Introduction to Spatial Regression Analysis

Paul Voss
UNC Chapel Hill

Day 3

Review of yesterday

- Global & local spatial autocorrelation
  - Moran's I
  - Geary's c
  - LISA statistics
  - Moran scatterplot
- Weights matrices
- Spatial lag operator
- Spatial processes
  - Spatial heterogeneity
  - Spatial dependence

Questions?

Plan for today

- Spatial processes
  - spatial heterogeneity
  - spatial dependence
- Spatial regression models
- Various specifications for spatial dependence
  - spatial lag model
  - spatial error model
  - higher-order models
- Afternoon lab
  - spatial regression modeling in *GeoDa* & R

Recall, we said yesterday: When spatial autocorrelation in our data is indicated…

- At least one assumption of the standard linear regression model is violated (the classical independence assumption)
- The latent information content in the data is diminished
- We need to do something about it:
  - get rid of it; model it away
  - take advantage of it; bring it into the model
- Either spatial dependence or spatial heterogeneity (or both) should be entertained as potential data-generating models

Worth repeating…

For many spatial analysts, the term spatial heterogeneity refers to variation in *relationships* over space.

So, how do we proceed?

There's no agreed-upon formal roadmap for how to conduct a spatial data analysis, but certainly some steps must precede other.

Usually it goes something like this…

Recommended Steps in Spatial Data Analysis (1)

- EDA on variables; ESDA on variables; look for global and local patterns of spatial autocorrelation under different neighborhood specifications
  - put your theory hat on, consider possible structural covariates of dependent variable
  - transform variables as necessary; outliers?
  - visually inspect your maps; outliers?
  - test different weights matrices
  - global and local tests for spatial autocorrelation
  - examine Moran scatterplot; outliers?
  - decisions about outliers
  - look for extent of, and possible amelioration of, spatial heterogeneity

Recommended Steps in Spatial Data Analysis (2)

- EDA on variables; ESDA on variables; look for global and local patterns of spatial autocorrelation under different neighborhood specifications
- OLS baseline model and accompanying diagnostics
  - Specify model and run in OLS; iterate this for other specifications
  - map residuals & be on lookout for such things as geographic clustering, variance nonstationarity, possible spatial regimes; outliers?
  - examine the diagnostics; where are your problems?
  - What do the LM diagnostics suggest wrt spatial dependence modeling
  - run model using GWR to further understand spatial structural variance

Recommended Steps in Spatial Data Analysis (3)

- EDA on variables; ESDA on variables; look for global and local patterns of spatial autocorrelation under different neighborhood specifications
- OLS baseline model and accompanying diagnostics
- Correct for spatial heterogeneity if indicated
  - carefully select covariates
  - surface trend fitting
  - spatial regime analysis

Recommended Steps in Spatial Data Analysis (4)

- EDA on variables; ESDA on variables; look for global and local patterns of spatial autocorrelation under different neighborhood specifications
- OLS baseline model and accompanying diagnostics
- Correct for spatial heterogeneity if indicated
- With possible controls for spatial heterogeneity, estimate and compare spatial models
  - spatial lag model?
  - spatial error model?
  - mixed lag & error model (SARAR)?
  - what's your theory?
  - estimator?

Recommended Steps in Spatial Data Analysis (5)

- EDA on variables; ESDA on variables; look for global and local patterns of spatial autocorrelation under different neighborhood specifications
- OLS baseline model and accompanying diagnostics
- Correct for spatial heterogeneity if indicated
- With possible controls for spatial heterogeneity, estimate and contrast spatial error and spatial lag model results
- Iterate these steps as necessary

So, that's where we're headed today

Questions?

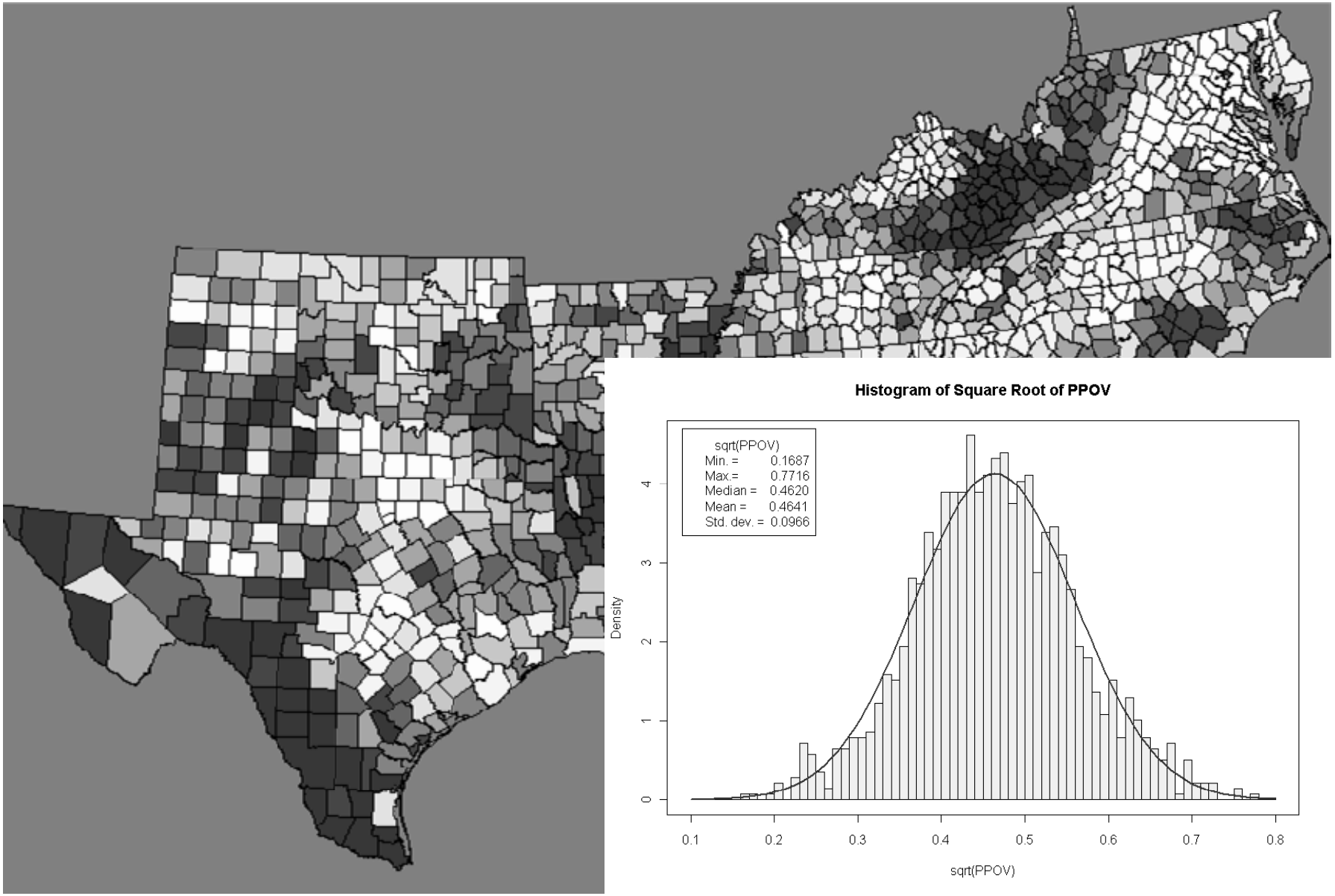
Carrying out a Spatial Data Analysis. Recall Step 1…

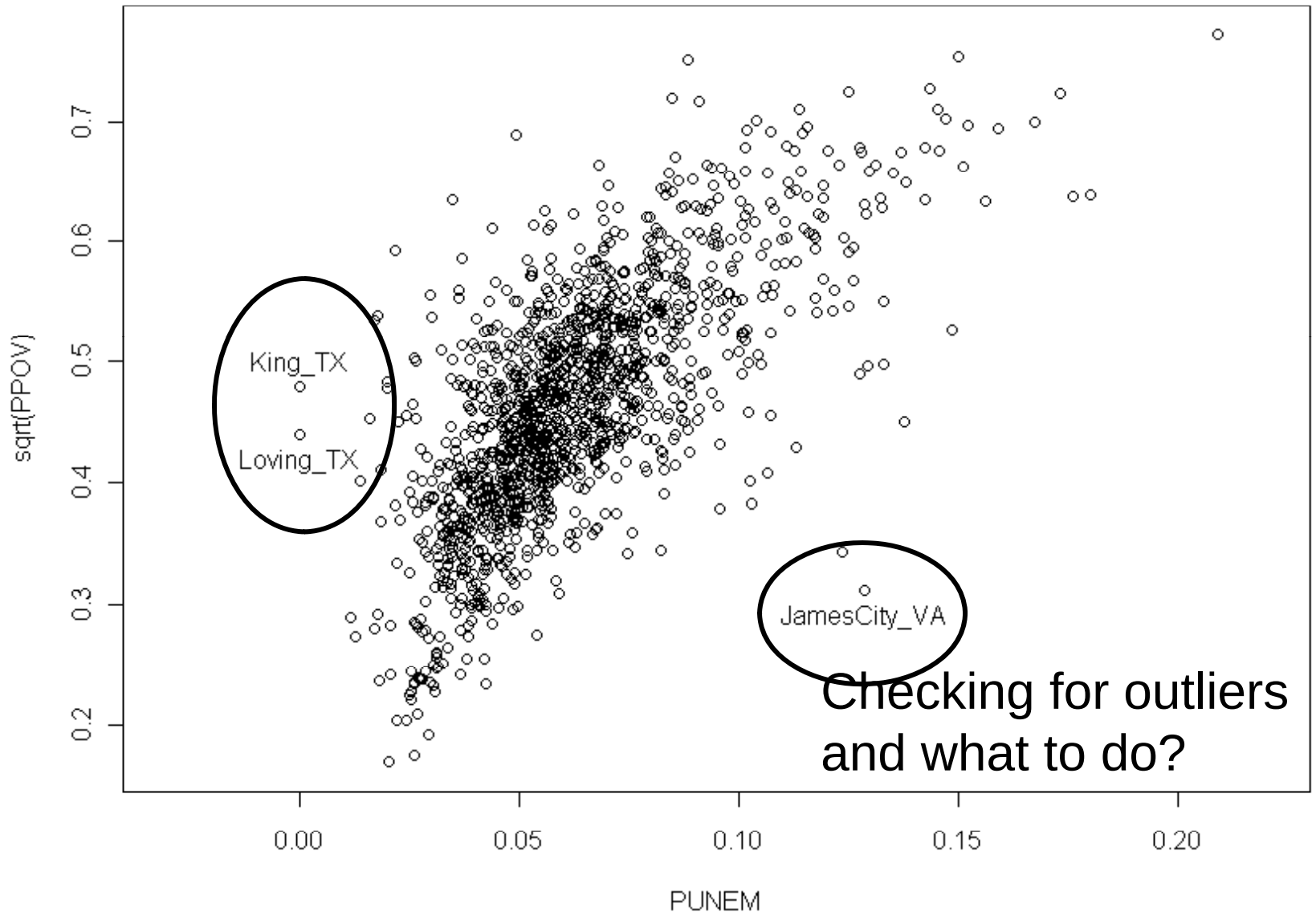
- EDA on variables; ESDA on variables; look for global and local patterns of spatial autocorrelation under different neighborhood specifications
  - put your theory hat on, consider possible structural covariates of dependent variable
  - transform variables as necessary; outliers?
  - visually inspect your maps; outliers?
  - test different weights matrices
  - global and local tests for spatial autocorrelation
  - examine Moran scatterplot; outliers?
  - decisions about outliers
  - look for extent of, and possible amelioration of, spatial heterogeneity

Visualizing Spatial Data

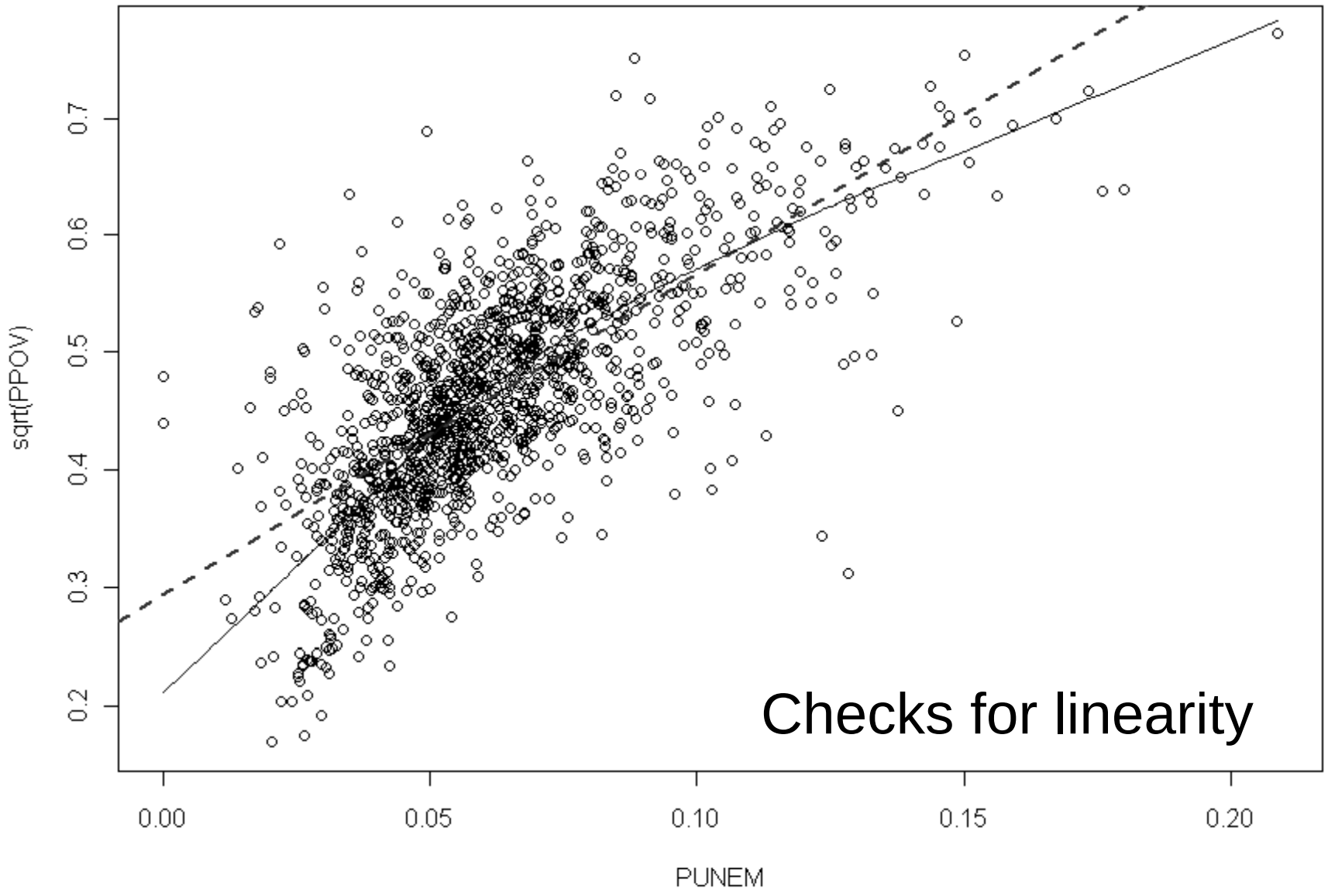
- Part of your ESDA
- Goal is to “see” the data; map the data; plot the data; look for patterns
- Mapping software is a fundamental tool
- Statistical analysis software is a fundamental tool

Square root transformation of PPOV variable

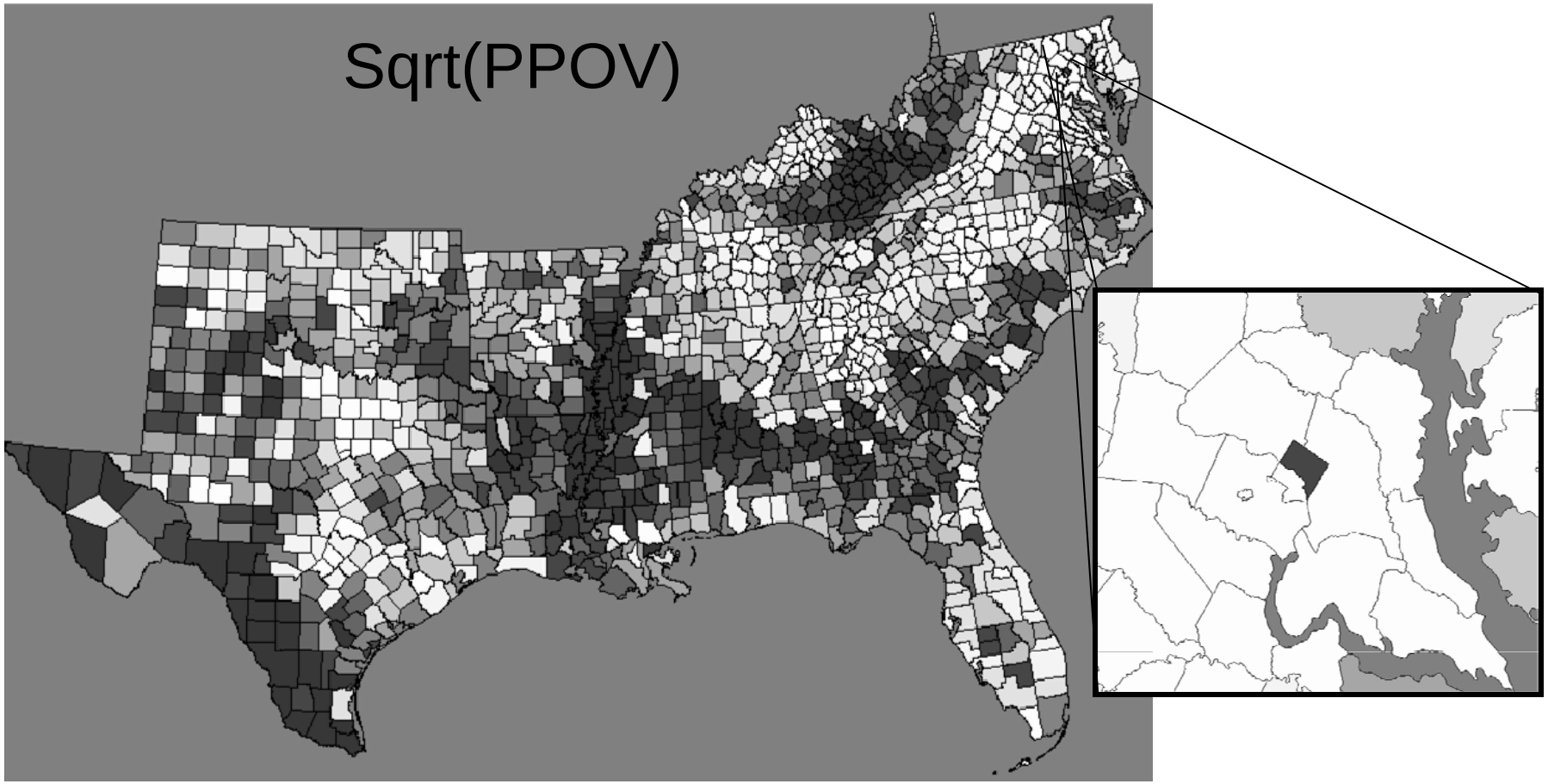




Checking for outliers and what to do?


sqrt(PPOV) by PUNEM
sqrt(PPOV)
0.2
0.3
0.4
0.5
0.6
0.7
0.00
0.05
0.10
0.15
0.20
PUNEM
Checks for linearity

Very often a unit of observation may not stand out as an outlier in any univariate or bivariate plots, but might be a “spatial outlier”



Exploring Spatial Data with an eye on spatial processes:
Spatial Heterogeneity
Spatial Dependence

Exploring 1st Order Variation Spatial Heterogeneity

- Mapping
  - Looking for & gaining some understanding of patterns in the variables
    - Similar map patterns among different variables
    - "Opposite" map patterns for some variables
  - Looking for global trend or "drift" in the data (especially in your response variable)
    - Might there be something to model using our spatial coordinates?
  - Looking for spatial outliers

Exploring 1st Order Variation Spatial Heterogeneity

- Mapping
- Clustering
  - Geodemographic clustering
    - Mapping of clusters
    - Very useful device for spatial sampling

Exploring 1st Order Variation Spatial Heterogeneity

- Mapping
- Clustering
- Spatial moving averages

$$\hat{\mu}_i = \bar{y}_i = \frac{\sum_{j=1}^{n} w_{ij} y_j}{\sum_{j=1}^{n} w_{ij}}$$

Exploring 1st Order Variation Spatial Heterogeneity

- Mapping
- Clustering
- Spatial moving averages
- Regression
  - Trend surface
  - Geographically Weighted Regression

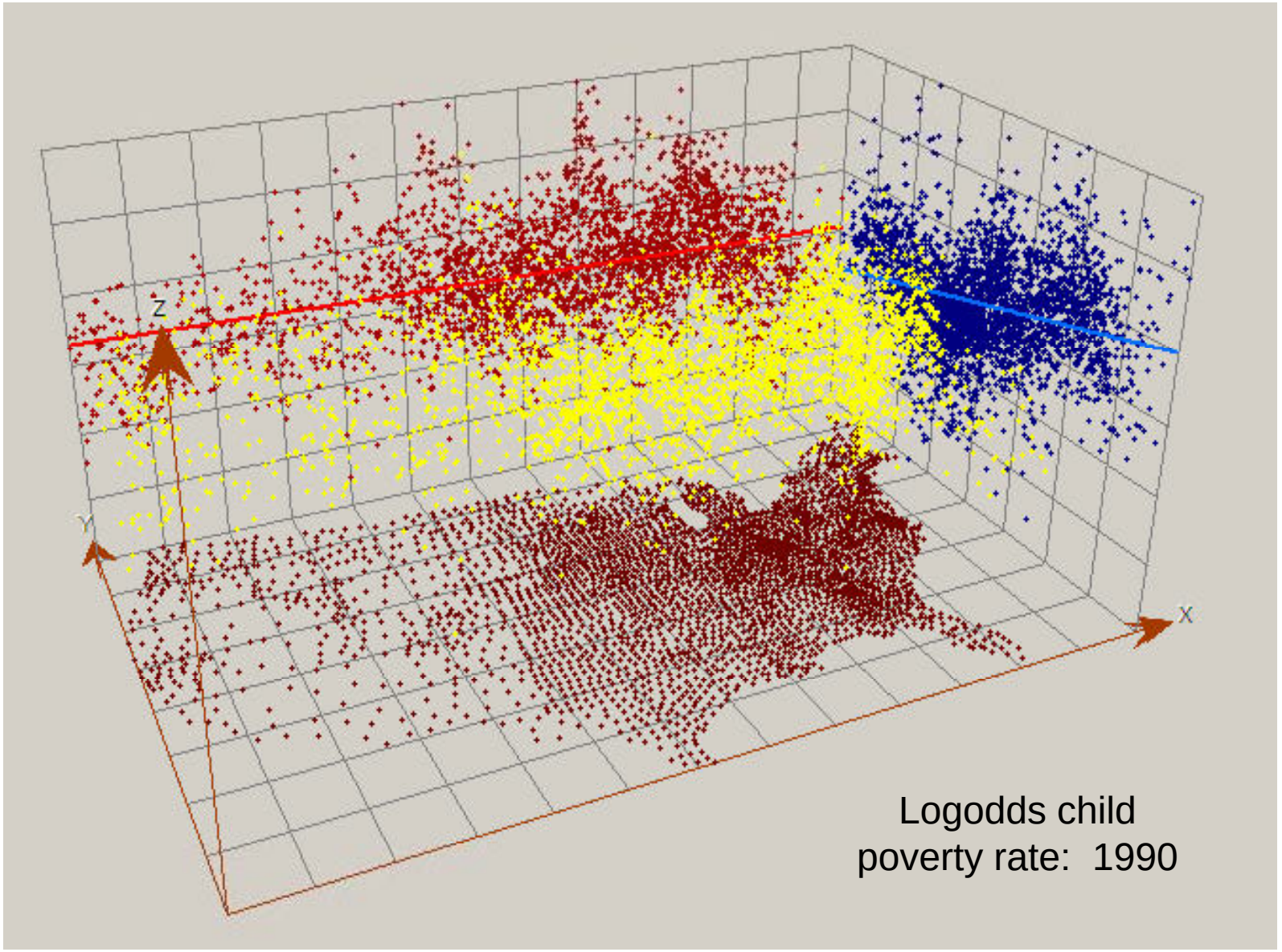
Trend Surface Regression

- Spatial drift in mean
  - polynomial regression in coordinates of the observations (x,y)

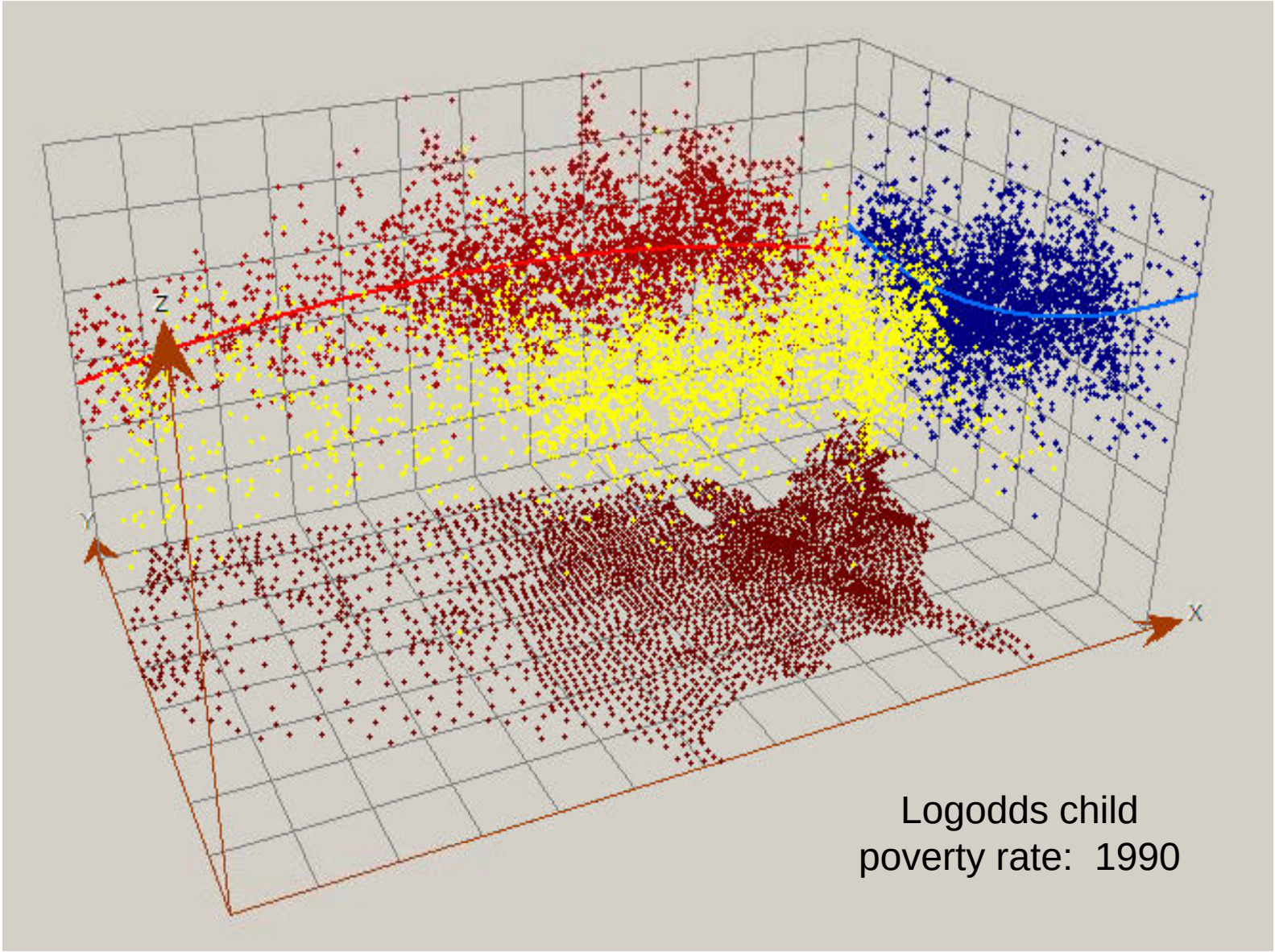
$$z = \alpha + \beta_1 x + \beta_2 y + \beta_3 x^2 + \beta_4 y^2 + \beta_5 xy + \varepsilon$$

- Interpretation/problems
  - spatial interpolation
  - no meaningful substantive interpretation (geographic determinism)
  - multicollinearity
  - problems at the boundaries of study area

First-Order Trend Surface?

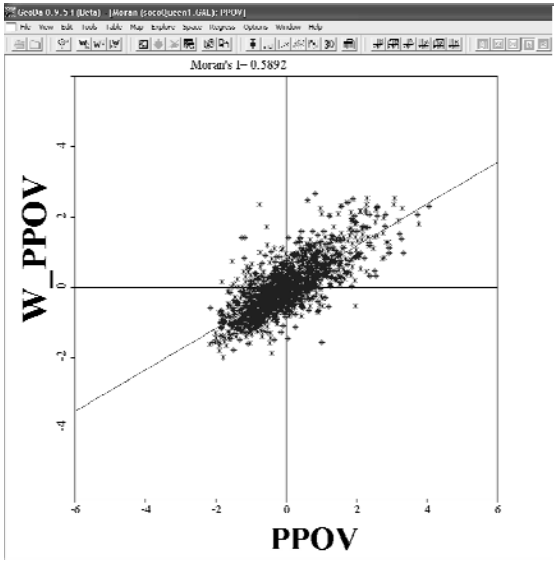


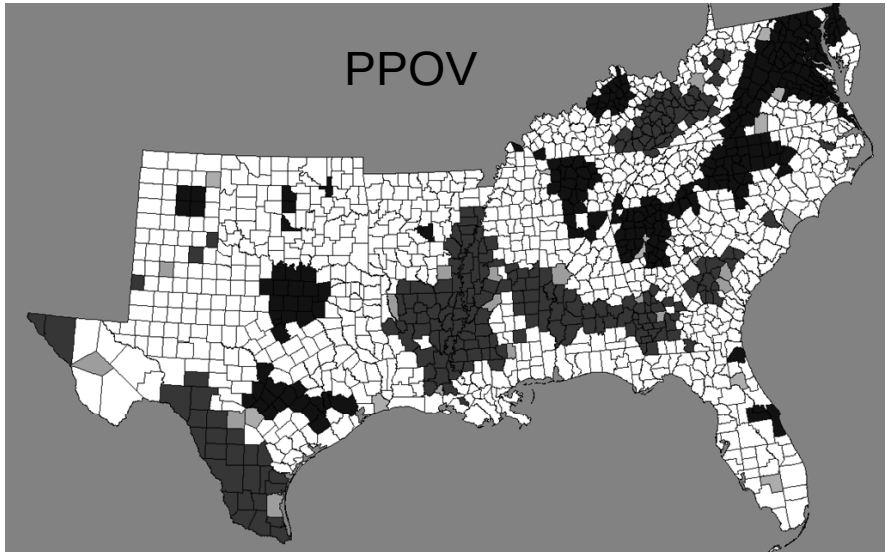
Second-Order Trend Surface?



Two useful devices for exploring local spatial autocorrelation in ESDA (reminder from yesterday)

- Moran scatterplot
- LISA statistics
- Both are based on the notion of a *local* spatial autocorrelation statistic





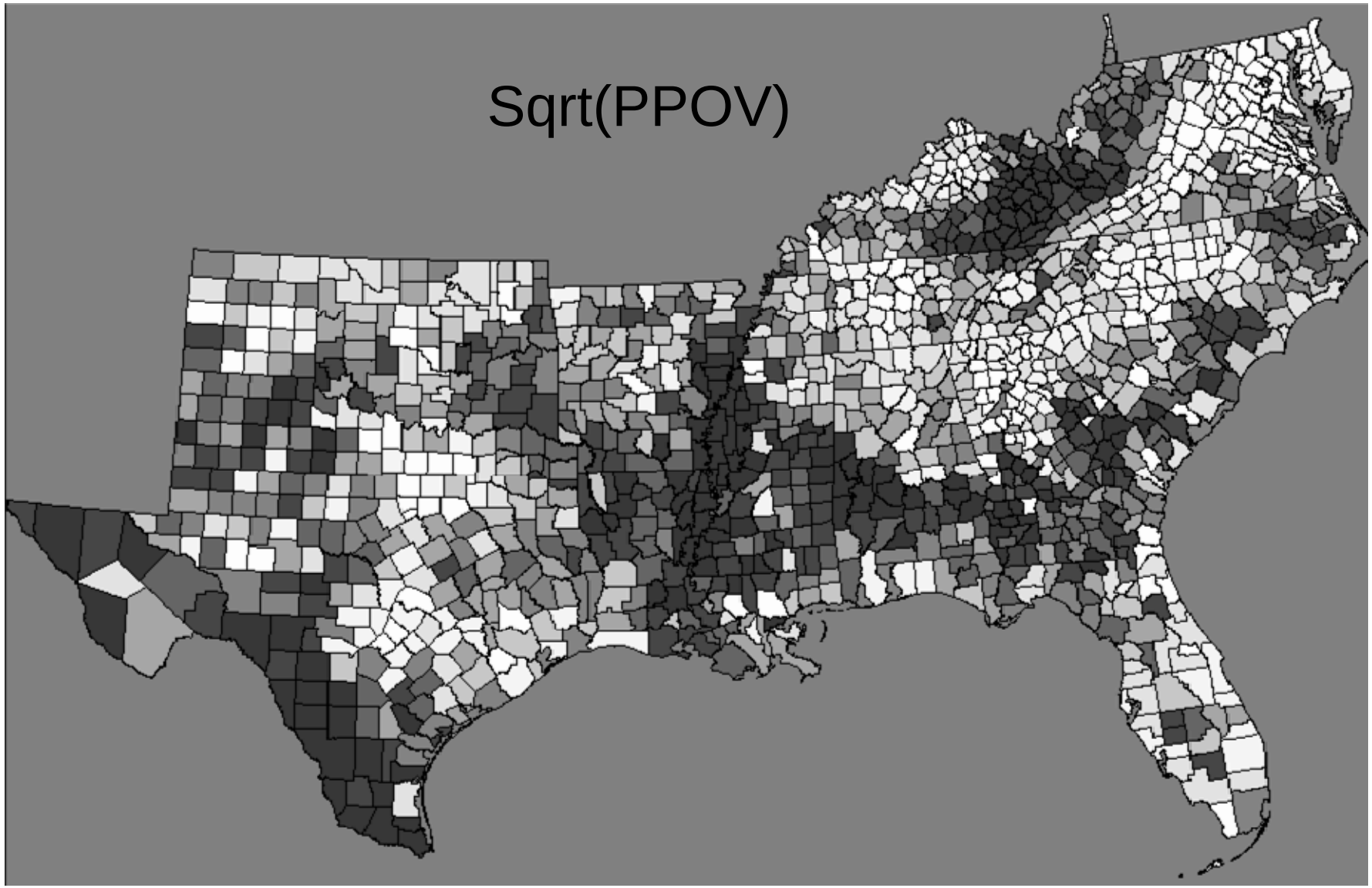
Local Indicators of Spatial Association (LISA)

- Assess assumptions of stationarity
- Indicate local regions of non-stationarity ("hotspots" or "pockets")
- Allow for decomposition of global measure into contributions of individual observations
- Identify outliers or spatial regimes

Spatial autocorrelation as a nuisance

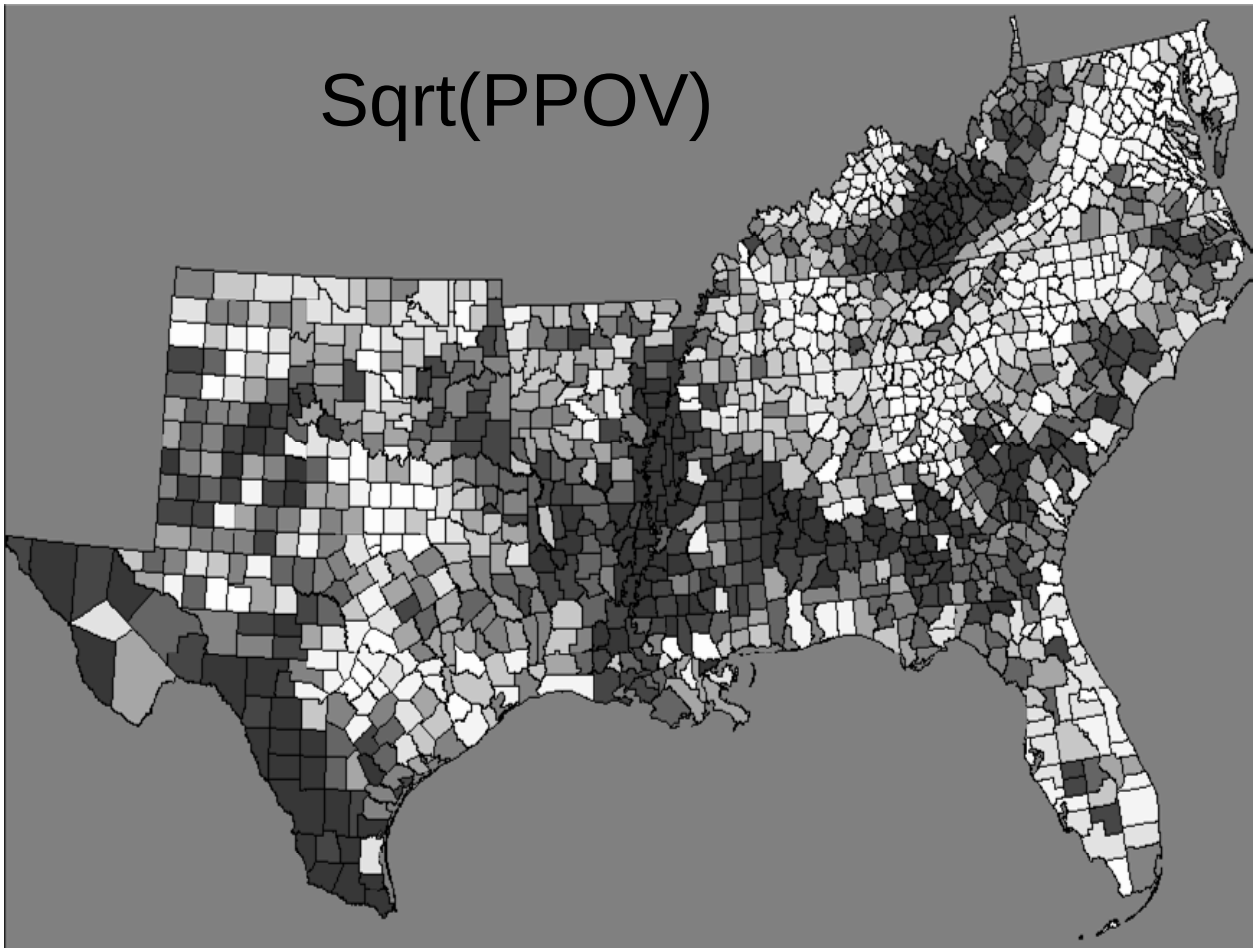
Or, better said, spatial autocorrelation arising from a mismatch between a spatial process and your particular window on that process

Nuisance autocorrelation: Mismatch between the spatial process and the unit of observation



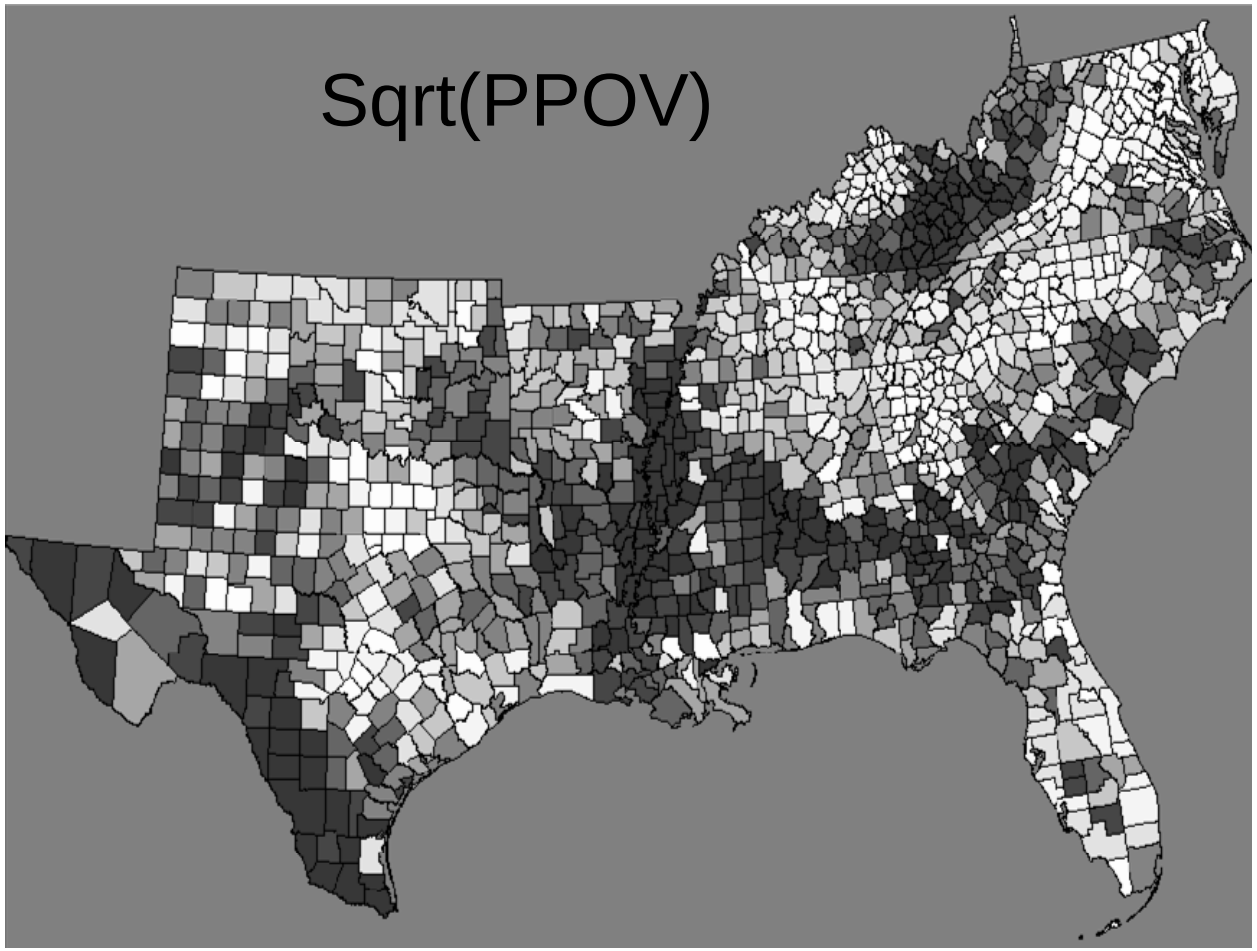
Spatial autocorrelation as a substantive process

Spatial autocorrelation as a substantive process



- Grouping processes
- Group-dependent processes
- Feedback processes

Said another way: consider some (unknown) spatial process and associated attribute values for areas across region



- Interaction?
- Reaction (to some other set of variables)?
- Nuisance?

If reaction…

- Then a regression structure is appropriate to think about
- Focus is on spatial heterogeneity

If interaction…

- Then we must consider a model with a non-diagonal covariance structure
- Focus is on spatial dependence (spatial interaction)

If both reaction and interaction are believed to be at work:

- Spatial regression model; spatial heterogeneity in the design matrix; spatial interaction in the residuals (“spatial error model”); or…
- Spatial regression model; spatial heterogeneity in the design matrix; with an explicit expression controlling for spatial interaction in the dependent variable (“spatial lag model”)
- Both?

We’ve arrived (finally) at the topic Spatial Modeling

One of the earliest spatial econometric models explored was the "Autocorrelated Errors Model" or Spatial Error Model

$$y = X\beta + u$$

$$u = \lambda W u + \varepsilon$$

First-order variation comes only through $X\beta$; second-order variation is represented as an autoregressive, interactive effect through $\lambda W u$

$$E(u) = 0 \qquad E(\varepsilon) = 0$$

$$E(uu') = C \qquad E(\varepsilon\varepsilon') = \sigma^2 I$$

From this basic specification several different equivalent expressions can be derived

Substitution of the lower equation into the top equation yields:

$$u - \lambda W u = \varepsilon$$

$$(I - \lambda W)u = \varepsilon$$

$$u = (I - \lambda W)^{-1}\varepsilon$$

$$y = X\beta + (I - \lambda W)^{-1}\varepsilon$$

It turns out that…

$$(I - \lambda W)^{-1} = I + \lambda W + \lambda^2 W^2 + \lambda^3 W^3 \ldots$$

and therefore that…

An alternative (reduced form) expression for the Spatial Error Model becomes:

$$y = X\beta + [I + \lambda W + \lambda^2 W^2 + \lambda^3 W^3 + \ldots]\varepsilon$$

Alternatively, going back to the original (structural form) specification of the spatial error model, substitution of the top equation into the lower equation yields a slightly different, equivalent, specification…

Substitution of top into bottom:

$$y - X\beta = u$$

$$y - X\beta = \lambda W(y - X\beta) + \varepsilon$$

$$y = X\beta + \lambda Wy - \lambda WX\beta + \varepsilon$$

This particular substitution process leads to what is often called a “Spatial Durbin Model” (or “Common Factors Model”)

Discuss later… time permitting

Spatial Lag Model

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

Here, first-order variation comes only through $X\beta$; second-order variation is represented as an autoregressive, interactive effect through $\rho W y$

Analogous to a distributed lag in a time-series model

Let's rearrange the terms in this spatial lag model just a bit…

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

$$y - \rho Wy = X\beta + \varepsilon$$

$$(I - \rho W)y = X\beta + \varepsilon$$

$$y = (I - \rho W)^{-1} X\beta + (I - \rho W)^{-1} \varepsilon$$

and recalling that…

$$(I - \rho W)^{-1} = I + \rho W + \rho^2 W^2 + \rho^3 W^3 ...$$

We thus have this revised (reduced form) expression for the Spatial Lag Model:

$$y = \left[I + \rho W + \rho^2 W^2 + \rho^3 W^3 + ...\right] X\beta$$
$$+ \left[I + \rho W + \rho^2 W^2 + \rho^3 W^3 + ...\right] \varepsilon$$

Can you say in words what this model is telling us?

Comparing the two models (structural specification)

Spatial Error Model:

$$y = X\beta + u$$

$$u = \lambda W u + \varepsilon$$

Spatial Lag Model:

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

Comparing the two models (reduced form specification)

Spatial Error Model:

$$y = X\beta + [I + \lambda W + \lambda^2 W^2 + \lambda^3 W^3 + \ldots]\varepsilon$$

Spatial Lag Model:

$$y = [I + \rho W + \rho^2 W^2 + \rho^3 W^3 + \ldots]X\beta + [I + \rho W + \rho^2 W^2 + \rho^3 W^3 + \ldots]\varepsilon$$

Because the spatial error and spatial lag models are not nested specifications, i.e., they cannot be derived from some general specification by setting terms to zero, they are usually presented (e.g., in *GeoDa* as alternative model specifications: either/or

So how do we know which model to use?

GeoDa output from an OLS regression run looks like this

```
REGRESSION
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Data set            : south00E
Dependent Variable  :    SQRTPPOV  Number of Observations: 1387
Mean dependent var  :    0.464095  Number of Variables   :    4
S.D. dependent var  :   0.0965369  Degrees of Freedom    : 1383

R-squared           :    0.676912  F-statistic           :     965.854
Adjusted R-squared  :    0.676211  Prob(F-statistic)     :           0
Sum squared residual:     4.17624  Log likelihood        :     2058.04
Sigma-square        :  0.00301969  Akaike info criterion :    -4108.08
S.E. of regression  :   0.0549517  Schwarz criterion     :    -4087.14
Sigma-square ML     :  0.00301098
S.E of regression ML:   0.0548724
------------------------------------------------------------------------
    Variable    Coefficient     Std.Error    t-Statistic   Probability
------------------------------------------------------------------------
    CONSTANT   -0.06198439     0.01071833      -5.783029    0.0000000
    SQRTUNEM     0.8280865     0.03843687       21.54406    0.0000000
    SQRTPFHH     0.3865141     0.02413307       16.01595    0.0000000
    LOGHSPLS    -0.1416951    0.005916132      -23.95064    0.0000000
------------------------------------------------------------------------
```

For now, we want only the next page

Part of the *GeoDa* output from an OLS regression run looks like this

```
REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER   19.70728
TEST ON NORMALITY OF ERRORS
TEST                     DF          VALUE              PROB
Jarque-Bera               2           568.7069          0.0000000

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                     DF          VALUE              PROB
Breusch-Pagan test        3           101.1902          0.0000000
Koenker-Bassett test      3           41.82912          0.0000000
SPECIFICATION ROBUST TEST
TEST                     DF          VALUE              PROB
White                     9           300.2165          0.0000000

DIAGNOSTICS FOR SPATIAL DEPENDENCE
FOR WEIGHT MATRIX : socoQueen1.GAL  (row-standardized weights)
TEST                           MI/DF       VALUE           PROB
Moran's I (error)           0.450740     28.0766483       0.0000000
Lagrange Multiplier (lag)          1     492.1368675       0.0000000
Robust LM (lag)                    1      38.1573548       0.0000000
Lagrange Multiplier (error)        1     775.6052246       0.0000000
Robust LM (error)                  1     321.6257118       0.0000000
Lagrange Multiplier (SARMA)        2     813.7625794       0.0000000
```

This will be covered in more detail in this afternoon's lab

Questions for now?

Readings for today

- Anselin, Luc, and Anil Bera. 1998. "Spatial Dependence in Linear Regression Models with an Introduction to Spatial Econometrics." Chapter 7 (pp. 237-289) in Aman Ullah and David Giles (eds.) *Handbook of Applied Economic Statistics* (New York: Marcel Dekker).
- Anselin, Luc. 2002. "Under the Hood: Issues in the Specification and Interpretation of Spatial Regression Models." *Agricultural Economics* 27(3):247-267.
- Baller, Robert D., and Kelly K. Richardson. 2002. "Social Integration, Imitation, and the Geographic Patterning of Suicide." *American Sociological Review* 67(6):873-888.
- Sparks, Patrice Johnelle, & Corey S. Sparks. 2010. "An Application of Spatially Autoregressive Models to the Study of US County Mortality Rates." *Population, Space and Place* 16:465-481.
- Anselin, Luc. 2005. *Exploring Spatial Data with GeoDa: A Workbook*, (chapters 22-25).
- Anselin, Luc. 2005. *Spatial Regression Analysis in R: A Workbook,* (chapter 6).

Afternoon Lab

Spatial Regression Modeling in *GeoDa* & R

Areas of needed research:

- Spatial panel models; space-time interactions
  - Jihai Yu
  - Yanbing Zheng
- Latent continuous variables; binary dependent variable; counts; etc.
- Flow models
- Endogenous weighting matrices

Thanks for your participation!!

See you this afternoon