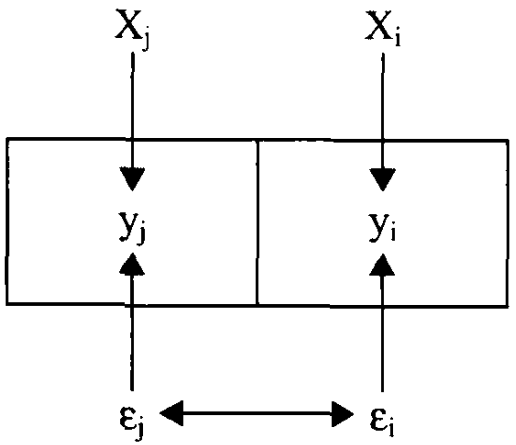
No influence from neighbors

SPATIAL LAG



Dependent variable influenced by neighbors

SPATIAL ERROR



Residuals influenced by neighbors

Which Model to Choose?

Spatial Lag

- You believe space matters in the relationships hypothesized
- You want to allow the model to build interaction into its workings

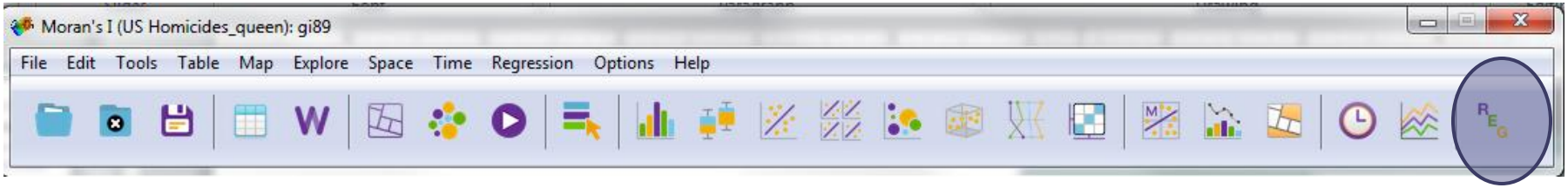
We will formally test which model is more appropriate by using LM tests (in GeoDa)

Spatial Error

- No theoretical reason to believe Y is affected by neighboring Y or X
- Missing variables difficult to quantify but have likely spatial footprint
- Size / density of observations vary widely and (potentially) systematically

Spatial Regression in GeoDa

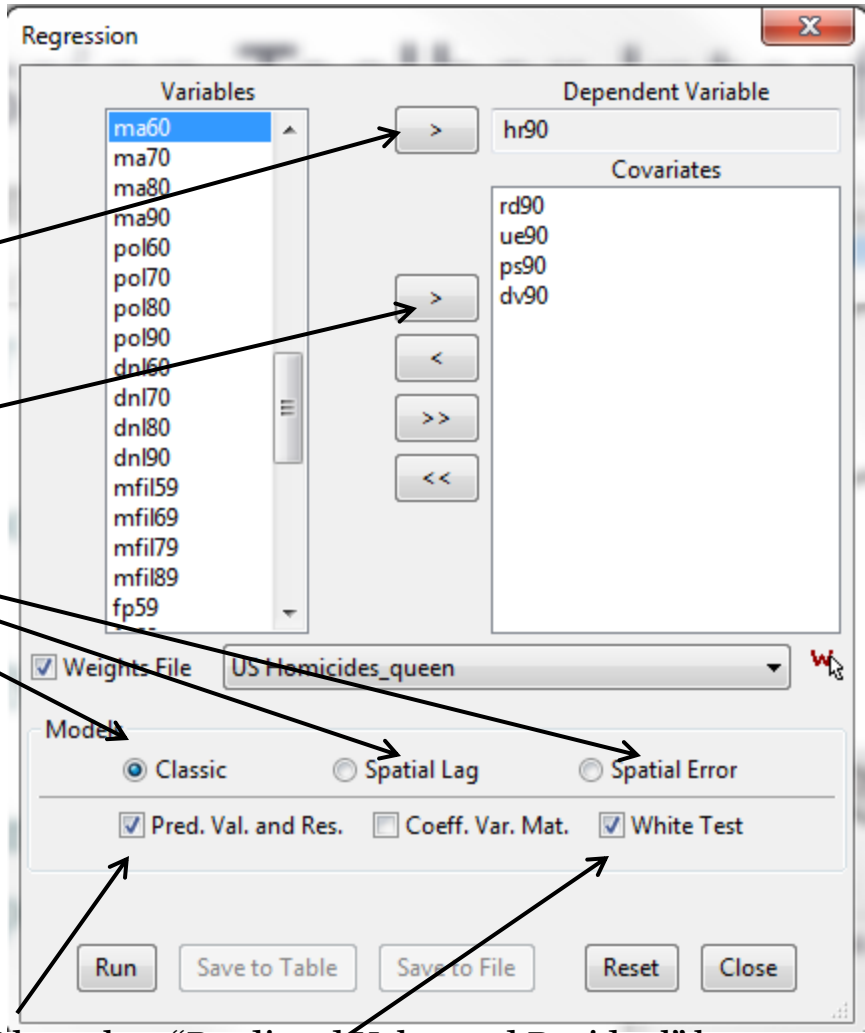
- OLS with diagnostics for spatial effects
  - "Decision Tree" for which spatial model to choose
- ML regression of spatial lag / spatial error
  - Saving / using residuals & predicted values



Regression Toolbar Interface

- Regression model specification window
  - Select DV first (using > button)
  - Then choose IVs (using > button)
  - 3 Model types listed here
  - IF you run a spatial model (lag, error), specify weight matrix

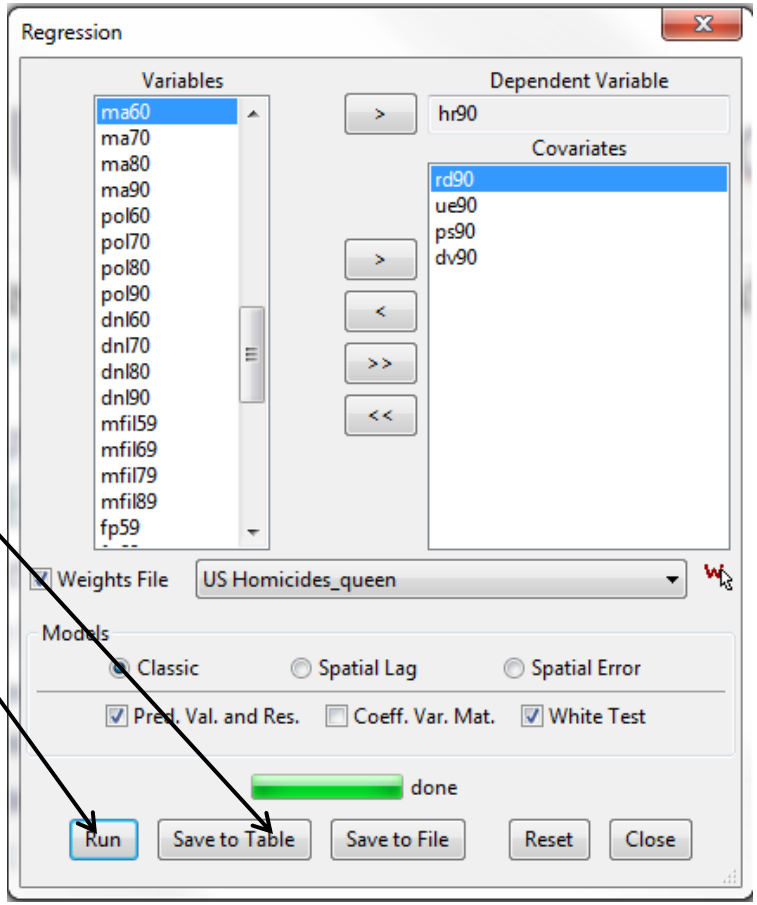
You will also need a weight matrix if you want to test for spatial dependence with OLS

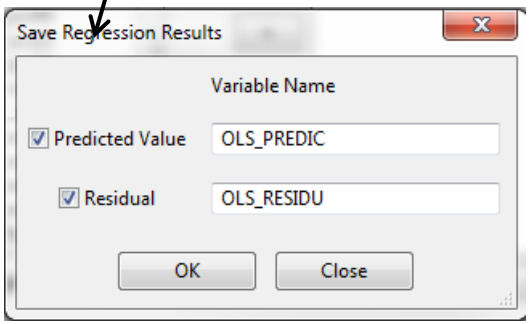


Also select "Predicted Value and Residual" box
And "White Test" box (for heteroskedasticity)

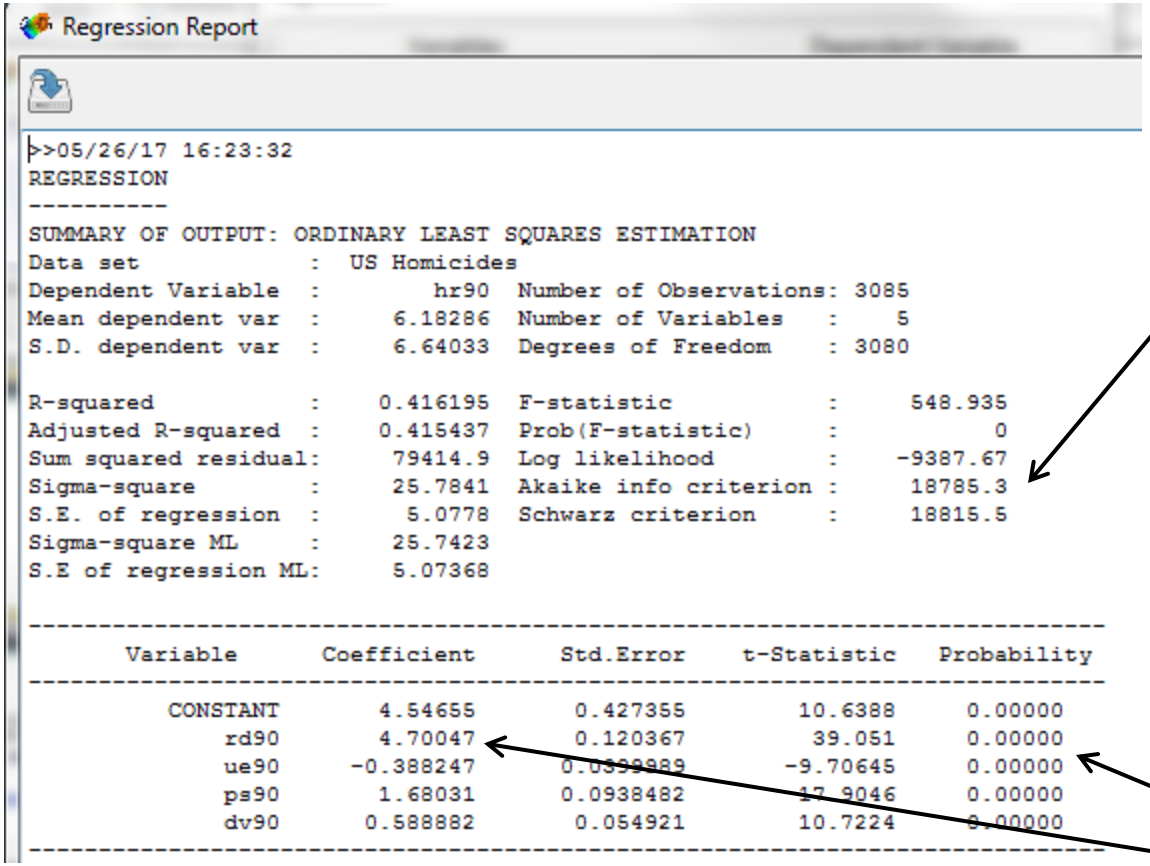
Spatial Regression in GeoDa

- Regression model specification window (cont'd)
  - Click "Run"
    - Progress bar will fill up
  - MAKE SURE TO:
    - Select "Save to Table"
    - Save the regression results
      - Predicted Value
      - Residuals
    - Name these based on the model
      - Ex: LAG_PREDIC or LAG_RESIDU





OLS Output



Regression Report

```
>>05/26/17 16:23:32
REGRESSION
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Data set            :  US Homicides
Dependent Variable  :        hr90  Number of Observations: 3085
Mean dependent var  :     6.18286  Number of Variables   :    5
S.D. dependent var  :     6.64033  Degrees of Freedom    : 3080

R-squared           :    0.416195  F-statistic           :     548.935
Adjusted R-squared  :    0.415437  Prob(F-statistic)     :           0
Sum squared residual:     79414.9  Log likelihood        :    -9387.67
Sigma-square        :     25.7841  Akaike info criterion :     18785.3
S.E. of regression  :      5.0778  Schwarz criterion     :     18815.5
Sigma-square ML     :     25.7423
S.E of regression ML:     5.07368

------------------------------------------------------------------------
       Variable      Coefficient      Std.Error    t-Statistic   Probability
------------------------------------------------------------------------
          CONSTANT       4.54655       0.427355        10.6388      0.00000
              rd90       4.70047       0.120367         39.051      0.00000
              ue90     -0.388247      0.0399989       -9.70645      0.00000
              ps90       1.68031      0.0938482        17.9046      0.00000
              dv90      0.588882       0.054921        10.7224      0.00000
------------------------------------------------------------------------
```

Measures of fit

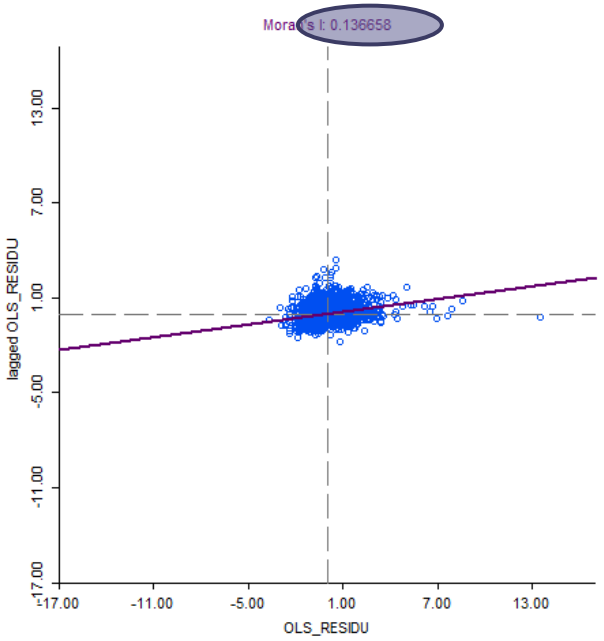
Spatial Diagnostics Included After Basic Results

Parameter values & significance

Moran's I of Residuals

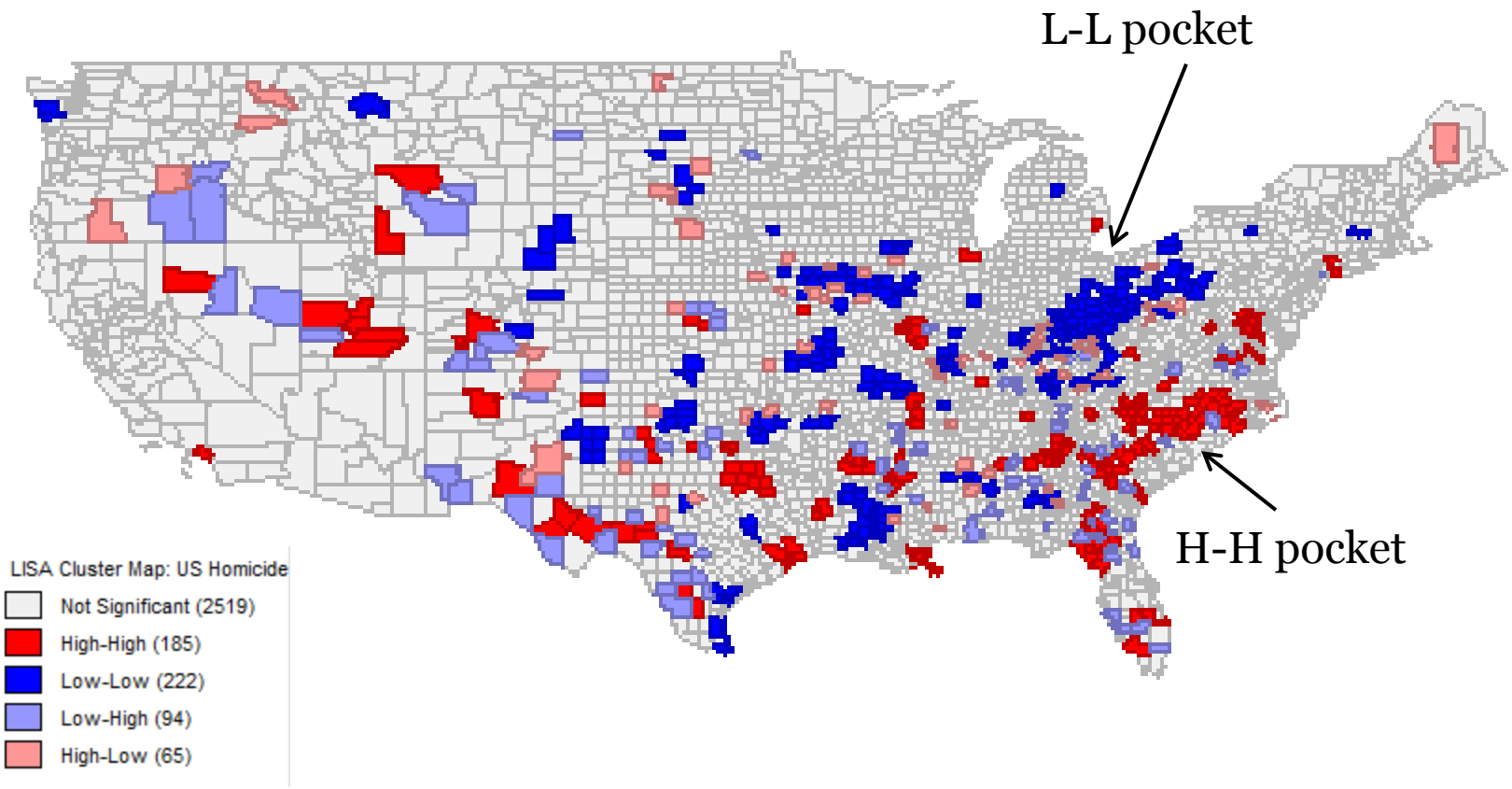
- The Moran's I reported in the results will be EXACTLY THE SAME if you run a Moran's I command for the OLS_RESIDU

```
DIAGNOSTICS FOR SPATIAL DEPENDENCE
FOR WEIGHT MATRIX : US Homicides_queen
   (row-standardized weights)
TEST                            MI/DF        VALUE           PROB
Moran's I (error)               0.1367       12.8235         0.00000
Lagrange Multiplier (lag)            1      178.3324         0.00000
Robust LM (lag)                      1       28.0817         0.00000
Lagrange Multiplier (error)          1      160.9723         0.00000
Robust LM (error)                    1       10.7216         0.00106
Lagrange Multiplier (SARMA)          2      189.0540         0.00000
```



If the Moran's I of the errors is significant – OLS is <u>not</u> the appropriate model!

Can also look at LISA map of OLS residuals



Decision Tree

Observe significance levels for LM Error (λ) and LM Lag (ρ)

Only look at Robust LM results if <u>BOTH</u> LM Error and LM Lag are significant

Figure 23.24: Spatial regression decision process.

Spatial Lag & Error Model Results

Spatial Lag

Regression Report

```
>>05/26/17 15:42:42
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIAL LAG MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : US Homicides
Spatial Weight      : US Homicides_queen
Dependent Variable  :        hr90  Number of Observations: 3085
Mean dependent var  :     6.18286  Number of Variables   :    6
S.D. dependent var  :     6.64033  Degrees of Freedom    : 3079
Lag coeff.   (Rho)  :    0.263898

R-squared           :    0.451098  Log likelihood        :   -9312.37
Sq. Correlation     : -           Akaike info criterion :    18636.7
Sigma-square        :     24.2033  Schwarz criterion     :    18672.9
S.E of regression   :     4.91968

------------------------------------------------------------------------------
      Variable       Coefficient     Std.Error       z-value     Probability
------------------------------------------------------------------------------
          W_hr90      0.263898      0.0217783       12.1175      0.00000
        CONSTANT      2.72049       0.431226         6.30875     0.00000
            rd90      3.79979       0.13599         27.9416      0.00000
            ue90     -0.294169      0.0389342       -7.55554     0.00000
            ps90      1.3956        0.0929926       15.0077      0.00000
            dv90      0.527921      0.053493         9.86897     0.00000
------------------------------------------------------------------------------
```

Spatial Error

Regression Report

```
>>05/26/17 15:43:38
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIAL ERROR MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : US Homicides
Spatial Weight      : US Homicides_queen
Dependent Variable  :        hr90  Number of Observations: 3085
Mean dependent var  :    6.182860  Number of Variables   :    5
S.D. dependent var  :    6.640331  Degrees of Freedom    : 3080
Lag coeff. (Lambda) :    0.331370

R-squared           :    0.454691  R-squared (BUSE)      : -
Sq. Correlation     : -           Log likelihood        :-9314.411113
Sigma-square        :     24.0448  Akaike info criterion :    18638.8
S.E of regression   :     4.90355  Schwarz criterion     :      18669

------------------------------------------------------------------------------
      Variable       Coefficient     Std.Error       z-value     Probability
------------------------------------------------------------------------------
        CONSTANT      3.75341       0.505358         7.42722     0.00000
            rd90      4.32221       0.143587        30.1017      0.00000
            ue90     -0.244072      0.0449439       -5.43059     0.00000
            ps90      1.42044       0.110808        12.8189      0.00000
            dv90      0.565266      0.0623751        9.06237     0.00000
          LAMBDA      0.33137       0.0255112       12.9892      0.00000
------------------------------------------------------------------------------
```

ρ

Compare parameter values

λ

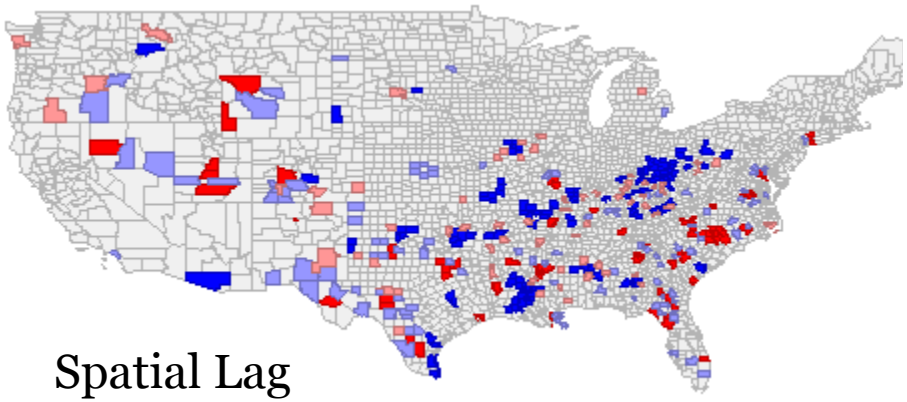
Measures of Fit

- Psuedo R^2 of Spatial models NOT directly comparable to R^2 of OLS
- 3 Measures of Fit to Compare models
  - Log-likelihood
    - Higher values (i.e. less negative) better
  - Akaiki Information Criteria (AIC)
    - Lower measure = better fit
  - Schwarz Criterion (SC)
    - Lower measure = better fit

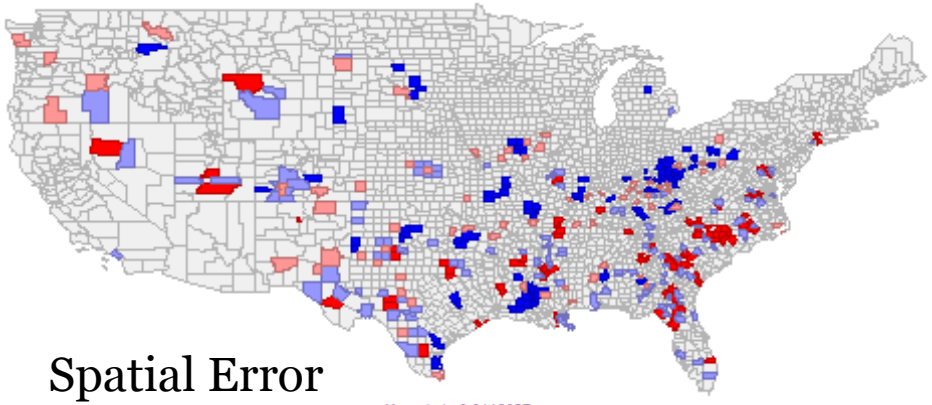
Model Comparison

	OLS		Lag		Error	
Const	4.546	***	2.720	***	3.753	***
rd90	4.700	***	3.799	***	4.322	***
ue90	-0.388	***	-0.294	***	-0.244	***
ps90	1.680	***	1.395	***	1.420	***
dv90	0.588	***	0.528	***	0.565	***
Rho			0.264	***		
Lambda					0.331	***
Log Likelihood	-9387.67		-9312.37		-9314.4	
AIC	18785.3		18636.7		18638.8	
Schwarz	18815.5		18672.9		18669	

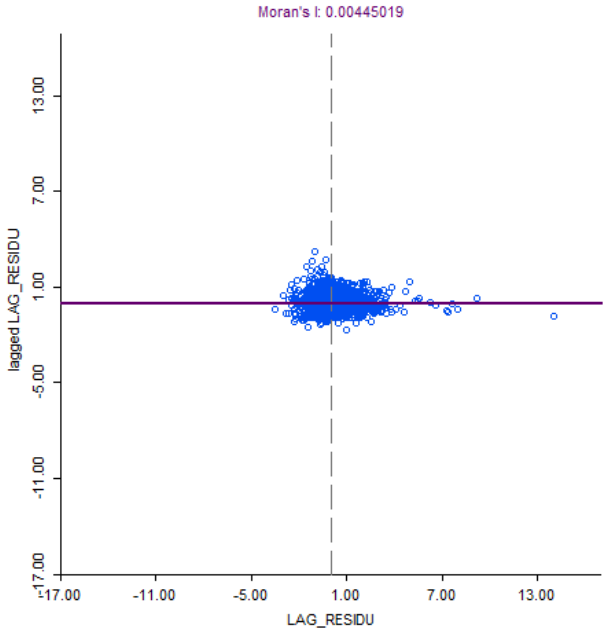
Look at Residuals!

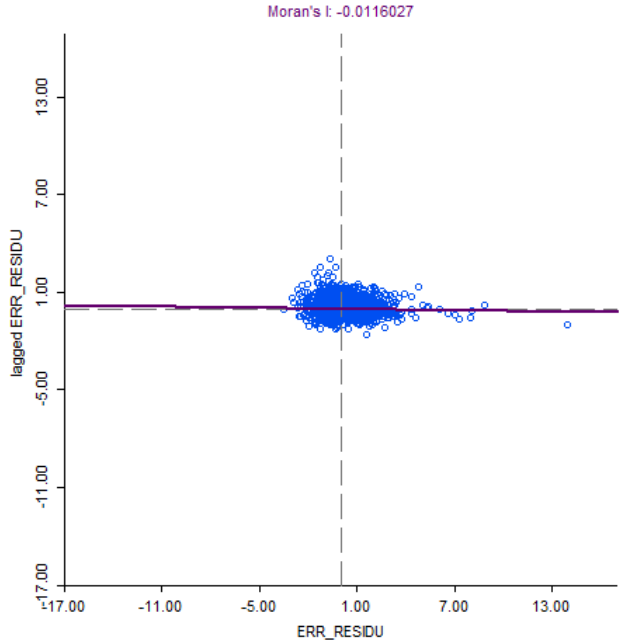


Spatial Lag



Spatial Error





Assignment

- Complete the worksheet using Las Rosas precision agriculture data
  - Run OLS, spatial error, spatial lag (and compare)

5. Future Spatial Econometrics Work

- Spatial Panel Data
- Geographically Weighted Regression (GWR)
- Bayesian Spatial Models

Spatial Panel Data

- Currently, most spatial analysis is done on cross-sectional data
- This is changing…

$$\boldsymbol{Y}_t = \delta \boldsymbol{W}\boldsymbol{Y}_t + \alpha \boldsymbol{\iota}_N + \boldsymbol{X}_t\boldsymbol{\beta} + \boldsymbol{W}\boldsymbol{X}_t\boldsymbol{\theta} + \boldsymbol{u}_t$$

$$\boldsymbol{u}_t = \lambda \boldsymbol{W}\boldsymbol{u}_t + \boldsymbol{\varepsilon}_t$$

Simple extension of cross-sectional model over T time periods

$$\boldsymbol{Y}_t = \rho \boldsymbol{W}\boldsymbol{Y}_t + \alpha \boldsymbol{\iota}_N + \boldsymbol{X}_t\boldsymbol{\beta} + \boldsymbol{W}\boldsymbol{X}_t\boldsymbol{\theta} + \boldsymbol{\mu} + \xi_t \boldsymbol{\iota}_N + \boldsymbol{u}_t$$

$$\boldsymbol{u}_t = \lambda \boldsymbol{W}\boldsymbol{u}_t + \boldsymbol{\varepsilon}_t$$

Extension to allow for spatial and temporal heterogeneity

Elhorst (2014)

Spatial Panel Data

- Pros:
  - More data (N x T observations)
  - Easier to make case for causality when time dimension is included
  - Allows for random / fixed effects
- Cons
  - LOTS more correlation possible: Observation i_t may be correlated with i_{t-1} , j_t , or even j_{t-1}
  - What if W varies over time??

Geographically Weighted Regression

- Allows relationships in regression model to vary over space
  - What we have just done uses *constant* regression coefficients over space (β, λ, ρ)
  - GWR estimates regression coefficients for *each unit of analysis*

- Based on idea of estimating local models using subsets of observations around a point
  - Nonparametric

Wheeler (2014)

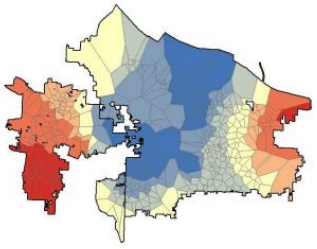
Geographically Weighted Regression

$$y_i = \beta_0 + \sum_{k=1}^{p-1} \beta_k x_{ki} + \varepsilon_i$$

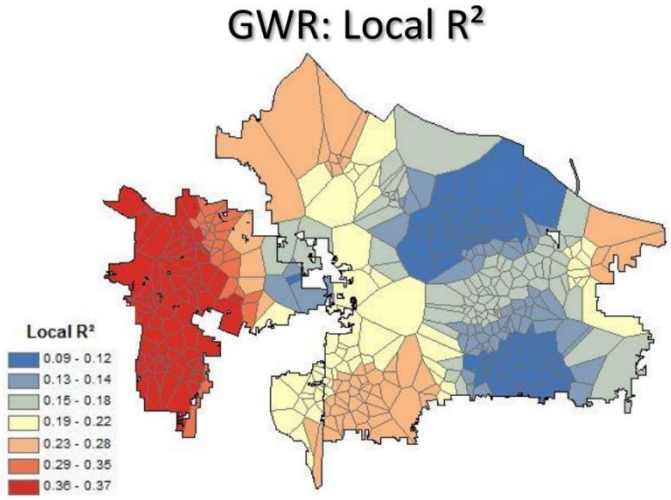
Traditional OLS – each variable x_k gets its own parameter β_k

$$y_i = \beta_{0i} + \sum_{k=1}^{p-1} \beta_{ki} x_{ki} + \varepsilon_i$$

GWR– each unit of analysis x_i gets its own parameter β_i



Global: 0.000025



Result: Coefficient estimates (and resulting R^2) that vary by unit of analysis!

GWR: Problems

- Significant published evidence that multicollinearity in estimated coefficients may bias results
- SW packages do not typically calculate t-stats

"notoriously unreliable"

http://r-sig-geo.2731867.n2.nabble.com/A-question-about-gwr-morantest-pvalue-td7292670.html

There is no "correct" here, as ?pchisq shows that if lower.tail= is not set it is taken by default to be TRUE. You will need to check this yourself. Note that GWR is a notoriously unreliable technique, and simulation studies indicate that it finds pattern in coefficients even when there is none. So any tests are doubtful anyway - it should only be used for exploring the data for possible missing variables or inappropriate functional forms.

Hope this clarifies,

Roger

--

Roger Bivand
Department of Economics, NHH Norwegian School of Economics,
Helleveien 30, N-5045 Bergen, Norway.
voice: +47 55 95 93 55; fax +47 55 95 95 43
e-mail: Roger.Bivand at nhh.no

Wrote GWR package (spgwr) in R

R package spgwr:
No standard errors, no t-stats

ArcGIS:
Includes standard errors but no t-stats

Bayesian Spatial Analysis

- Applying Bayesian methodology to spatial models
- Bayes' theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)},$$

Replace B with D to reflect spatial data / weight matrix

Replace A with Θ to represent spatial parameters

$$P(\Theta|D) = \frac{P(D|\Theta)P(\Theta)}{P(D)}.$$

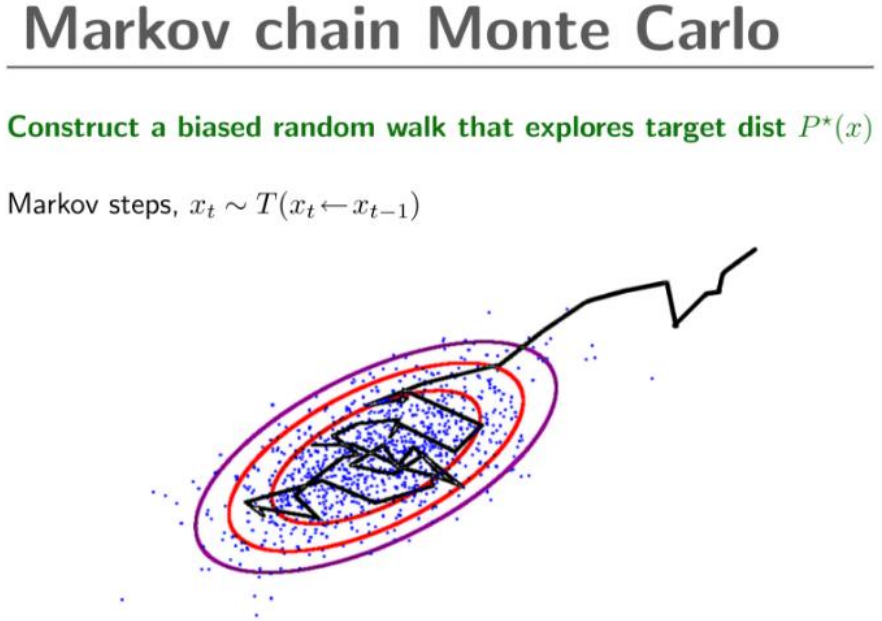
- So, P(D|Θ) is the likelihood of obtaining D under the spatial model that contains Θ

Bayesian Spatial Analysis

- Offers a more solid foundation relating to existing knowledge of unknown parameters
  - Bayesian approach assumes unknown parameters follow prior distributions, and uses these priors to update later distributions of the parameters
  - Alternative view of "frequentists": unknown parameters are 'fixed and knowable' because observed data is from a specific likelihood model
- Challenge: Obtaining posterior distributions requires integration – typically approximated numerically (and not easily computed)

Bayesian Spatial Analysis

- Typically requires use of Markov Chain Monte Carlo (MCMC) simulations
- Software programs
  - WinBUGS
  - GeoBUGS
  - geoR (R)
  - spBayes (R)



Recommended Readings

- LeSage, J. 2014. "What Regional Scientists Need to Know about Spatial Econometrics." *The Review of Regional Studies* 44(1), 13-32. Link

- Vega, S. and J. Elhorst. 2013. "On Spatial Econometric Models, Spillover Effects, and W." Working paper for ERSA Conference. Link

THANKS FOR ATTENDING!