Introduction to Spatial Regression Analysis

Paul Voss
UNC Chapel Hill

Day 1

Objective of course

Provide an introduction and overview of concepts and techniques of spatial regression analysis

with hands-on experience in the afternoon lab sessions

What’s the point?

Data that are referenced to location bring important additional information to your data analysis

But they also present some (possibly unfamiliar) pitfalls that require a new awareness as your analysis proceeds

Plan for next 3 days (1)

- Today: Broad overview of spatial data, spatial data analysis and core spatial concepts; OLS overview
  - Why "spatial is special"
  - Classical linear regression model
    - assumptions underlying OLS
    - consequences of violations of assumptions
  - Why spatial processes violate OLS assumptions
  - Introduction to EDA/ESDA
  - Lab: Introduction to *GeoDa* & R
- Tomorrow: ESDA & introduction spatial autocorrelation
  - Understanding & measuring global spatial autocorrelation
  - Weights matrices
  - Understanding & measuring local spatial autocorrelation
    - Moran scatterplot
    - LISA statistics
  - Lab: Global / local spatial autocorrelation with *GeoDa* & R

Plan for the week (2)

- Thursday: Spatial modeling
  - Firm understanding of spatial processes
    - spatial heterogeneity
    - spatial dependence
  - Spatial regression models
    - OLS in *GeoDa*
    - understanding *GeoDa* regression diagnostics
    - spatial lag model
    - spatial error model
  - Lab: Spatial regression modeling with *GeoDa* & R

Welcome & Introductions

Plan for today

- A brief motivational example
- Why “spatial is special”
  - characteristics of spatial data
  - potential complications when using spatial data
- Spatial analysis vs. spatial *data* analysis
- Broad overview of spatial data and spatial data analysis
- Classes of problems in spatial data analysis
- Review OLS assumptions & violations
- Exploratory Data Analysis
- Afternoon lab

Any questions as we get started?

One thing worth mentioning…

A dynamic field of

© 2009

Some recent general books

© 2010

A dynamic field of study!

© 2006

Some recent specialized books

© 2010

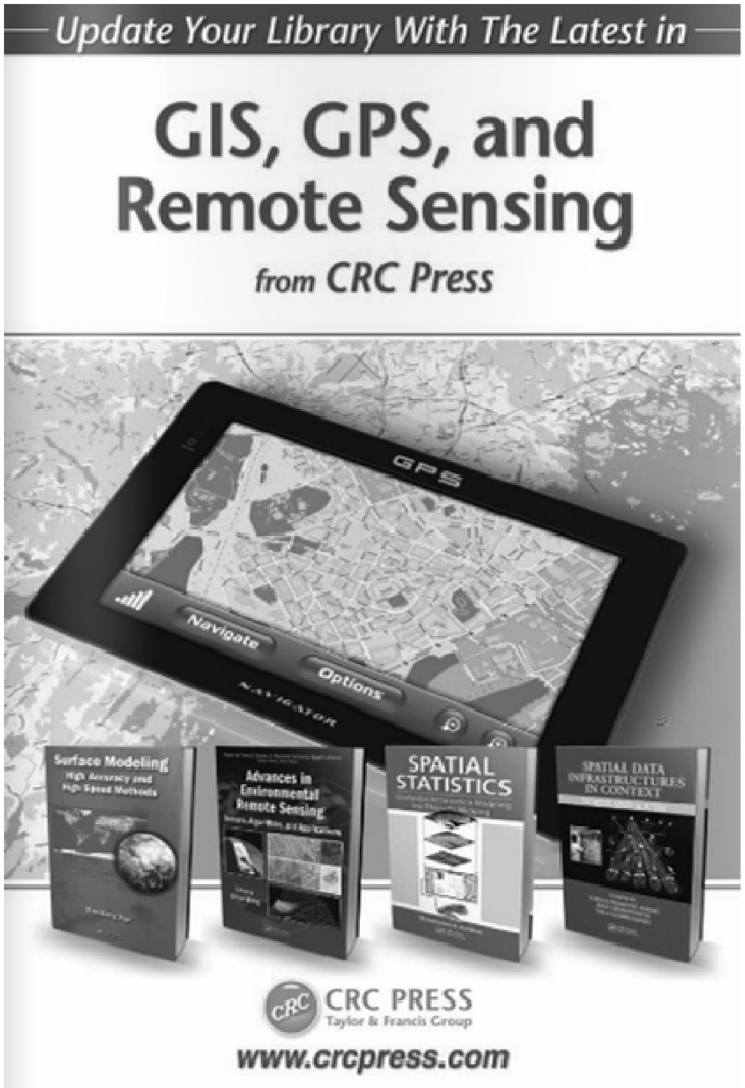
A dynamic field of study!

© 2008

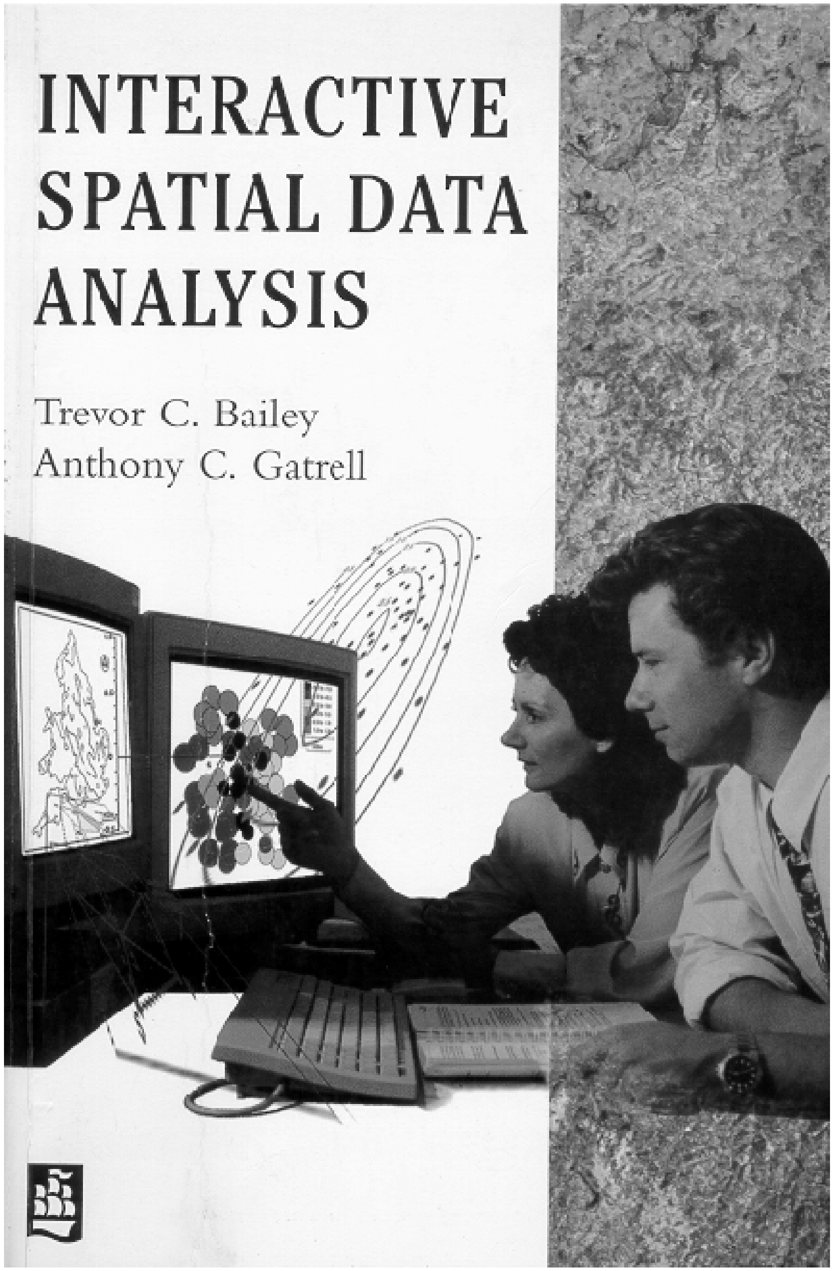
More…

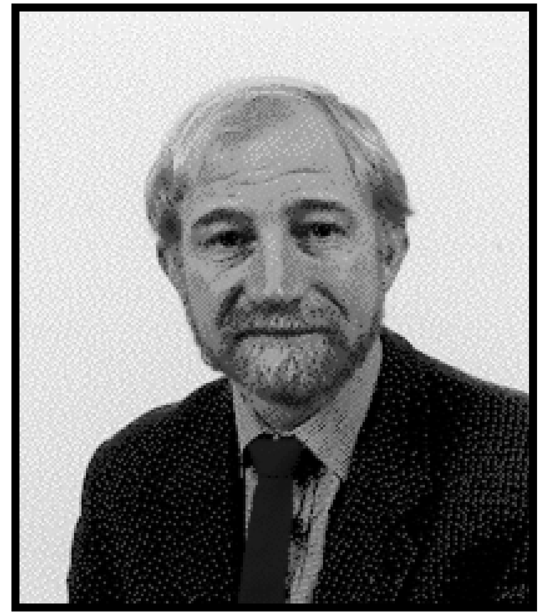
© 2011

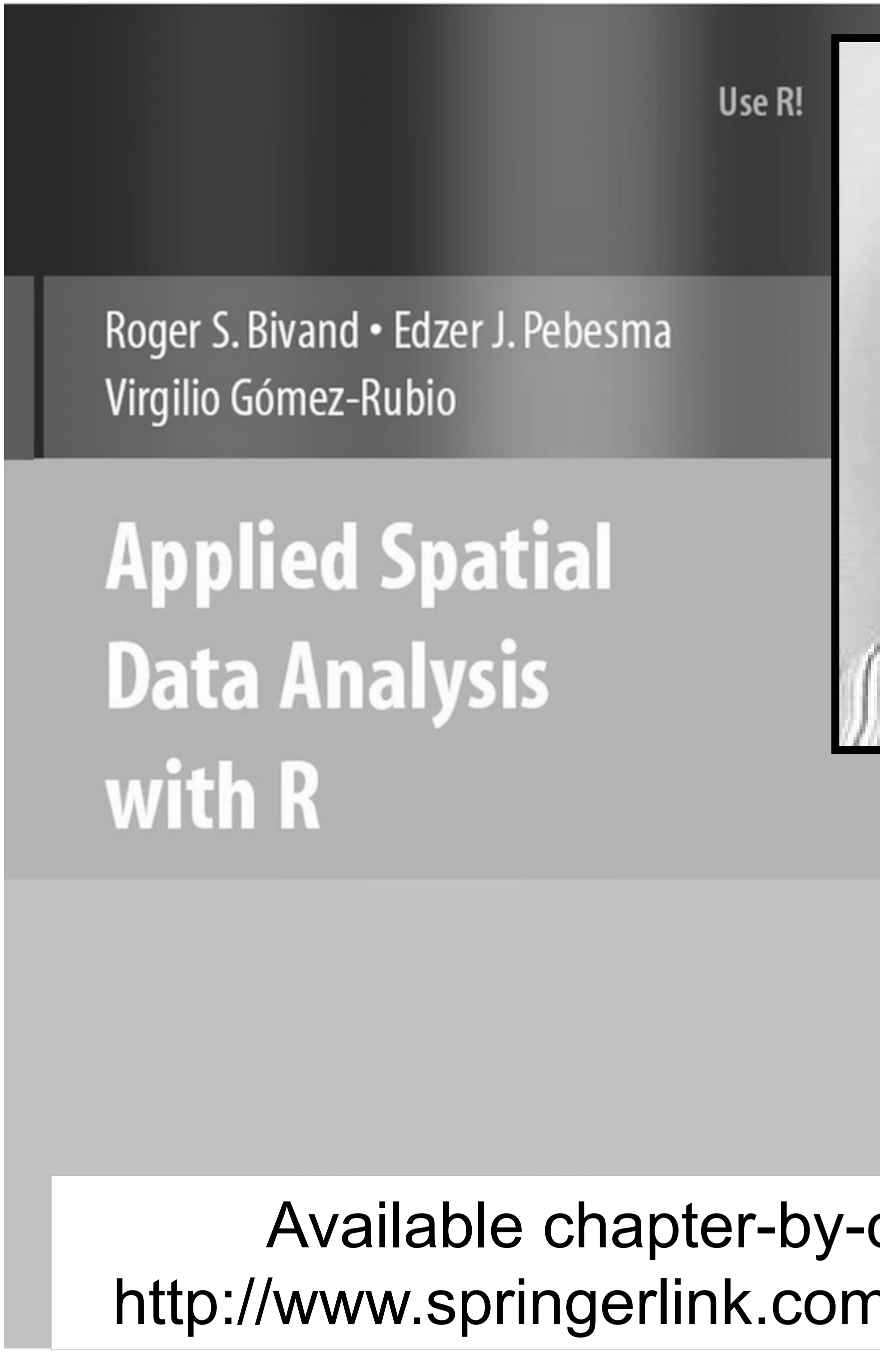
A dynamic field of study!



Recent catalogue from CRC Press. Several dozen new, or fairly new, books focusing largely on spatial applications of remotely sensed data


INTERACTIVE
SPATIAL DATA
ANALYSIS
Trevor C. Bailey
Anthony C. Gatrell





Available chapter-by-chapter as PDF files at:
http://www.springerlink.com/content/978-0-387-78170-9

GeoDa


Luc Anselin

GEODA CENTER
FOR GEOSPATIAL ANALYSIS
AND COMPUTATION

ASU

ARIZONA STATE UNIVERSITY

http://geodacenter.asu.edu

R: “The GNU S”

- S language developed at Bell Labs beginning in late 1960s (Chambers, Becker, Wilks & several others)
- Late 1980s, commercial version of S (S-Plus) is launched
- Early 1990s, Ross Ihaka & Robert Gentleman (University of Auckland) develop a reduced version of S for classroom
- 1995: Ihaka & … open source General Publ…
- Since 1997, c… small group of statisticians, … R version 1.0.0 released on … 2.13.1
- Free; extensi… …asis on graphics; obj… …ing curve

Data Analysts Captivated by R’s Power

Stuart Isett for The New York Times

R first appeared in 1996, when the statistics professors Robert Gentleman, left, and Ross Ihaka released the code as a free software package.

Okay… some beginning facts

- Regression is the workhorse of quantitative social science
- Much social science data is spatially referenced
- Spatially referenced data bring special problems to an analysis
  - heterogeneity of observational units → heteroskedasticity
  - spatial autocorrelation → residual dependence
- A consequence of these “special problems” is that the assumption of *iid* errors in a standard OLS regression specification is violated, and statistical inference from such a model is not valid

Motivation

- Omer R. Galle, Walter R. Gove, & J. Miller McPherson. 1972. "Population Density and Pathology: What Are the Relations for Man" *Science* 176(4030):23-30
  - data: 75 community areas in Chicago for 1960
  - 5 measures of "social pathology" as function of crowding, controlling for social class & ethnicity
  - *"…the greater the density, the greater the fertility"* (p. 176)
- Colin Loftin & Sally K. Ward. 1983. "A Spatial Autocorrelation Model of the Effects of Population Density on Fertility" *American Sociological Review* 48(1):121-128
  - *"…the GGM findings with regard to fertility are an artifact of the failure to recognize the presence of disturbance variables which are spatially autocorrelated"* (p. 127)
- Moral: When analyzing spatially referenced data, it's highly useful to know something about the rudiments of spatial data analysis (i.e., some understanding of why "spatial is special")

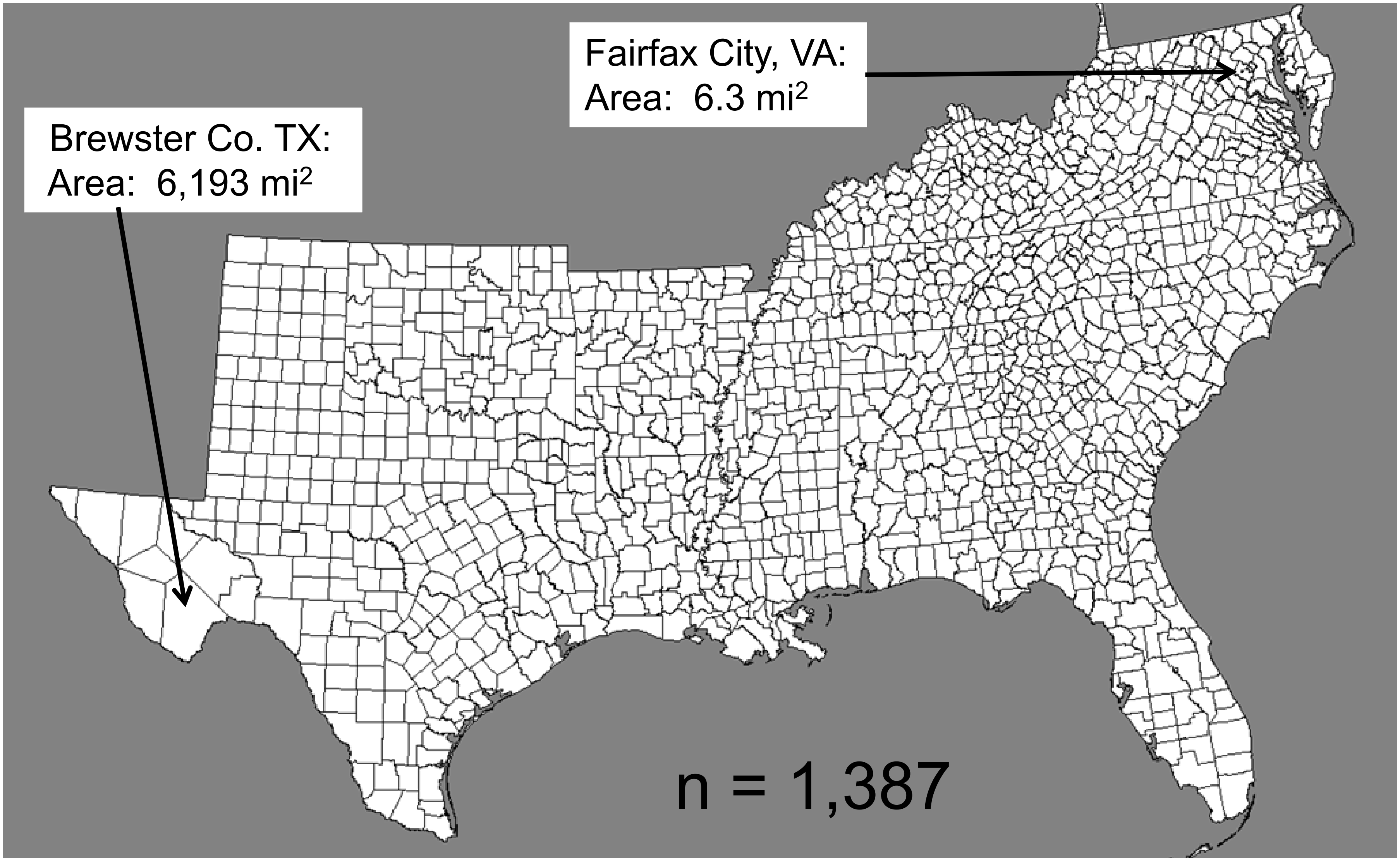
And why *is* spatial special?

- Scale dependency
  - Robinson (*ASR*, 1950) → "Ecological Fallacy"
  - *"The relationship between ecological and individual correlations which is discussed in this paper provides a definite answer as to whether ecological correlations can validly be used as substitutes for individual correlations. They cannot."* (p. 357)
  - MAUP (Modifiable Areal Unit Problem)
  - *"Habitual users of ecological correlations know that the size of the coefficient depends to a marked degree upon the number of sub-areas. …[T]he size of the ecological correlation [will increase numerically as consolidation of smaller areas into larger areas takes place]."* (Robinson, pp. 357-8)

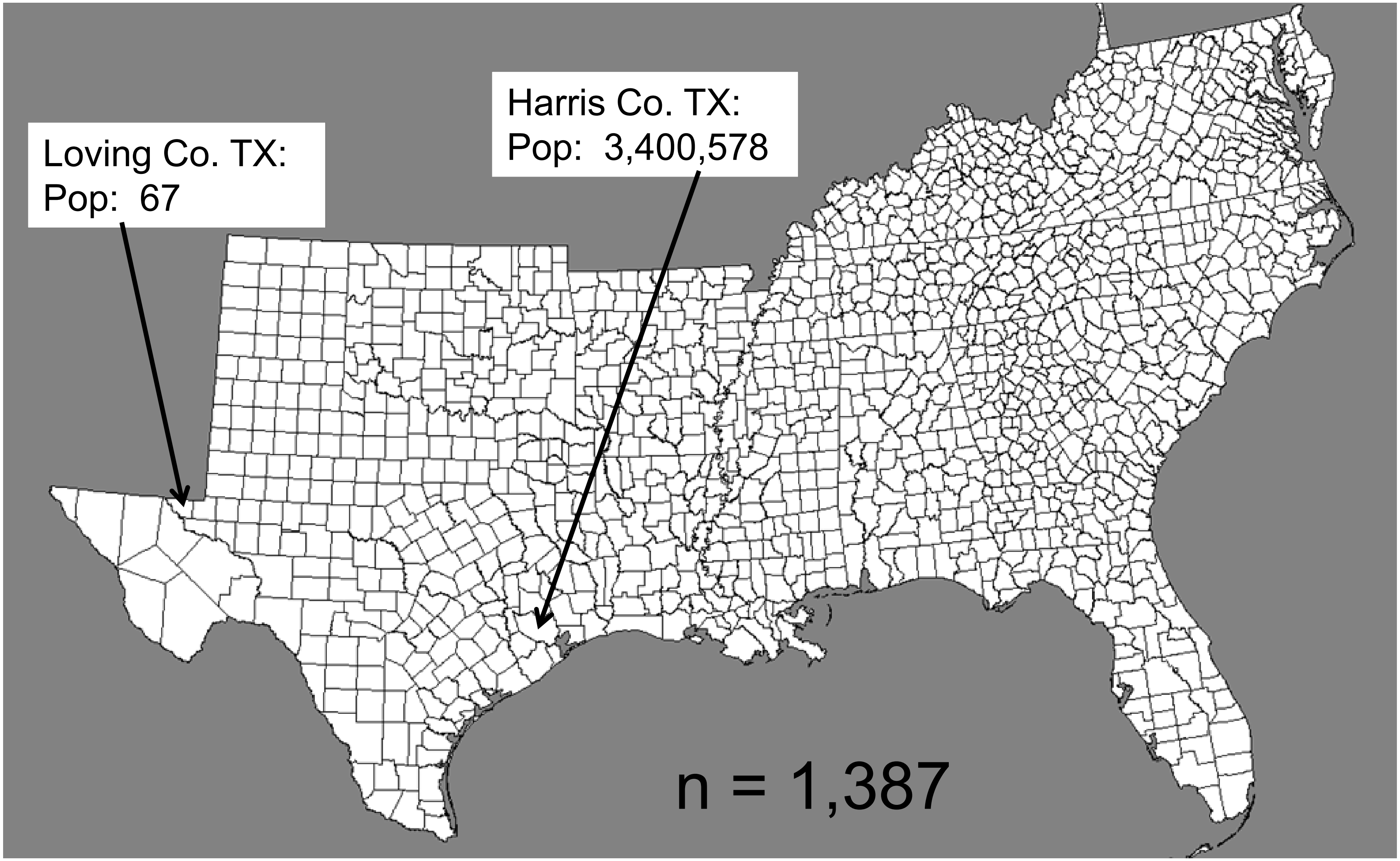
Why is spatial special? (2)

- Scale dependency
- Observational areas are generally of different size
  - heterogeneity → heteroskedasticity

Counties in U.S. South: 2000 Census



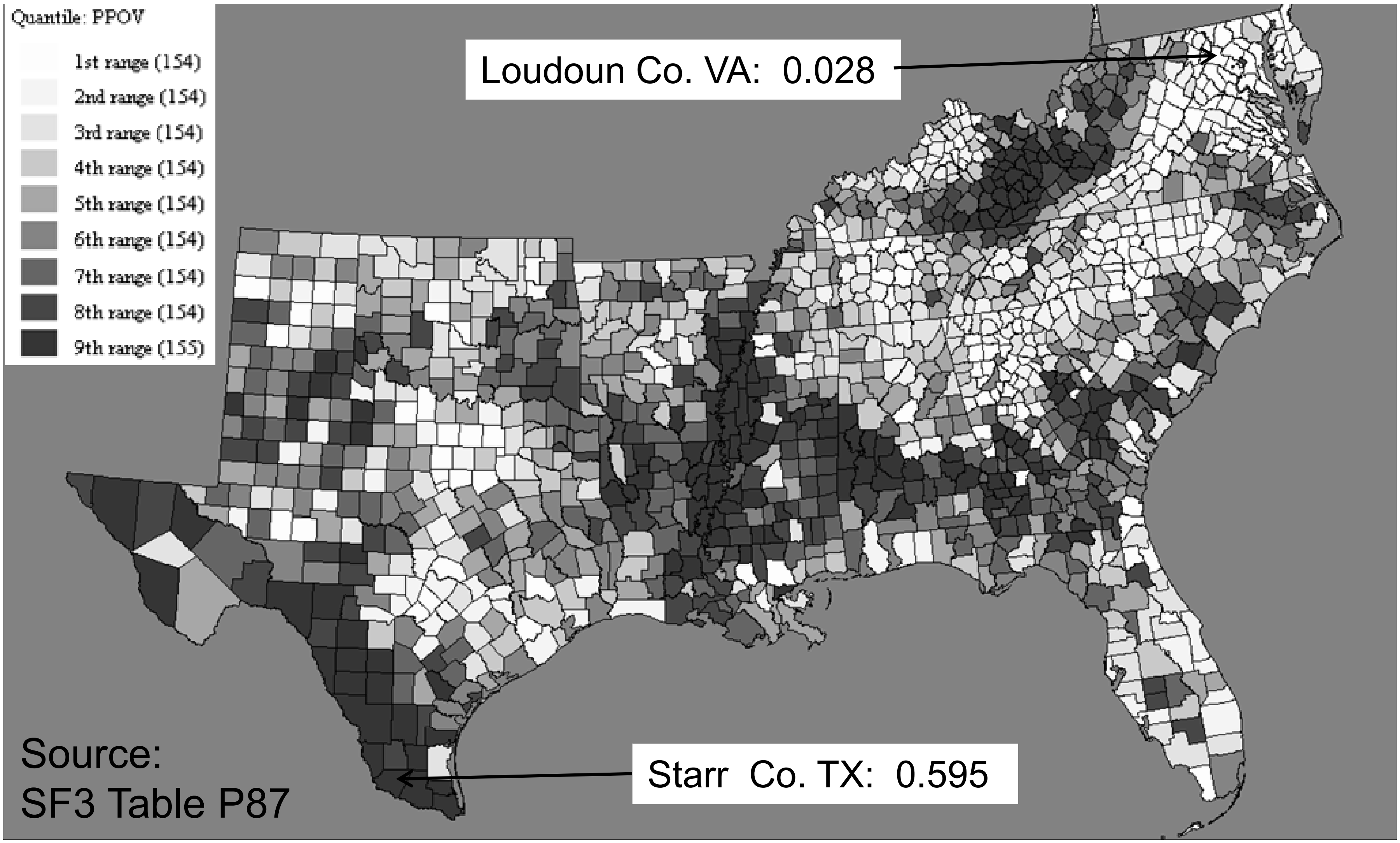
Counties in U.S. South: 2000 Census



Why is spatial special? (3)

- Scale dependency
- Observational areas are generally of different size (geographic size; population size)
  - heterogeneity → heteroskedasticity
- Neighboring areas are similar
  - Tobler's 1st law of Geography: *"Everything is related to everything else, but near things are more related than distant things."* (1970:236)
  - (positive) spatial autocorrelation

Proportion of Children (under age 18) in Poverty: 2000 Census

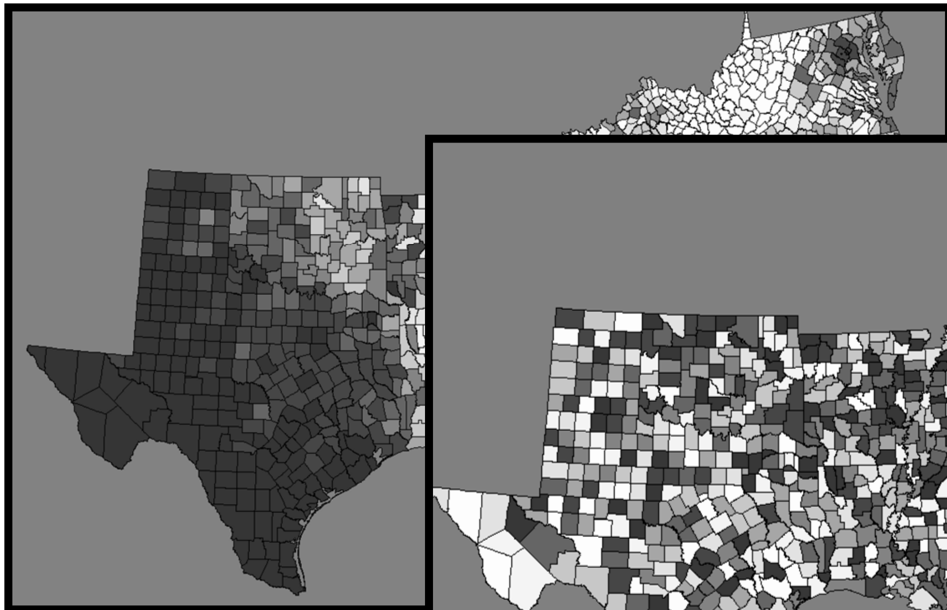


Why is spatial special? (4)

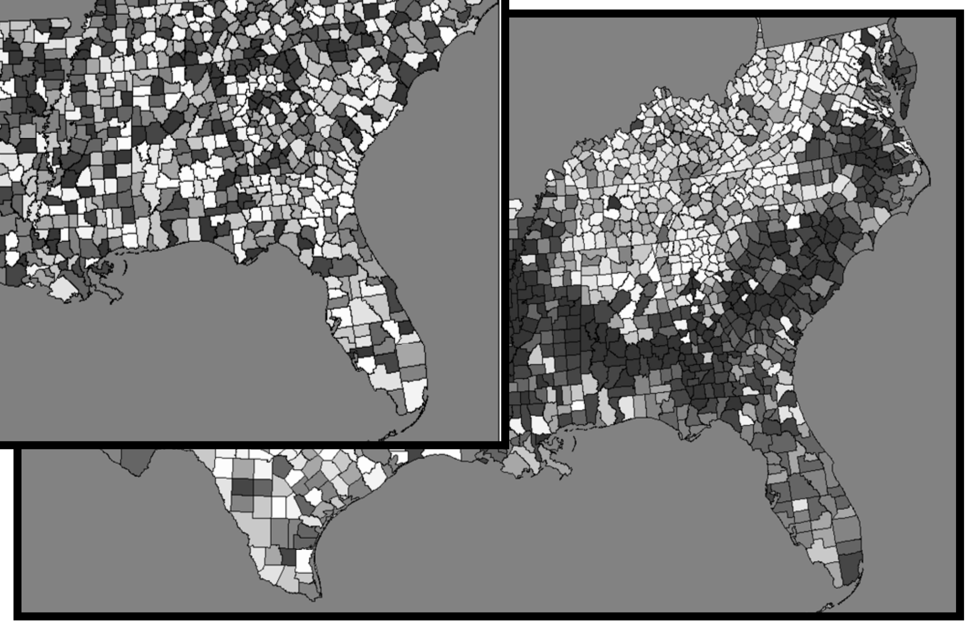
- Scale dependency
- Observational areas are generally of different size
  - heterogeneity → heteroskedasticity
- Neighboring areas are similar
  - Tobler's 1st law of Geography: *"Everything is related to everything else, but near things are more related than distant things."* (1970:236)
  - (positive) spatial autocorrelation
- Probable stumbling blocks when modeling the data
  - again… the assumption of *iid* errors in a standard OLS regression specification is violated and statistical inference is not valid

Spatial versus Non-Spatial Data Analysis

Take these maps, for example



or even this one…



Any traditional data analysis that does not utilize the location & spatial arrangement (topological information) of the data will lead to identical results for the three maps

When data have spatial structure, even simple statistical measures likely have problems

- Consider n independent samples $\{y_1, y_2, \ldots, y_n\}$ from a normal distribution with mean μ and known variance σ^2, say… n = 35 and sample mean = 21.2 and sample variance of 12
- The most efficient unbiased estimator of μ is the sample average which follows a normal distribution with mean μ and variance σ^2/n

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

- Suppose we want to test H_0: $\mu = 20$
- Can do simple 2-tailed z test: $p = 0.040$ (reject null at $\alpha = 0.05$)

$$z = \frac{21.2 - 20}{\sqrt{12} / \sqrt{35}} = 2.049$$

- Thus a 2-sided 95% confidence interval for μ is:

$$(\bar{y} - 1.96\sigma/\sqrt{n},\ \bar{y} + 1.96\sigma/\sqrt{n}) =$$
$$(21.2 - 1.15,\ 21.2 + 1.15) = (20.05,\ 22.35)$$

When data have spatial structure, even simple statistical measures likely have problems (2)

- Now, instead of independent data, suppose we have spatial data, and our data have a covariance structure
- Let's assume data is now gathered door-to-door along a street ("linear transect"; $\mathcal{R}^1$), and let's further suppose that $\text{corr}(y_i,y_j) = \rho^{|i-j|} = (0.4)^{|i-j|}$ (correlation a function of separation of housing units)
  - $\text{cor}(y_1,y_2) = (0.4)$
  - $\text{cor}(y_1,y_3) = (0.4)^2 = 0.16$
  - $\text{cor}(y_1,y_4) = (0.4)^3 = 0.64$
  - $\text{cor}(y_4,y_7) = (0.4)^3 = 0.64$

- The sample mean will still follow a normal distribution with mean μ, but now with variance that adjusts for ρ:

$$\text{var}(\bar{y}) = \frac{\sigma^2}{n}\left[1 + 2\left(\frac{\rho}{1-\rho}\right)\left(1-\frac{1}{n}\right) - 2\left(\frac{\rho}{1-\rho}\right)^2\left(\frac{1-\rho^{n-1}}{n}\right)\right]$$

Cressie (1993:14)

When data have spatial structure, even simple statistical measures likely have problems (3)

- So… for these data, we have observations $\{y_1, y_2, \ldots, y_n\}$ from a normal distribution with mean μ and known variance σ^2 (n = 35, sample mean = 21.2, sample variance of 12)… but with a correlation structure defined by $\text{cor}(y_i, y_j) = \rho^{|i-j|} = (0.4)^{|i-j|}$
- This is equivalent to $\text{cov}(y_i, y_j) = \sigma^2 \rho^{|i-j|} = \sigma^2 (0.4)^{|i-j|}$
- Using the formula on the previous slide, we have:

$$\text{var}(\bar{y}) = \frac{\sigma^2}{n}\left[1 + 2\left(\frac{\rho}{1-\rho}\right)\left(1 - \frac{1}{n}\right) - 2\left(\frac{\rho}{1-\rho}\right)^2\left(\frac{1-\rho^{n-1}}{n}\right)\right]$$

$$= \frac{12}{35}\left[1 + 2\left(\frac{0.4}{1-0.4}\right)\left(1 - \frac{1}{35}\right) - 2\left(\frac{0.4}{1-0.4}\right)^2\left(\frac{1-0.4^{34}}{35}\right)\right]$$

$$= \frac{12}{35}\left[2.270\right] = 0.778$$

c.f. 0.343 for independent sample

$$z = \frac{21.2 - 20}{\sqrt{12/35}\sqrt{2.270}}$$

$$= 1.202$$

$$p = 0.229$$

Can't reject null

c.f. p = 0.040 for independent sample

When data have spatial structure, even simple statistical measures likely have problems (4)

- To summarize…
- For observations $\{y_1, y_2, \ldots, y_n\}$ from a normal distribution with mean μ and known variance σ^2 (n = 35, sample mean = 21.2, sample variance of 12)…
- We had 95% confidence interval of:
  - (20.05 , 22.35) if independence is assumed
    reject H_0 : μ = 20 at α =0.05 (2-tailed test)
  - (19.47 , 22.93) when the correlation structure is taken into account
    can't reject H_0
- Our illustration was for a 1-dimensional transect; life gets even more interesting in 2-dimensions!

When data have spatial structure, even simple statistical measures likely have problems (5)

- One final comment on this. When our data have spatial structure, models that assume independent observations will give us estimates from which statistical inference is invalid
- Failure to account for the underlying (positive) autocorrelation structure spuriously narrows our traditional confidence intervals. The reverse is true under (more rare) instances of negative spatial autocorrelation
- Another way of looking at this is to consider how much information coming to model is "lost" when our observations are correlated; points to the topic of "effective sample size"

$$n^* = n\left[1+2\left(\frac{\rho}{1-\rho}\right)\left(1-\frac{1}{n}\right)-2\left(\frac{\rho}{1-\rho}\right)^2\left(\frac{1-\rho^{n-1}}{n}\right)\right]^{-1}$$

$$n^* = 35/2.270 = 15.419$$

So…

“What makes the methods of modern [spatial data analysis] different from many of their predecessors is that they have been developed with the recognition that spatial data have unique properties and that these properties make the use of methods borrowed from aspatial disciplines highly questionable”

Fotheringham, Brunsdon & Charlton
Quantitative Geography:
Perspectives on Spatial Data Analysis
Sage, 2000 p. xii

Because of these unique p
we blithely carry out an OLS
using aggregated geograp

Some large subset of the f
undesirable horrors almost cer
us (the curse of Tobler's 1

- Our estimated regression coefficients are biased and inconsistent, or…
- Our estimated regression coefficients are inefficient
- Our R^2 statistic is exaggerated
- We've made incorrect inferences
- We'll *never* get it published
  - or shouldn't!

Given these problems, why would anyone bother to analyze spatial data?

- There's lots of it
- Occasional need for non-disclosure of individual-level data
- Space is important
  - space as a means of organizing human activities
  - location as a means of integrating interesting data
- There are strong arguments against using ecological regression
- Yet…some interesting questions can (only) be examined with spatial data

Now, Let's Define Some Terms

What exactly are “spatial (aggregated, geographic, ecological) data”?

…data where, in addition to attribute values relating to the primary phenomena of interest, the relative spatial locations of observations are also recorded

Such as…

Examples of spatial data

- Housing prices for census block groups
- Median household income for census tracts
- Poverty rates for counties
- Accident counts by intersection
- Cancer incidence reports for health districts
- County-to-county migration streams for persons 65+
- etc.

And what is spatial (ecological) regression analysis?

Regression using spatial data where *explicit attention* is given to location and arrangement of geographic units

Even if we don't really care about spatial processes

Spatial Analysis versus Spatial *Data* Analysis

GIS

Spatial
Analysis

- P-median problems
- Maximal covering problem
- Location set covering problem
- Traveling salesman problem

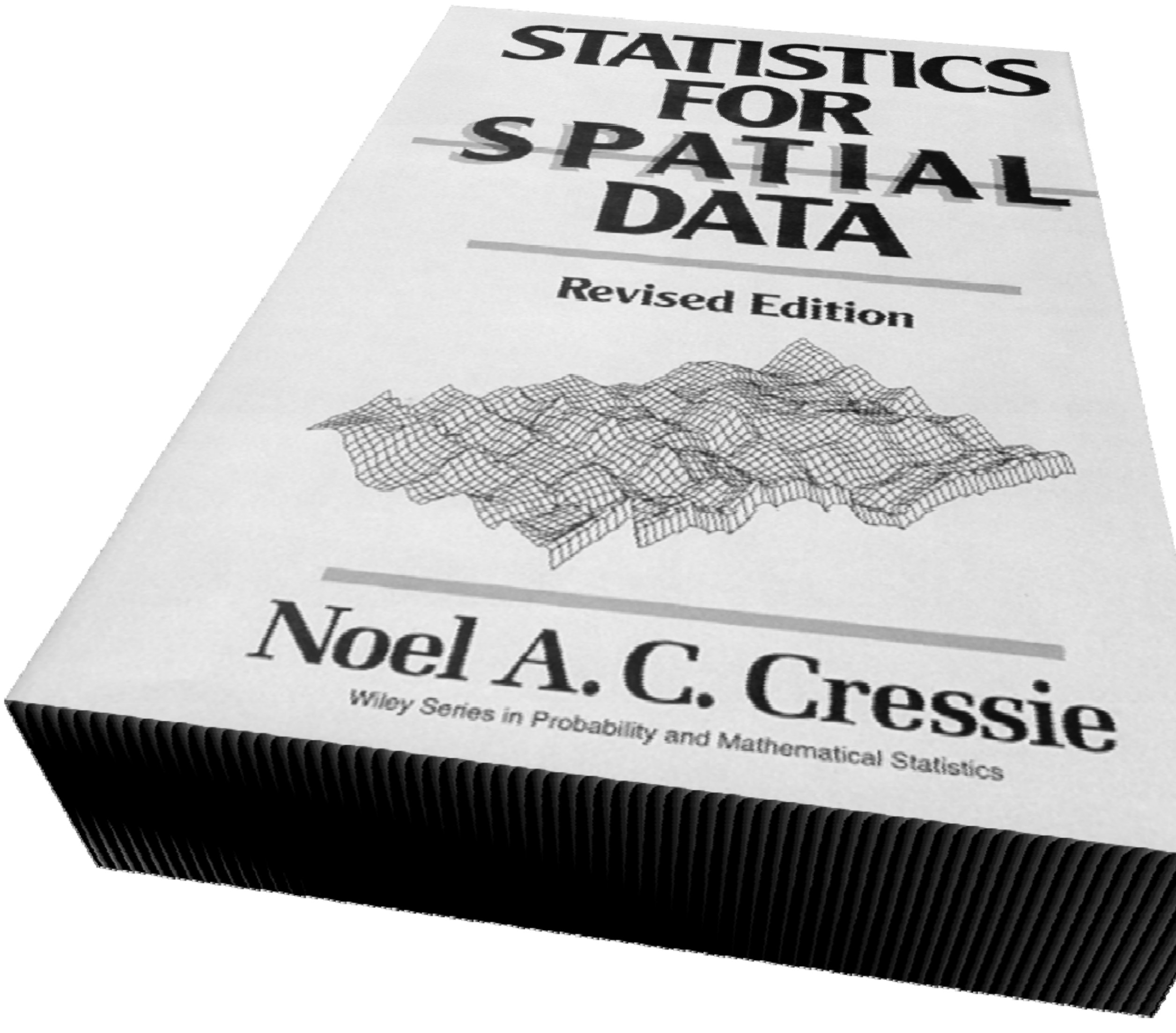
GIS

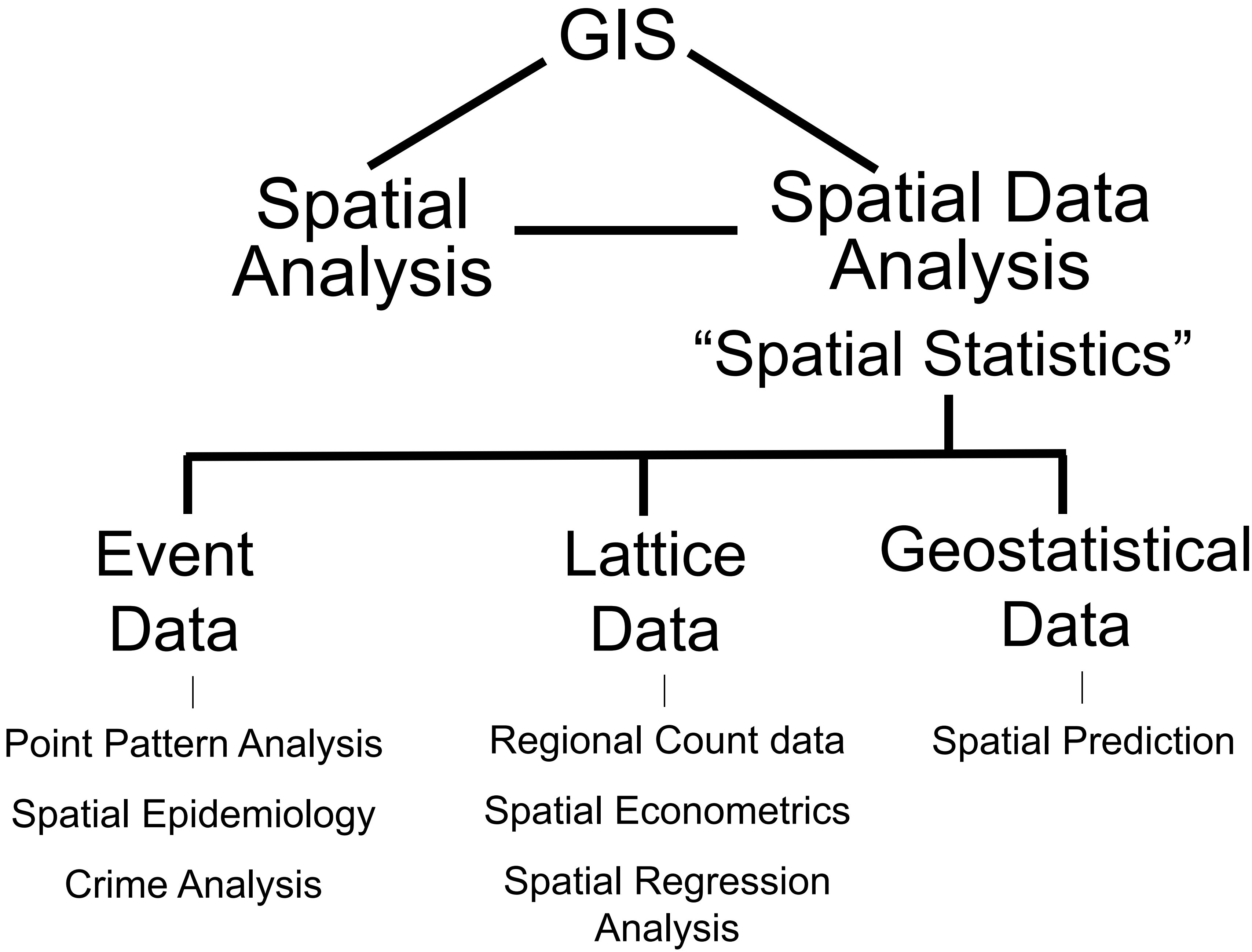
Spatial Analysis

Spatial Data Analysis

"Spatial Statistics"

Spatial Data Analysis




GIS
Spatial Analysis
Spatial Data Analysis
“Spatial Statistics”
Event Data
Lattice Data
Geostatistical Data
Point Pattern Analysis
Spatial Epidemiology
Crime Analysis
Regional Count data
Spatial Econometrics
Spatial Regression Analysis
Spatial Prediction

Types of Spatial Data

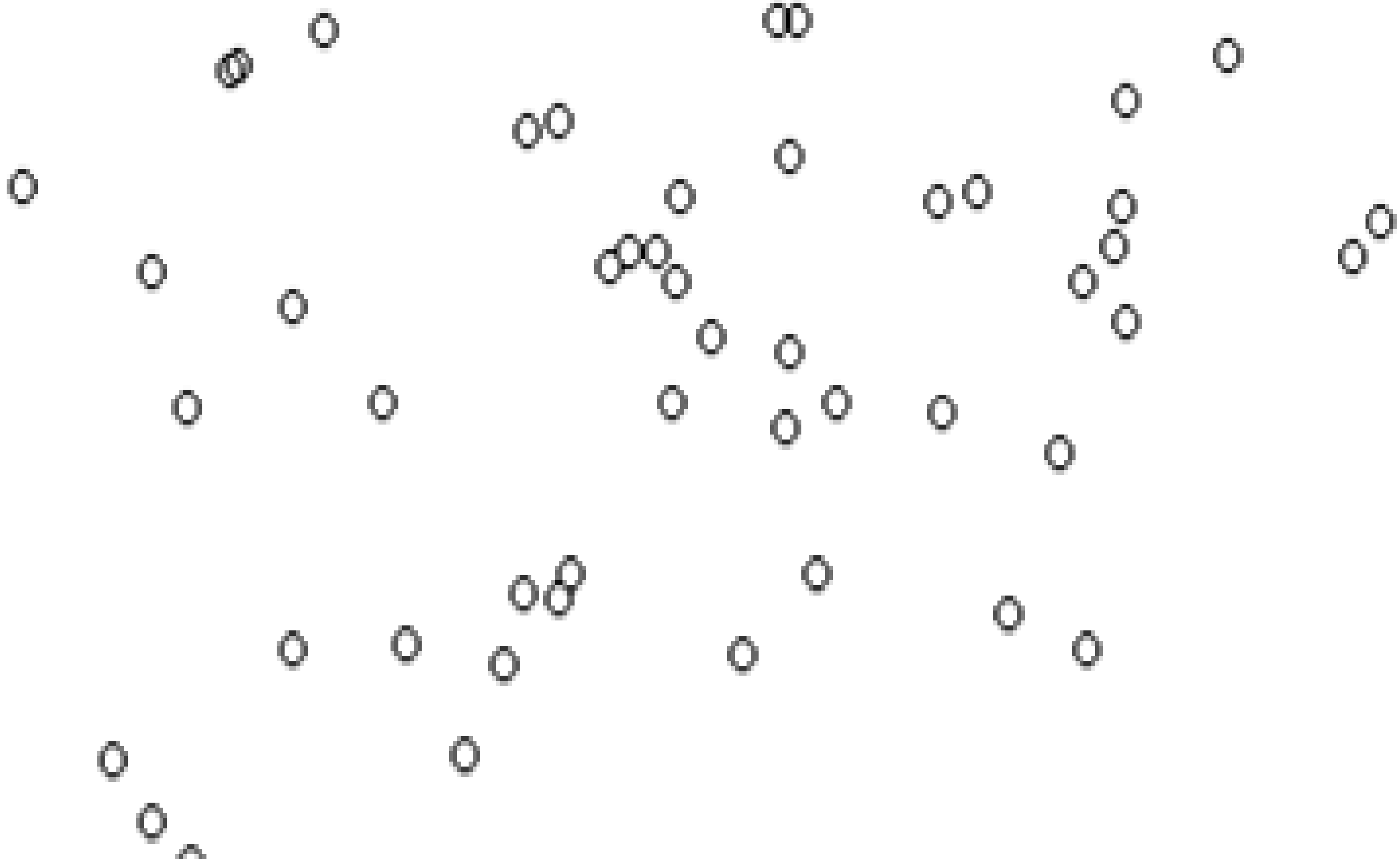
- Event data (point data)
- Spatially continuous data (geostatistical data)
- Lattice data (regionalized data)
- Spatial interaction data (flow data)

Focus this week

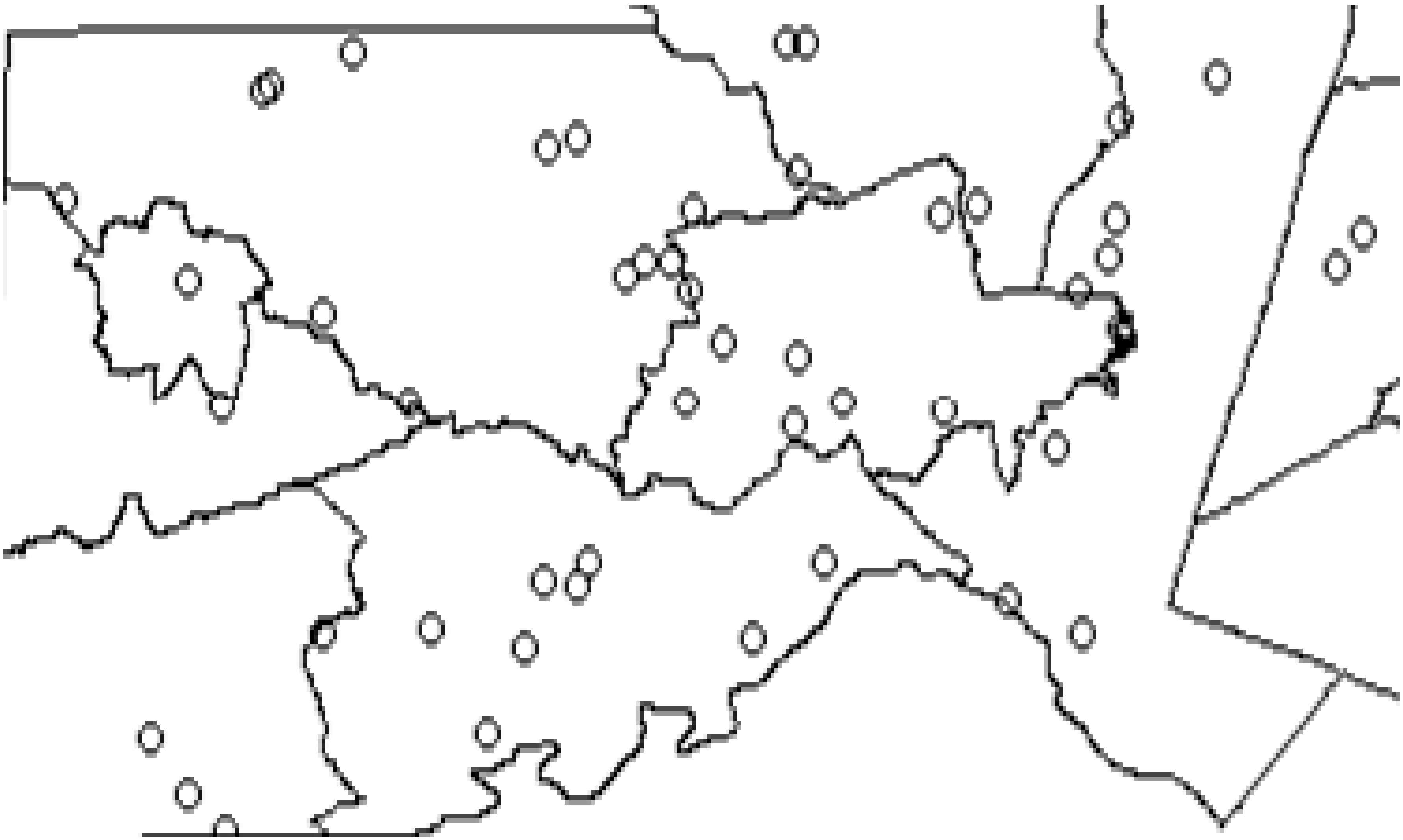
Discussion of these data types makes them appear as if they were strict, mutually exclusive, categories

Not necessarily the case

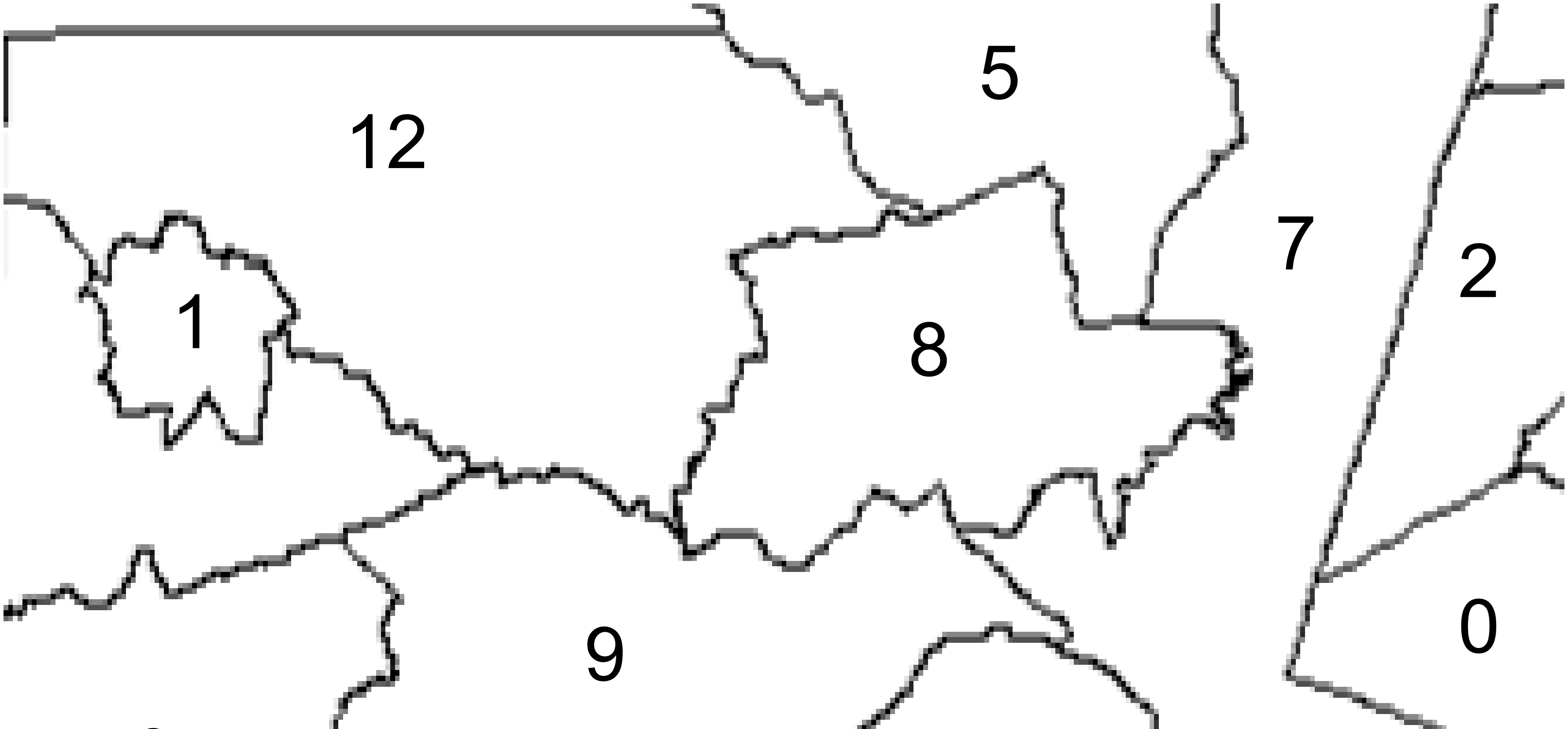
Let's say we observed this point pattern of events (of some type)



Let's also say we can geolocate these events in some relevant areal units

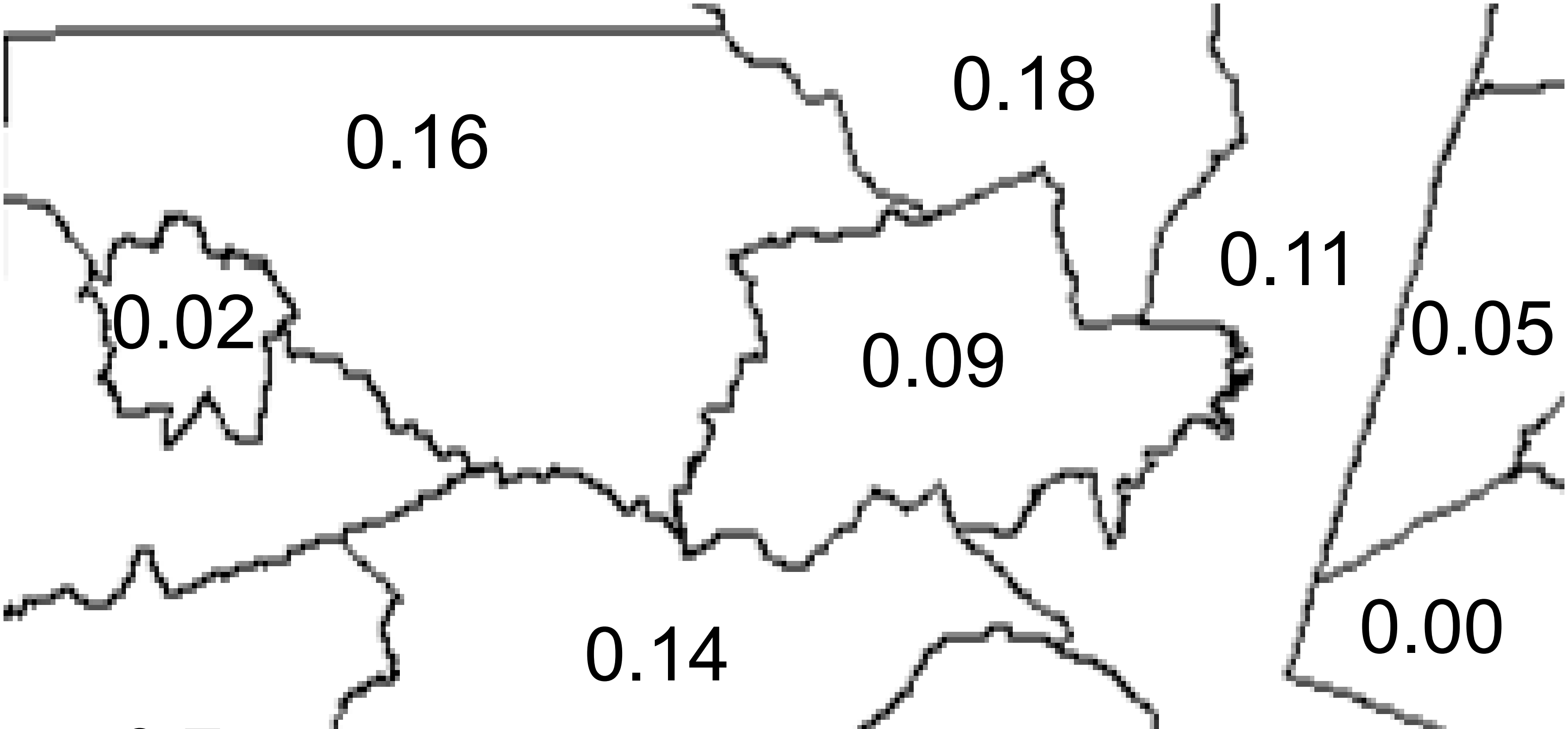


We then could generate the following regional count data



Regional count data are analyzed with a different toolkit than that used with event data

Finally, we might generate rates (e.g., prevalence proportions) from the regional count data



Analysis of rates opens up yet one more spatial analytic toolkit

Let's pause to clarify formally how spatial data (a spatial process or spatial random field) are often defined

- We will use the (2 x 1) column vector $\boldsymbol{s} = (s_1, s_2)^T$ to refer to a point location (coordinates in two dimensional space)
- Two point locations would be referred to by the two vectors $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$, where, in terms of coordinates, we have $\boldsymbol{s}_1 = (s_{11}, s_{12})^T$ and $\boldsymbol{s}_2 = (s_{21}, s_{22})^T$
- We will represent random variable Z at locations $\boldsymbol{s}_i$ in domain D in region $\mathcal{R}^d$ as the set of (possibly non-independent) random variables:

$$\{Z(\boldsymbol{s}),\ \boldsymbol{s} \in D \subset \mathcal{R}^d\}$$

- We will represent *realizations* of random variable Z as z_i , $\{i = 1,\ldots,n\}$ ($\boldsymbol{s}$ is implicit), or, in vector algebra, simply as the (n x 1) column vector $\boldsymbol{z}$

We generally will think of our task as wishing to understand (model) the structured aspects of our data, leaving behind a vector of random noise

Data = Structure + Error $\longleftarrow$ A common formulation in the statistical sciences

Data = Signal + Noise

$\boldsymbol{z}(\boldsymbol{s}) = f(\boldsymbol{X},\boldsymbol{s},\boldsymbol{\beta}) + \boldsymbol{u}(\boldsymbol{s})$

where $\boldsymbol{u}(\boldsymbol{s})$ is a random vector with mean $\mathbf{0}$ and variance $\mathrm{Var}[\boldsymbol{u}(\boldsymbol{s})] = \boldsymbol{\Sigma}(\boldsymbol{\theta})$

Data = Smooth + Spatial + Rough $\longleftarrow \eta_i + \varepsilon_i$

Data = 1^{st}-order process + 2^{nd}-order process + residual random effect

Lattice Data

- Have one or more variables whose values are measured over a set of areas (ideally mutually exclusive & exhaustive)
- Interest focuses on the attribute values, not on the locations which are known and unchanging
- Objective: Understand the spatial arrangement of attribute values, detect patterns, and examine relationships among the set of variables taking into account any spatial effects present
- Approach
  - exploratory spatial data analysis
  - confirmatory spatial data analysis (modeling & hypothesis testing)

The next 2-3 days will focus on analyzing data on a lattice with the previous data formulation either latent or explicit in our models

Some early cautions:

- Our goal is to correctly model and draw proper inferences about an unobserved, random DGP (random field)
  - spatial process?
    - spatial heterogeneity?
    - spatial dependence?
    - time?
  - sampling perspective?
- Spatial autocorrelation
- Scale issues; scale dependency; aggregation bias; boundary issues
- The tools are pretty good, but along the way many subjective decisions are made
  - defining “neighborhood”
  - choosing a weights matrix

Before taking a closer look at the meanings of Spatial Autocorrelation, Spatial Heterogeneity and Spatial Dependence, let’s take a closer look at the traditional OLS regression model

Standard OLS Regression

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i$$

In matrix notation: $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$

$(n \text{ x } 1)$ $(n \text{ x } k+1)$ $(k+1 \text{ x } 1)$ $(n \text{ x } 1)$

and where: $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$

$$\hat{\sigma}^2 = \left[1/(n-k-1) \right] e^T e$$

Okay… but what about the assumptions underlying the OLS regression model?

We must establish some conditions both on the population and on the data to establish unbiasedness, consistency and efficiency

These conditions are embodied in the Gauss-Markov Theorem

The Gauss-Markov Theorem asserts that $\hat{\beta}$ is a "*Best Linear Unbiased Estimator*" (*BLUE*) of β, provided the following assumptions are met:

- Linearity
- Mean independence $\mathrm{E}[\varepsilon_i \,|x_i] = 0$ (implies $\mathrm{E}[\boldsymbol{\varepsilon}] = 0$)
- Homoskedasticity and uncorrelated disturbances $\mathrm{Var/Cov}[\boldsymbol{\varepsilon}] = \mathrm{E}[\boldsymbol{\varepsilon}\,\boldsymbol{\varepsilon}'] = \sigma^2\boldsymbol{I}$
- $\boldsymbol{X}$ is of rank $k+1$ (k = no. of "independent" vars.)
- $\boldsymbol{X}$ is non-stochastic (or stochastic with finite second moments, and $\mathrm{E}[\boldsymbol{X}'\boldsymbol{\varepsilon}] = 0$ for unbiasedness)
- Normal disturbance

It is partly with these OLS assumptions in mind that we commence to get to know our data

EDA / ESDA

For example…

- Are we starting out by maximizing the probability of obtaining normal error structure?
  - Why do we care about this?
  - How can we check for this?
  - What can we do about it?
- Do we have good linear relationships between our dependent variable and independent variables?
  - Why do we care about this?
  - How can we check for this?
  - What can we do about it?
- Should any of our variables be transformed?
  - Do you know how to proceed?
- Do we have any outliers?
  - What kind of outliers?
  - What options are available to us?
- Fortunately almost everything we'll want to do can be done within *GeoDa* ™ And what we do in *GeoDa*, we'll try to replicate with R

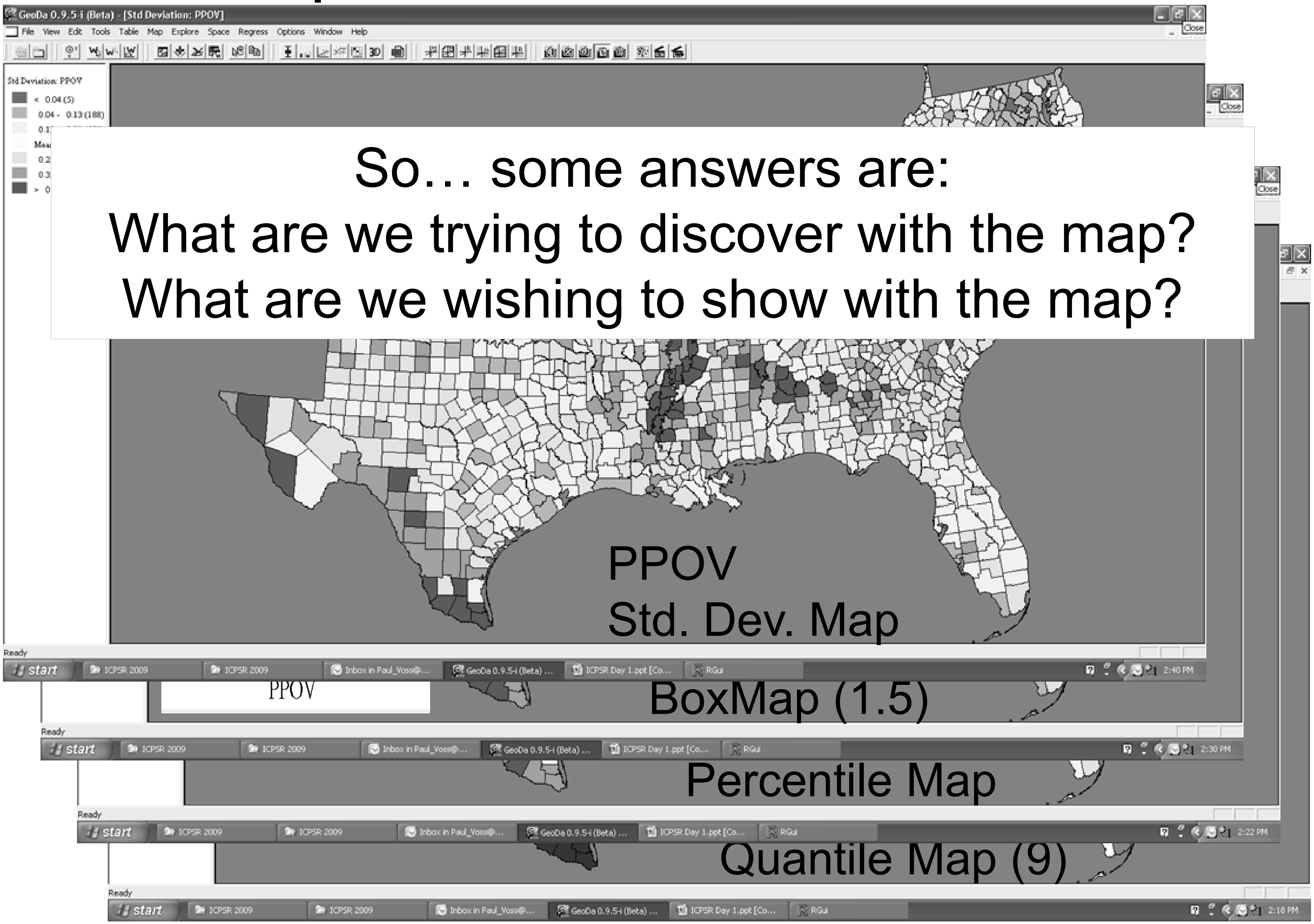
Exploring Spatial Data

- Goal is to seek good understanding and description of the data, thus suggesting hypotheses to explore
- Not much emphasis here on p-values, which are so ubiquitous in most of our training & our statistical instincts
- Look especially for clues to “spatial heterogeneity” or “spatial dependence”
- Few *a priori* assumptions about the data
- Analysis may sometimes end here
- EDA/ESDA: Despite good tools, at heart EDA is a philosophy; an attitude; “best practice” way of thinking about your data; a way of staying out of trouble

EDA / SDA Tools

- Maps
- Descriptive statistics
- Plots and graphics
- Classification and clustering methods
- Software with dynamically linked objects
- Global and local spatial autocorrelation

Maps? Sure, but what kind?



So… some answers are:

What are we trying to discover with the map?

What are we wishing to show with the map?

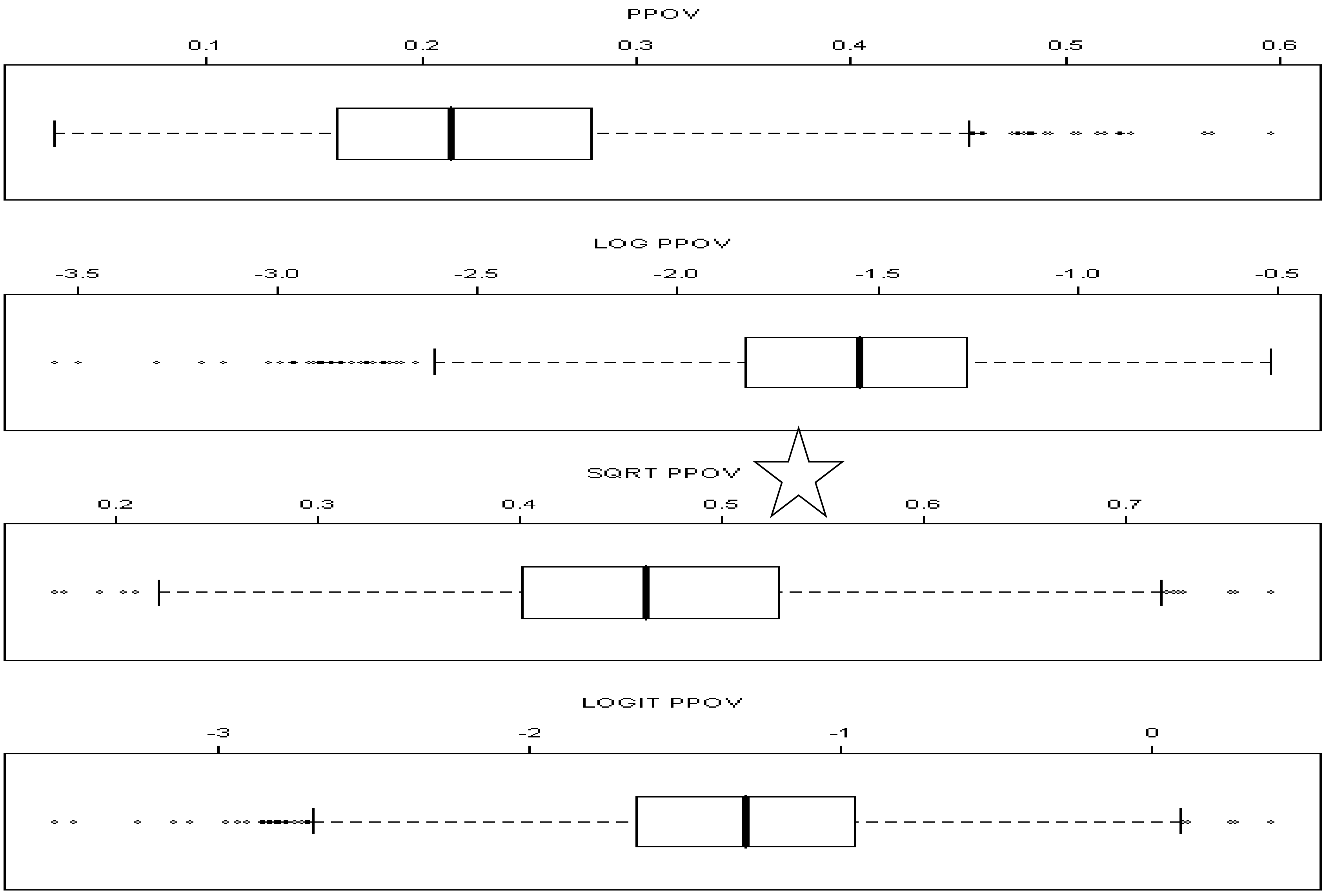
Descriptive Statistics

Untitled - Notepad

File Edit Format View Help

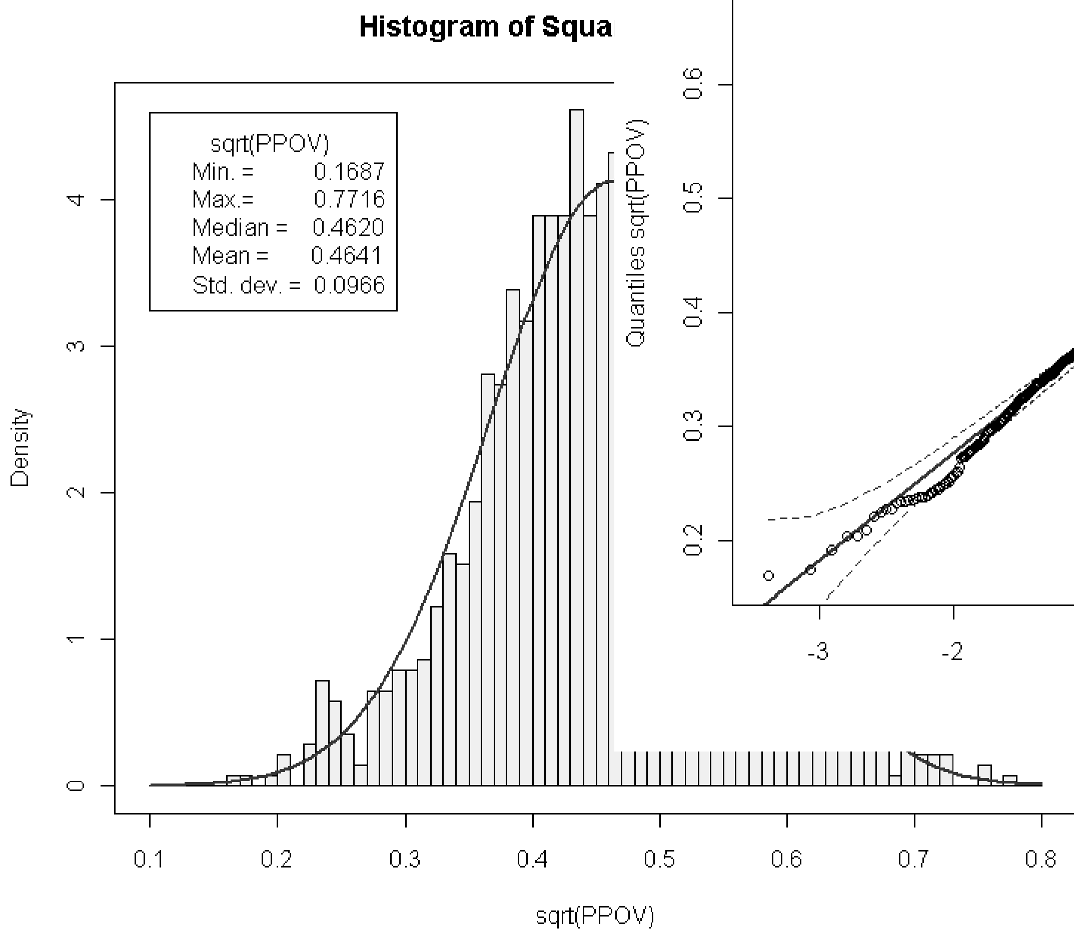
```
> summary(socotxt)

         NAME              STUSAB             FIPS              PPOV              PHSP                PFHH
 Abbeville_SC:   1    TX      :254    Min.   : 1001    Min.   :0.02846   Min.   :0.002459   Min.   :0.04444
 Acadia_LA   :   1    GA      :159    1st Qu.:13272    1st Qu.:0.16080   1st Qu.:0.010134   1st Qu.:0.17204
 Accomack_VA :   1    KY      :120    Median :37065    Median :0.21345   Median :0.019164   Median :0.20162
 Adair_KY    :   1    NC      :100    Mean   :31600    Mean   :0.22470   Mean   :0.072395   Mean   :0.22009
 Adair_OK    :   1    VA      : 99    3rd Qu.:48122    3rd Qu.:0.27864   3rd Qu.:0.056986   3rd Qu.:0.25408
 Adams_MS    :   1    TN      : 95    Max.   :54109    Max.   :0.59530   Max.   :0.975390   Max.   :0.51722
 (Other)     :1381    (Other):560
     PWKCO             PHSLS             PUNEM              PUDEM             PEXTR               PPSRV
 Min.   :0.1379   Min.   :0.2468   Min.   :0.00000   Min.   :0.1033   Min.   :0.000772   Min.   :0.2002
 1st Qu.:0.4858   1st Qu.:0.5636   1st Qu.:0.04573   1st Qu.:0.1810   1st Qu.:0.020914   1st Qu.:0.3054
 Median :0.6385   Median :0.6464   Median :0.05740   Median :0.2041   Median :0.043319   Median :0.3471
 Mean   :0.6260   Mean   :0.6257   Mean   :0.06219   Mean   :0.2105   Mean   :0.060918   Mean   :0.3506
 3rd Qu.:0.7603   3rd Qu.:0.7003   3rd Qu.:0.07380   3rd Qu.:0.2320   3rd Qu.:0.078534   3rd Qu.:0.3910
 Max.   :0.9680   Max.   :0.8276   Max.   :0.20883   Max.   :0.4537   Max.   :0.486631   Max.   :0.5953

     PMSRV              PNDMPG             PNHSPW             PMNRTY             WGHT                LO_POV
 Min.   :0.03499   Min.   :0.00000   Min.   :0.02019   Min.   :0.01043   Min.   :2.029e+00   Min.   :-3.5305
 1st Qu.:0.09525   1st Qu.:0.03500   1st Qu.:0.61559   1st Qu.:0.09670   1st Qu.:5.642e+02   1st Qu.:-1.6523
 Median :0.10959   Median :0.06371   Median :0.76509   Median :0.23491   Median :1.052e+03   Median :-1.3042
 Mean   :0.11339   Mean   :0.07533   Mean   :0.73602   Mean   :0.26398   Mean   :2.641e+03   Mean   :-1.3195
 3rd Qu.:0.12699   3rd Qu.:0.10169   3rd Qu.:0.90330   3rd Qu.:0.38441   3rd Qu.:2.166e+03   3rd Qu.:-0.9512
 Max.   :0.40364   Max.   :0.42973   Max.   :0.98957   Max.   :0.97981   Max.   :1.546e+05   Max.   : 0.3859

     YCOORD              XCOORD              PFRN                PNAT                PBLK              P65UP
 Min.   :-1537226   Min.   :-922316   Min.   :0.000000   Min.   :0.000000   Min.   :0.00000   Min.   :0.01801
 1st Qu.: -847128   1st Qu.: 204867   1st Qu.:0.008562   1st Qu.:0.001990   1st Qu.:0.02296   1st Qu.:0.11799
 Median : -580195   Median : 871489   Median :0.017704   Median :0.003029   Median :0.09455   Median :0.13858
 Mean   : -590247   Mean   : 707406   Mean   :0.033020   Mean   :0.010200   Mean   :0.16698   Mean   :0.14122
 3rd Qu.: -334488   3rd Qu.:1209457   3rd Qu.:0.039692   3rd Qu.:0.005393   3rd Qu.:0.27851   3rd Qu.:0.15942
 Max.   :  161823   Max.   :1712046   Max.   :0.509357   Max.   :0.424850   Max.   :0.86489   Max.   :0.34716
```

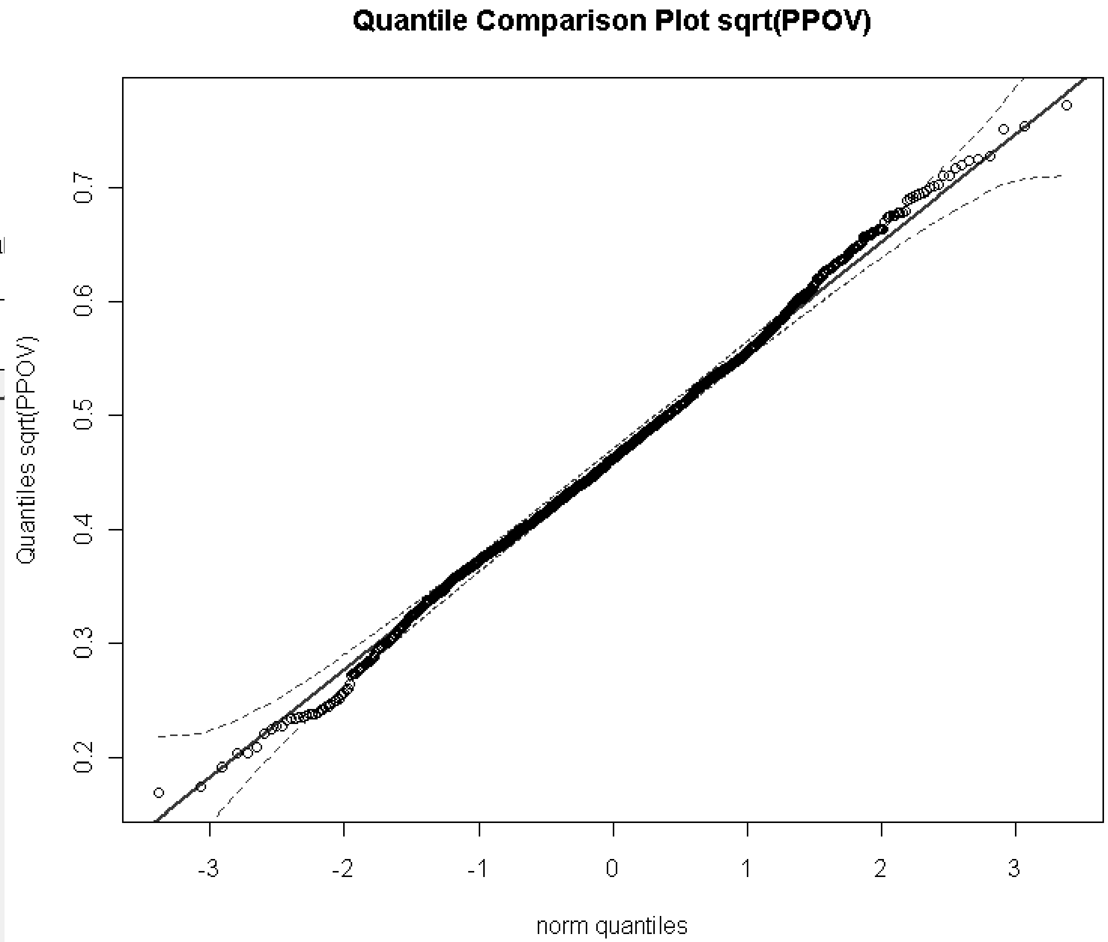
Plots and Graphs



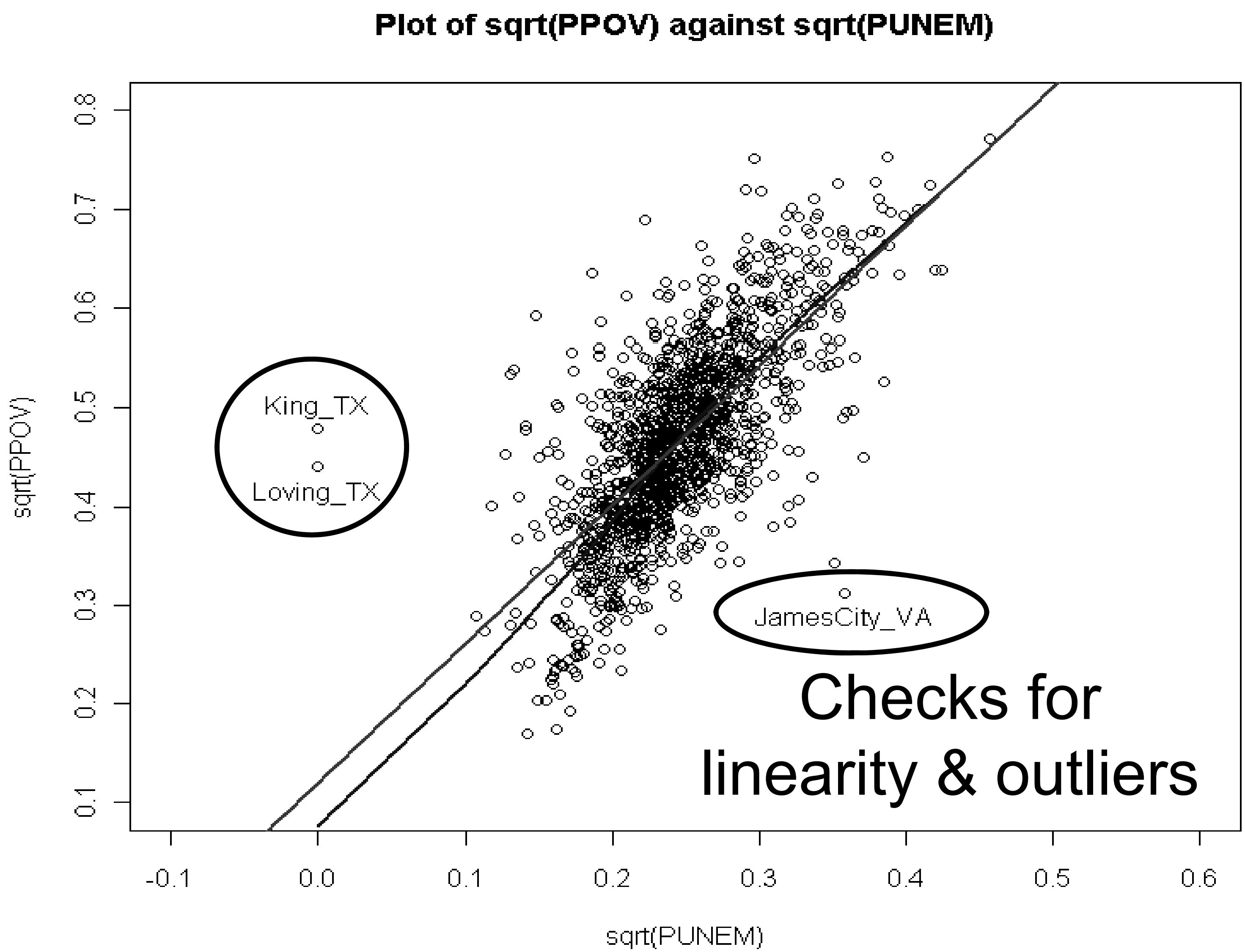
Plots and Graphs… (cont.)

Checks for normality and symmetry





Plots and Graphs… (cont.)



This afternoon we will take a look at how to do some of this EDA/ESDA in *GeoDa* & R

EDA/ESDA is less a toolkit than it is a philosophy or attitude regarding the task you face

Tomorrow we'll take a close look at the concept of spatial autocorrelation

… in the context of assumptions underlying the OLS model, this raises some serious problems

e.g., correlated disturbances:

$$\mathrm{Cov}[\boldsymbol{\varepsilon}] = \mathrm{E}[\boldsymbol{\varepsilon}\,\boldsymbol{\varepsilon}'] = \sigma^2\,\boldsymbol{\Sigma} \neq \sigma^2\,\boldsymbol{I}$$

- This makes OLS estimates of the t-test values unreliable; i.e., the OLS estimates are relatively inefficient
- *Second*, it inflates the value of the R^2 statistic

Correlation between $\boldsymbol{X}$ and $\boldsymbol{\varepsilon}$

$$\mathrm{E}[\boldsymbol{X}'\boldsymbol{\varepsilon}] \neq 0$$

$$\begin{aligned}
\mathrm{E}[\hat{\boldsymbol{B}}] &= \mathrm{E}[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}] \\
&= \mathrm{E}[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})] \\
&= \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\mathrm{E}[\boldsymbol{X}'\boldsymbol{\varepsilon}] \neq \boldsymbol{\beta}
\end{aligned}$$

OLS parameter estimates are biased

$$\mathrm{plim}[\hat{\boldsymbol{B}}] = \boldsymbol{\beta} + \mathrm{plim}[(\boldsymbol{X}'\boldsymbol{X}/n)^{-1}] \times \mathrm{plim}[\boldsymbol{X}'\boldsymbol{\varepsilon}/n]$$

OLS parameter estimates are inconsistent

Which means… we gotta do something about it!

and that's where we're headed over the next two days

Readings for today

- Anselin, Luc. 1989. "What is Special About Spatial Data? Alternative Perspectives on Spatial Data Analysis." *NCGIA Technical Paper 89-4*.
- Anselin, Luc. 2010. "Thirty Years of Spatial Econometrics." *Papers in Regional Science* 89(1):3-25.]
- Galle, Omer R., Walter R. Gove, & J. Miller McPherson. 1972. "Population Density and Pathology: What Are the Relations for Man?" *Science* (new series) 176:23-30.
- Loftin, Colin and Sally K. Ward. 1983. "A Spatial Autocorrelation Model of the Effects of Population Density on Fertility." *American Sociological Review*, 48(1):121-128.
- Anselin, Luc. 2005. *Exploring Spatial Data with GeoDa: A Workbook*, (chapters 2, 3, 7-12).
- Anselin, Luc. 2005. *Spatial Regression Analysis in R: A Workbook,* (chapters 1 & 2).
- Venables, W.N. & D.M. Smith. 2010. *An Introduction to R*.
- Messner, Steven F., et al. 1999. "The Spatial Patterning of County Homicide Rates: An Application of Exploratory Spatial Data Analysis." *Journal of Quantitative Criminology* 15(4):423-450

Afternoon Lab

Introduction to *GeoDa*™ and R

Questions?