Introduction to Spatial Regression Analysis

Paul Voss
UNC Chapel Hill

Day 2

Review of yesterday

- Overview of spatial data and spatial data analysis
- Why “spatial is special”
  - characteristics of spatial data
    - scale dependence
    - edge effects
    - heterogeneity
    - autocorrelation
  - problems caused by spatial data
    - iid assumptions of standard linear model violated
    - need for tools to deal with non-iid error variance-covariance among other problems (boundary problems)
- Classes of problems in spatial data analysis
- Review OLS assumptions & violations
- Importance of EDA/ESDA

Outline for today

- Spatial processes
  - spatial heterogeneity
  - spatial dependence
- Global spatial autocorrelation & weights matrices
  - Moran's I
  - Geary's c
- Understanding & measuring local spatial association
  - Moran scatterplot
  - LISA statistics
- Lab: spatial autocorrelation in *GeoDa* & R

Questions?

Recall from yesterday…

"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
Sage, 2000:xii

And what are these “unique properties”?

- Spatial heterogeneity
- Spatial dependence

} Lattice data

- Spatial inhomogeneity
- Contagion

} Event data

“Spatial Effects” or
“Spatial Processes”

"Spatial Effects" ("Spatial Processes")

- *Spatial effects* are properties of spatial data resulting in the tendency for spatially proximate observations of an attribute $Z(\boldsymbol{s})$ in $\mathcal{R}$ to be more alike than more distant observations (Tobler's 1st Law)
- Such clustering in space can result from *properties shared by some areas* in the study region that make them different from other areas in the region (identifiable or not) or from some type of *spatially patterned interaction* among neighboring units or both
- Think about it as:
  - reactive processes (which we will call "spatial heterogeneity")
    - We will try to model this process with covariates, but generally we will fail
  - interactive processes (which we will call "spatial dependence")

Spatial Heterogeneity

… exists when the mean, and/or variance, and/or covariance structure of the DGP "drifts" over a mapped process

- Typified by regional differentiation; a large scale spatial process expressing itself across the entire region under study; arises from regional differences in the DGP
- Reflects the "spatial continuities" of social processes which, "taken together help bind social space into recognizable structures" – a "mosaic of homogeneous (or nearly homogeneous)" areas in which each is different from its neighbors (Haining, 1990:22)
- we study using 1st-order analytical tools & perspectives (primarily linear or non-linear regression)

Spatial Heterogeneity (2)

- No *spatial interaction* is assumed in the process generating spatial heterogeneity. Follows from the "intrinsic uniqueness of each location" (Anselin, 1996:112)
- A troublesome property, because an assumption of spatial *homogeneity* (spatial stationarity) is assumed to provide the necessary replication for drawing inferences from the process
- Moreover, spatial stationarity is an assumption underlying spatial dependence testing & modeling
- Thus, we're going to work pretty hard to model the 1st-order spatial effects
- The definition also includes drift in *covariance structure*, and this is the crucial aspect of spatial heterogeneity for many analysts

If we assume spatial *heterogeneity* is a reasonable DGP to entertain…

- We are saying that apparent clustering in the data is not a result of spatial interaction among areas; no small scale neighbor influences; no contagion; no 2nd-order spatial effects
- For areal data, we are presuming that we can specify a regression model with suitable covariates such that residual autocorrelation evaporates
- We allow that purely spatial structural effects may have to be part of our model specification

Questions about the notion of spatial heterogeneity?

Spatial Dependence

… the existence of a *functional relationship* between what happens at one point in space and what happens elsewhere (Anselin, 1988:11)

- This sounds a lot like spatial autocorrelation (not yet formally defined)… but I do not use the terms interchangeably [not all authors are this cautious]
- It means a lack of independence among observations (by definition); but "functional relationship" is the key
- Expresses itself as a small-scale, localized, short-distance spatial process; 2nd-order spatial process

Spatial Dependence (2)

- For the examination of data on an irregular spatial lattice (e.g., counties), this spatial process generally is handled through the exogenous declaration of a "neighborhood" defined for each observation (and is operationalized by a "weights matrix")
- Returning to the formal expression for our data (defined yesterday), $\boldsymbol{z}(\boldsymbol{s}) = f(\boldsymbol{X},\boldsymbol{s},\boldsymbol{\beta}) + \boldsymbol{\varepsilon}(\boldsymbol{s})$, we assume that the 1st-order part of the model leaves behind a disturbance vector $\boldsymbol{\varepsilon}(\boldsymbol{s})$ with a spatial dependence process that is stationary (and, usually, isotropic – although environmental scientists will generally disagree with this last bit)

Spatial Dependence (3)

- This process follows *informally* from the so-called "First Law of Geography"
- Follows *formally* from a spatial property known as "ergodicity", where we permit spatial interaction to occur only over a very limited region
- We assume the process is ergodic in order to limit the number of parameters we estimate
- Recall from yesterday:

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

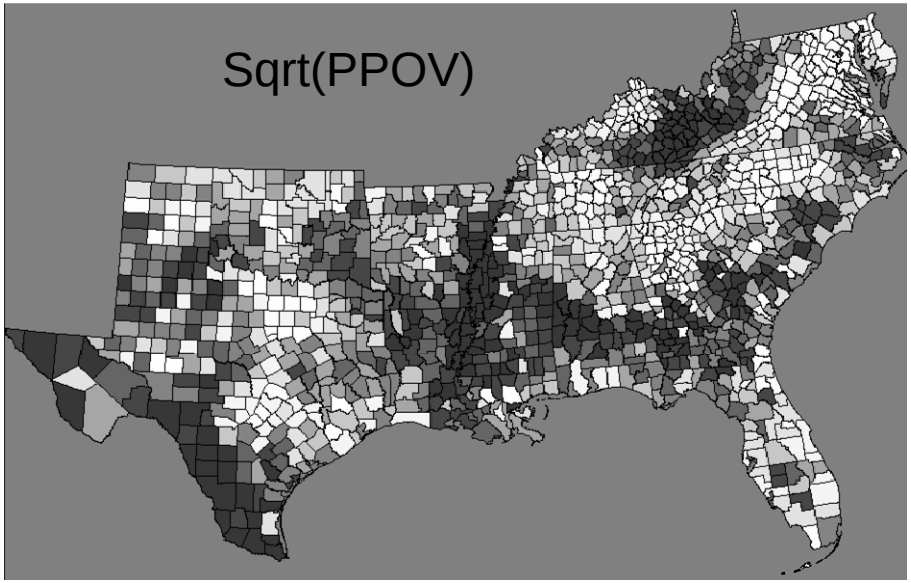
where $\boldsymbol{u}(\boldsymbol{s})$ is a random vector with mean $\mathbf{0}$ and variance $\mathrm{Var}[\boldsymbol{u}(\boldsymbol{s})] = \boldsymbol{\Sigma}(\boldsymbol{\theta})$

number of parameters = $((\mathrm{n} \times \mathrm{n}) - \mathrm{n})/2$ covariances
$+ \mathrm{n}$ variances
$+ \mathrm{k}+1$ parameters

If we assume spatial *dependence* is a reasonable DGP…

- We are saying that clustering in our data results from some type of spatial interaction; existence of small-scale neighbor influences
- We believe we can theoretically posit reasons why clustering results from spatial interaction ("contagion") among our units of observation
- We need to be thinking about what kind of dependence model should best introduce neighboring effects
- Time to dig into the 2nd-order spatial analysis toolkit

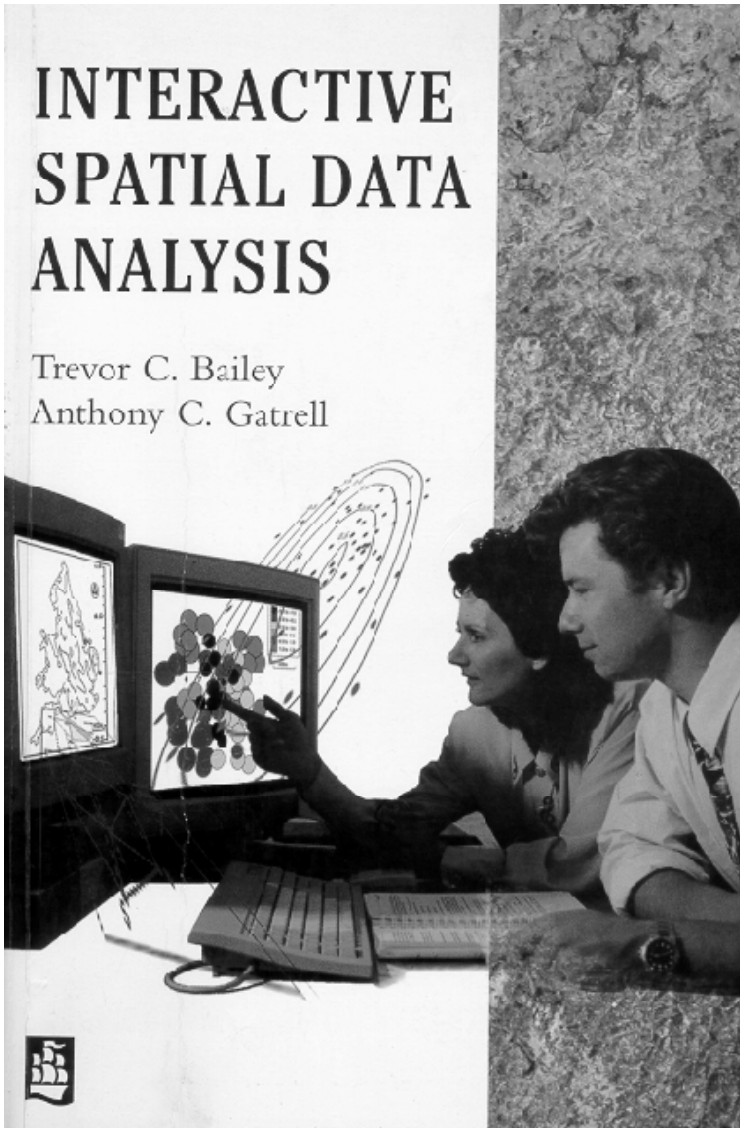
Questions about the notion of spatial dependence?

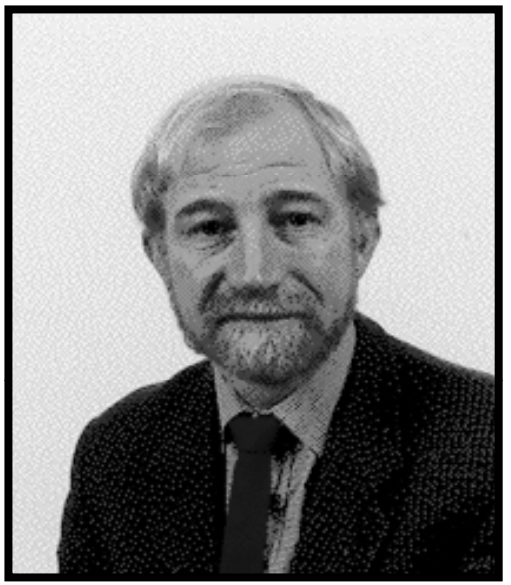


So, which is it… spatial dependence or spatial heterogeneity? (Frankly, it's not even a proper question! Why?)

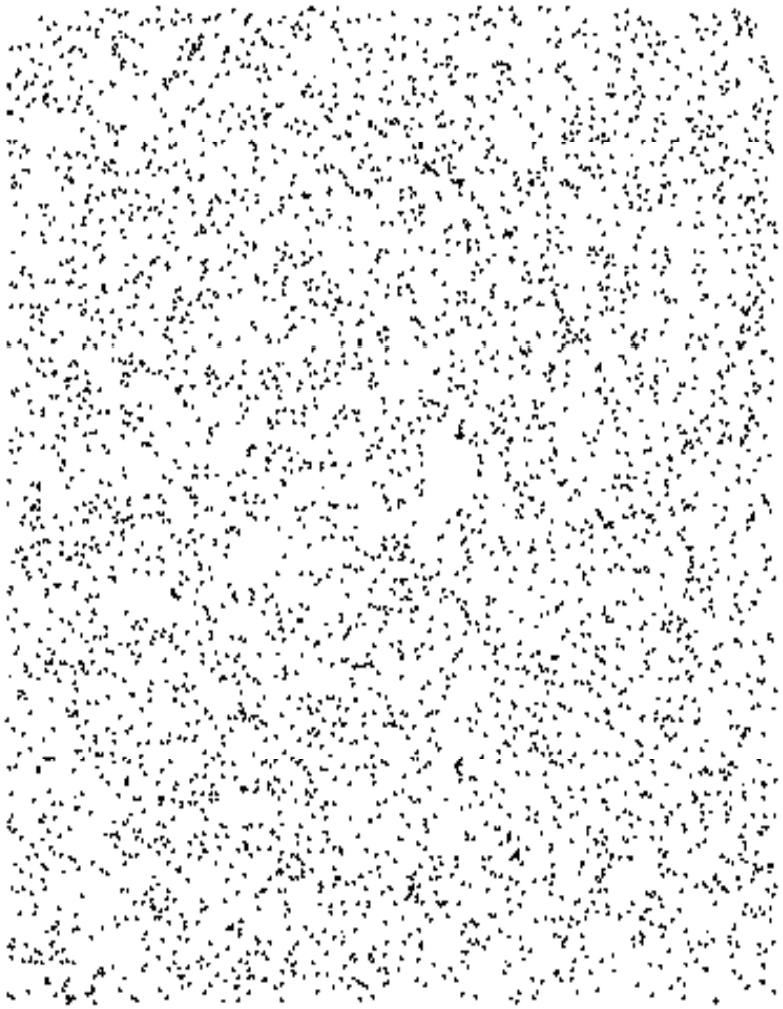
Better questions might be: "What do we have?" "What would we like to know?" "What questions can these data answer?" "What spatial tools do we need to turn to?" "How should we think about modeling these data given what we want to learn?"

Recall from yesterday…

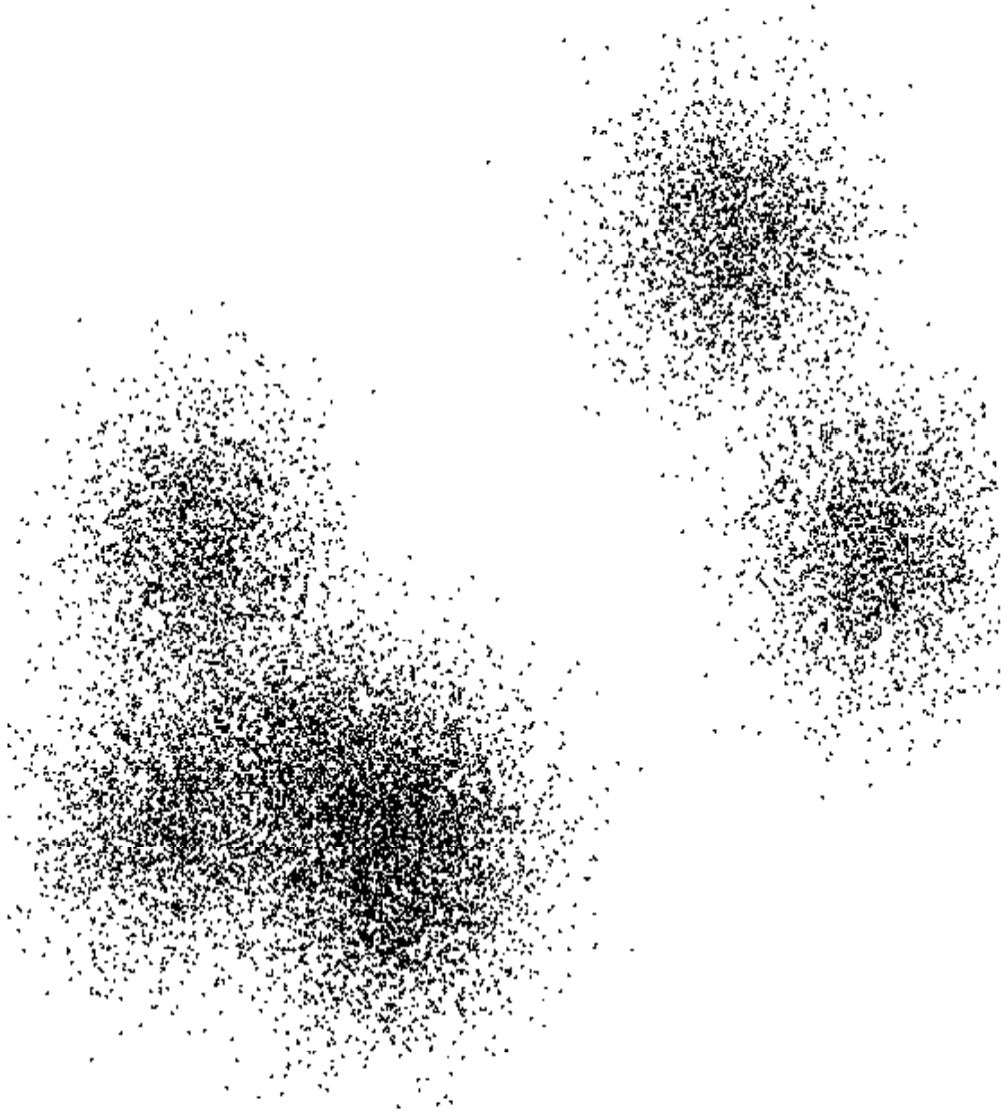




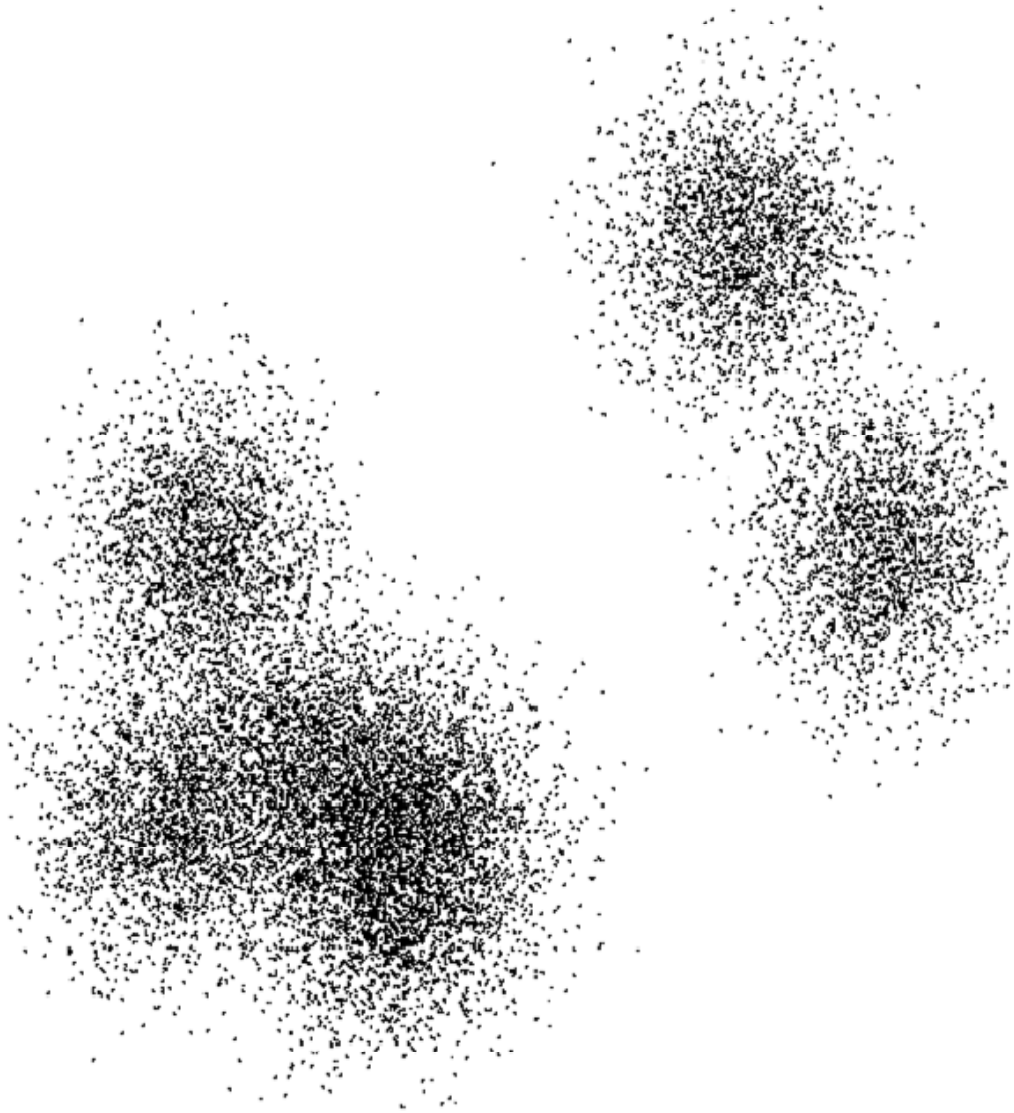
4,000 metal filings scattered on paper



4,000 metal filings, hidden magnets underneath

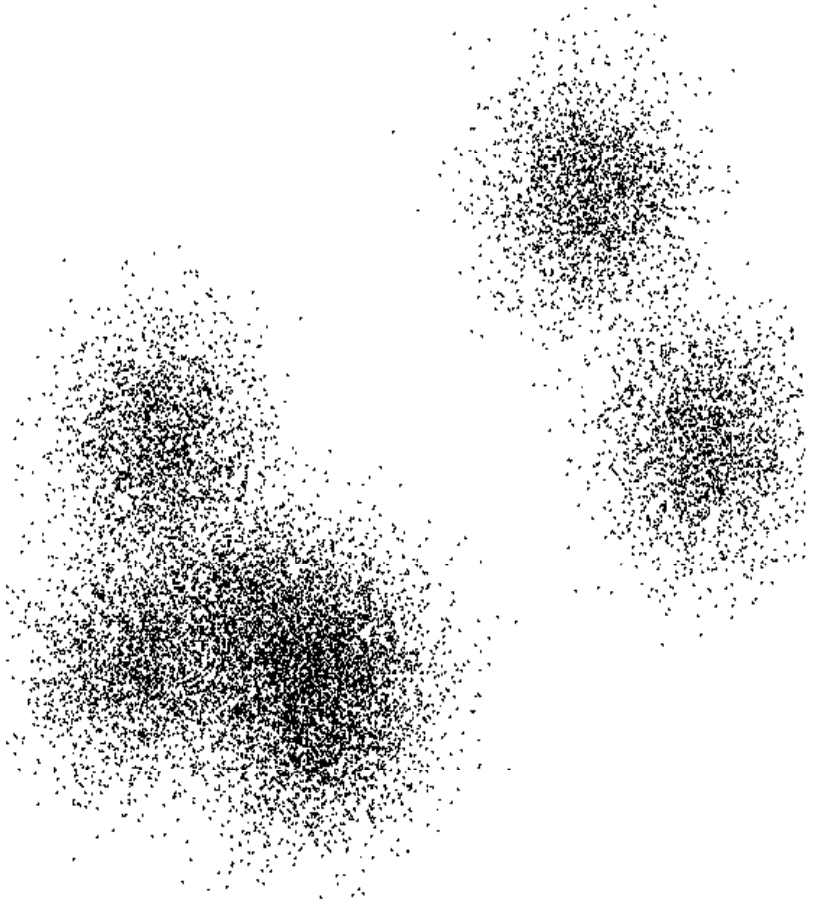


4,000 magentized metal filings on paper

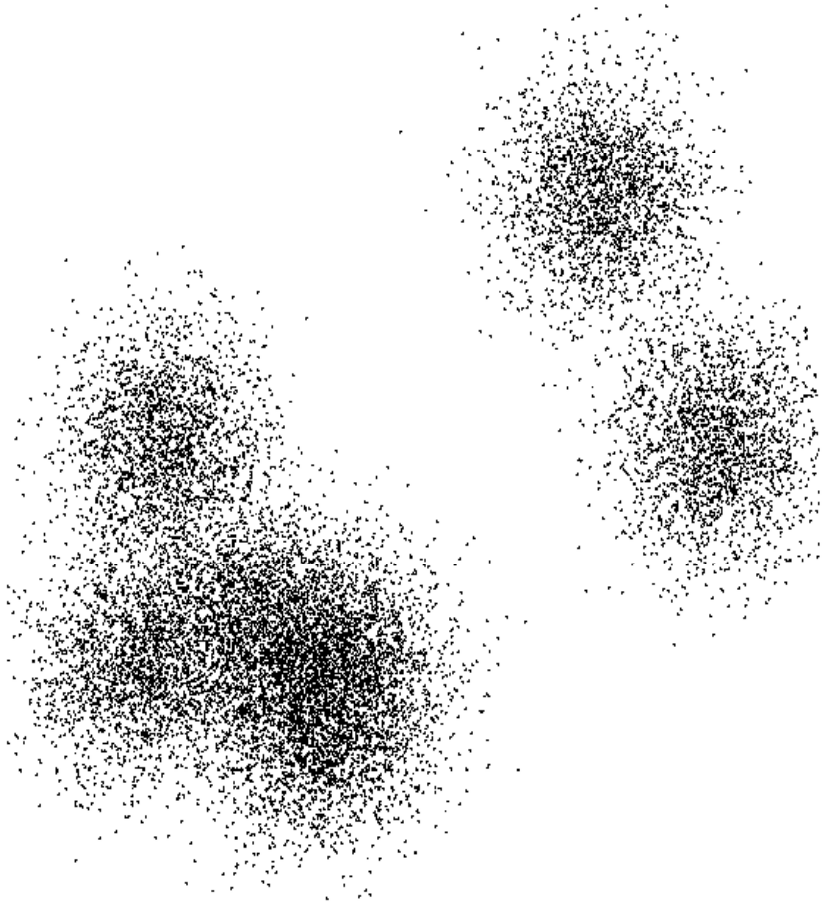


Which is which? How to proceed?

4,000 metal filings, hidden magnets underneath



4,000 magentized metal filings on paper



“reactive” process

“interactive” process

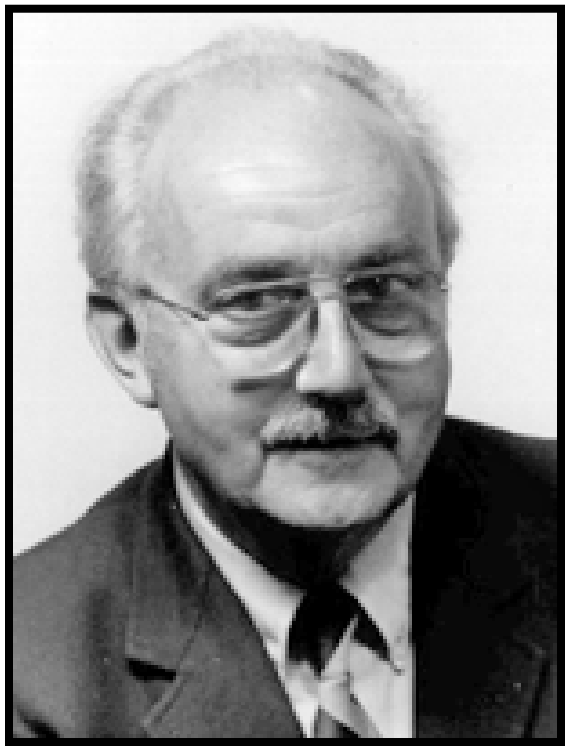
You almost never know

So… how *should* we proceed?

A useful admonition to keep in mind as we get started:

“All models are wrong, some models are useful”

G.E.P.Box
“Robustness in the Strategy of Scientific Model Building”
pp. 201-236 in Lanner and Wilkerson (eds.)
Robustness in Statistics
Academic Press, 1979



Questions?

Global Spatial Autocorrelation

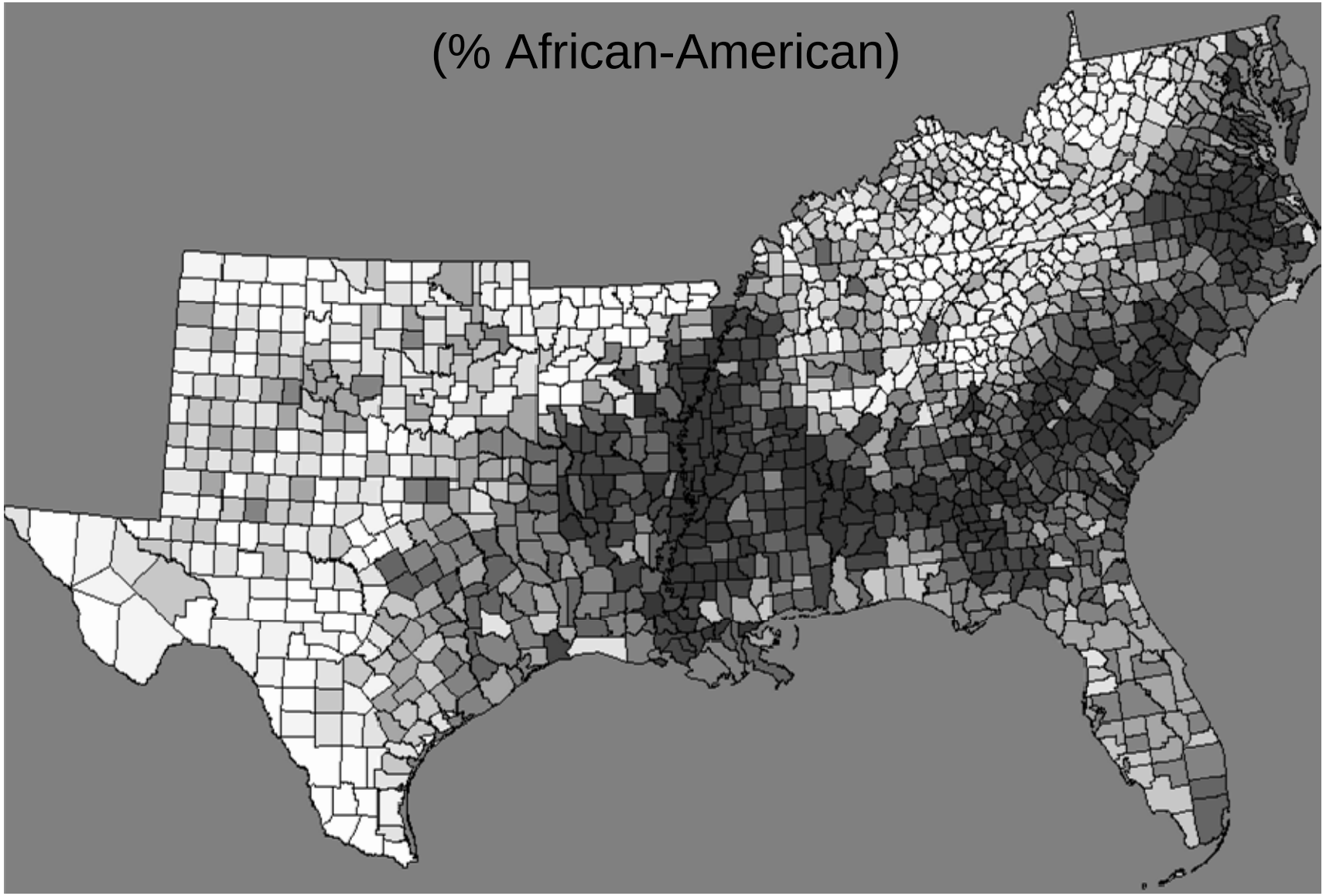
Recall from yesterday…

- When correlated errors arise from a specification with missing variables, OLS estimates of t-test values are unreliable
  - The OLS estimates are not efficient
  - Under positive spatial autocorrelation, the std. errors of the parameter estimates are biased downward
  - Informally, you can think of this as arising because the OLS model "thinks" it's getting more information from the observations than it is
  - Correlated errors inflate the value of the R^2 statistic
- When correlated errors result from endogeneity, OLS regression parameter estimates are biased and inconsistent

So, where do we go from here?

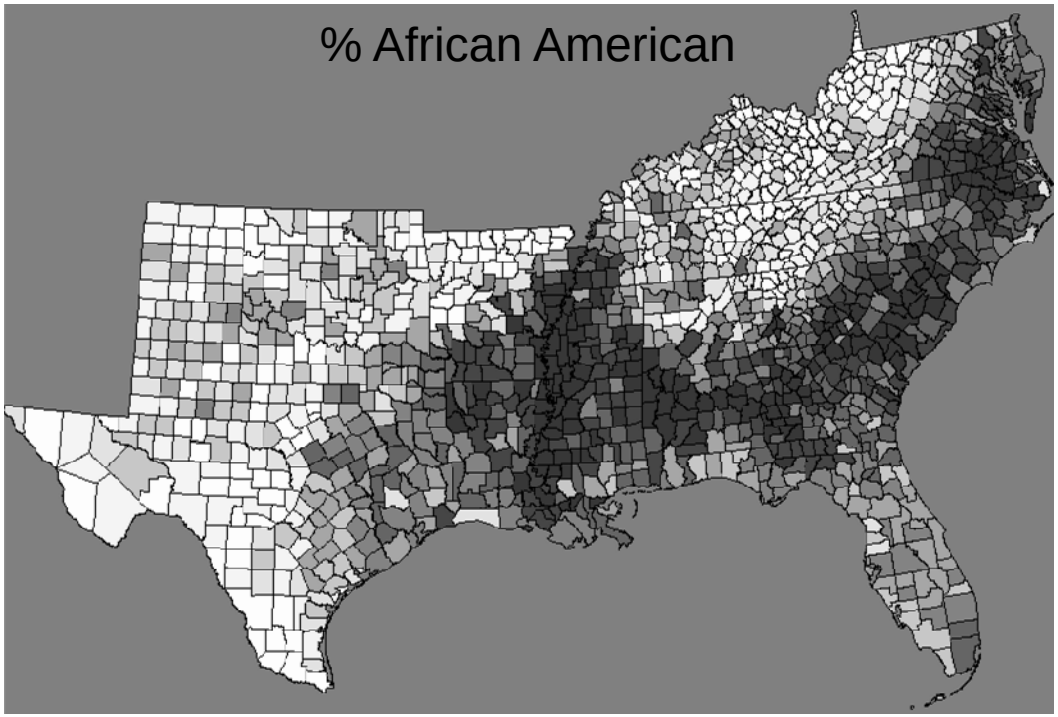
- Specifically, how do we develop a means to (statistically) differentiate among different kinds of maps?
- That is, can we *quantify* different kinds of map patterns?
- And once we develop a statistic for describing (quantifying) different kinds of map patterns, can we derive the sampling distribution for this statistic and thus make inferential claims about one map vs. another map?

Suppose we observe the following map of southern counties from the 2000 Census



The question for us then is this:

If African-Americans had somehow been allocated in a random fashion to the southern counties, would this observed spatial distribution be a likely outcome of such an allocation procedure?



But what does this question mean? Does it even make sense?

What does it mean to ask about the spatial distribution of a census variable as if the observations are an outcome from some type of sampling experiment?

- The data are… well, the data; right?
- We've got all the counties, not just a sample of them
- The data for each county (% African-American) are based on complete count census data, not on a sample.
- So what can it possibly mean to ask whether this percent has been allocated in a "random fashion" (or not)?
- We're looking at the complete "universe of observations." It's not a sample. Or is it??
- Sometimes asked: "Is it a sample of 1,387 (counties)?" Or is it, rather, a sample of 1 (single realization of a stochastic process)?

Answers to these questions point toward the concept of a data generating process (DGP)

This is the *conceptual* notion that our data actually represent just one realization of a very large number of possible outcomes

There are a number of formal perspectives on this topic

and a terrific quote…

"[Our data often render] the idea that one is working with a (spatial) sample somewhat remote. Great imagination has gone into turning what appears to be a population into a sample, thereby making statistical theory relevant…"

Graham J. G. Upton &
Bernard Fingleton
Spatial Data Analysis by Example, Vol. I
(Wiley & Sons, 1985:325)

Sampling Perspectives

Generally there are four spatial sampling perspectives discussed in the literature based on the sampling design:

with replacement?	Yes	No
order Important?	Yes	No

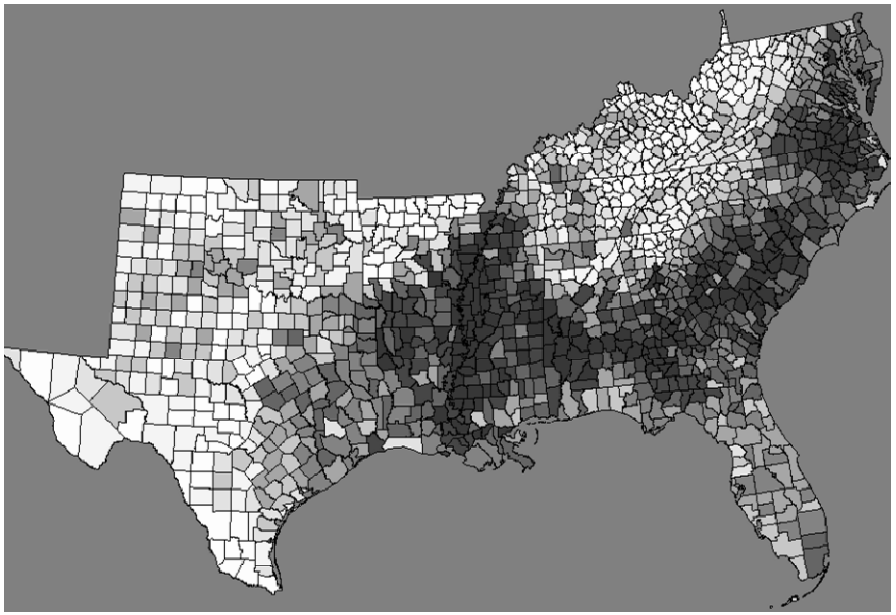
Two common of these sampling perspectives

Sampling with replacement, order is important ("free sampling" or "normalzation"): Perspective 1

Sampling without replacement, order is important ("nonfree sampling" or "randomization"): Perspective 2

Example using the southern counties

This was our "observation"



But here's just one other under nonfree sampling

The question becomes: How unusual is the *pattern* (the Moran statistic) in map 1 given the 1,387! possible permutations of these results under an assumption of nonfree sampling?

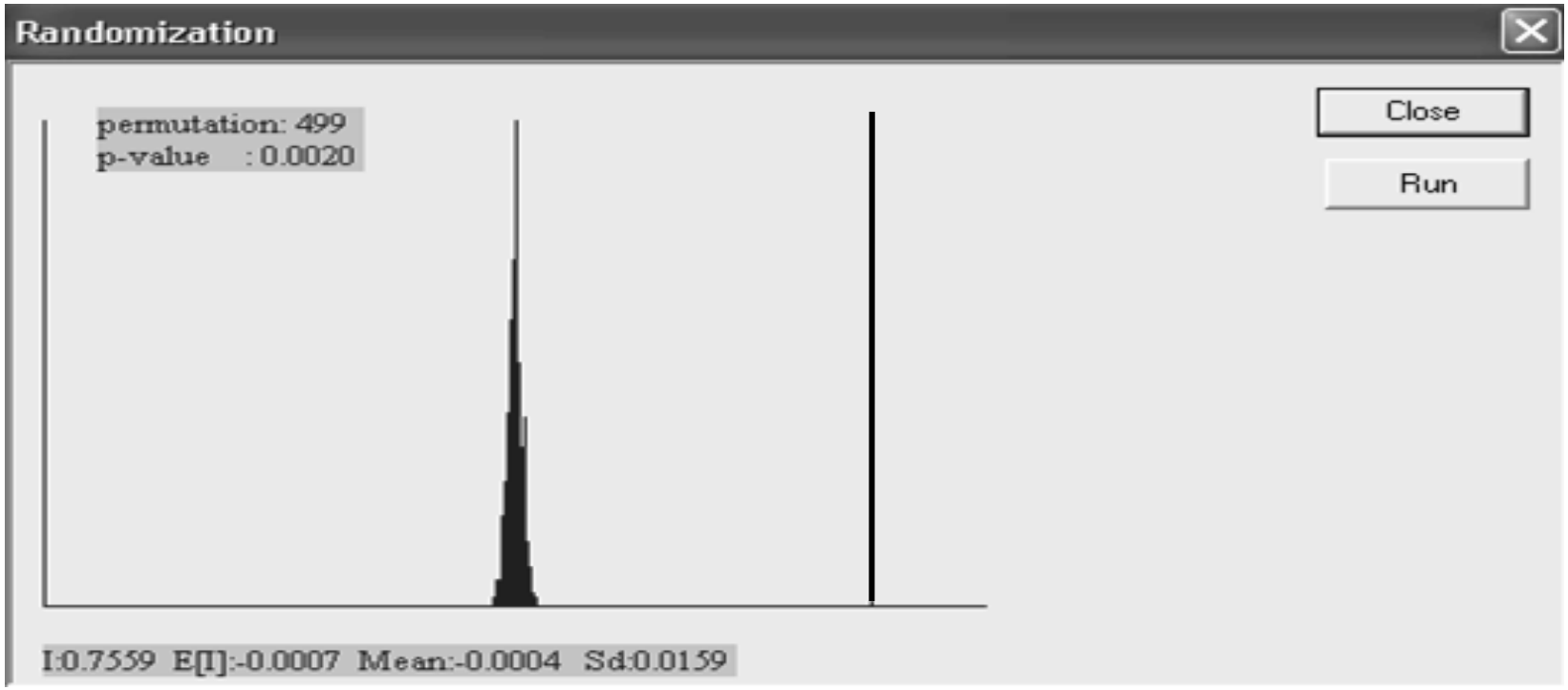
If you subscribe to the randomization approach…

To operationalize this, all (or many of) the different arrangements that are possible need to be identified in order to construct the sampling distribution for our statistic

(When the sampling distribution is simply too large to construct, we can estimate it. Such an approximate sampling distribution is sometimes called a "reference distribution")

Thus…

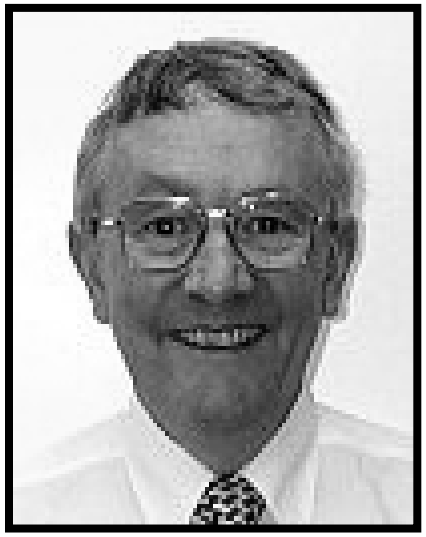
If we create 1,387! maps (or a large sample from this huge number) and derive our spatial autocorrelation statistic for each of these maps, we then have a reference distribution against which to weigh the one we actually observed. It might look something like this:

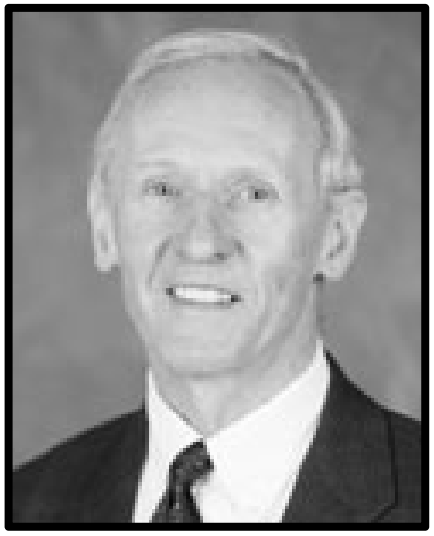


Why should I care about this?

By the way…

Australian, Pat Moran, published a version of what was to become known as the Moran test for spatial clustering in 1948





Andrew Cliff and J. Keith Ord generalized the Moran statistic to test for spatial autocorrelation among residuals from a linear regression model (under iid normal assumptions) and worked out both the large sample distribution and small sample moments in the 1970s

Spatial Autocorrelation

(Positive) spatial autocorrelation is the coexistence of attribute value similarity and locational similarity

It's the common, every day, confirmation of Tobler's first law

Formally expressed as a moment condition:

$$Cov[y_i, y_j] = E[y_i y_j] - E[y_i]E[y_j] \neq 0 \quad for\ i \neq j$$

Measuring Spatial Autocorrelation

Two classes of tests for spatial autocorrelation:

- Global spatial autocorrelation measures
  - do the data *as a whole* exhibit a spatial pattern, or are observations spatially random?
  - most common measure: Moran's statistic
- Local indicators of spatial association (LISA) statistics
  - identifies which units are significantly spatially autocorrelated with neighboring units
  - Identifies clustering ("hot spots," "cold spots")
  - localized Moran statistic

Global Moran's I

$$I = \left(\frac{n}{\left(\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \right)} \right) \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

covariance term

normalization term to scale I to the overall variance in the dataset

Moran's I coefficient as a measure of spatial autocorrelation

Pearson product-moment correlation

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

feasible range: -1 to +1

Moran's I coefficient

$$I_x = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{j=1}^{n}(x_j - \bar{x})^2}}$$

feasible range: -1 to +1 (sort of)

But… the calculation of the global Moran's I (or similar measures)requires the definition of a weights matrix

$$I = \left(\frac{n}{\left(\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \right)} \right) \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} (y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

And, as an aside, if you are interested in the intellectual history & background of where today's measures of autocorrelation originate, see special issue of…

Geographical Analysis (October, 2009) Vol. 41, Issue 4

We need to know something about weights matrices before we can proceed

Okay… again, the derivation of the global Moran's I statistic (and similar statistics) requires the specification of a weights matrix

$$I = \left(\frac{n}{\left(\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \right)} \right) \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} (y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Let's figure out what these w_{ij} elements are…

So, consider this “map”

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16

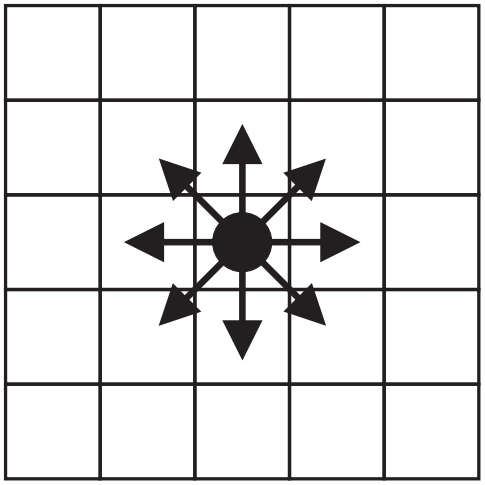
Let's say we're interested in area i = 6

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16

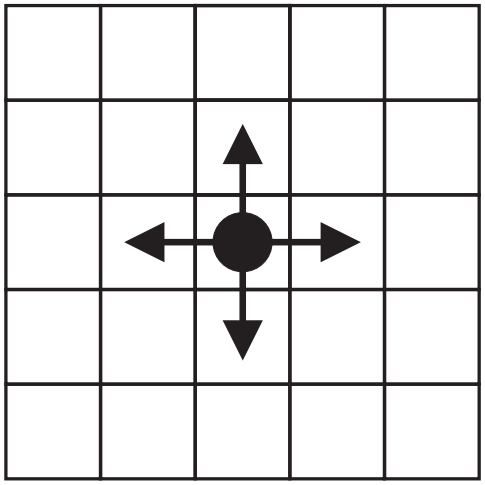
Which areas are "neighbors" of area 6?

Queens and Rooks (and—occasionally—Bishops)

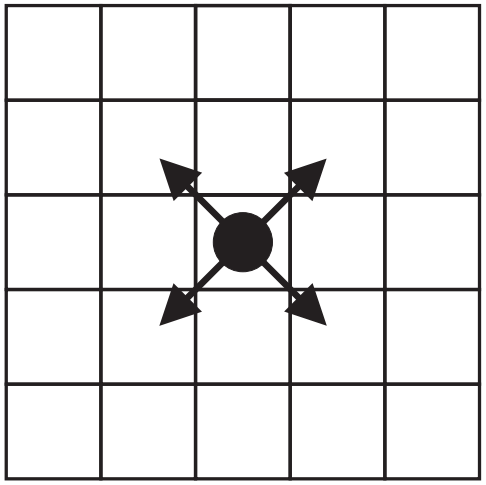
These terms are self explanatory, referring to which types of adjacent cells we choose to include as “neighbors”



Queen



Rook

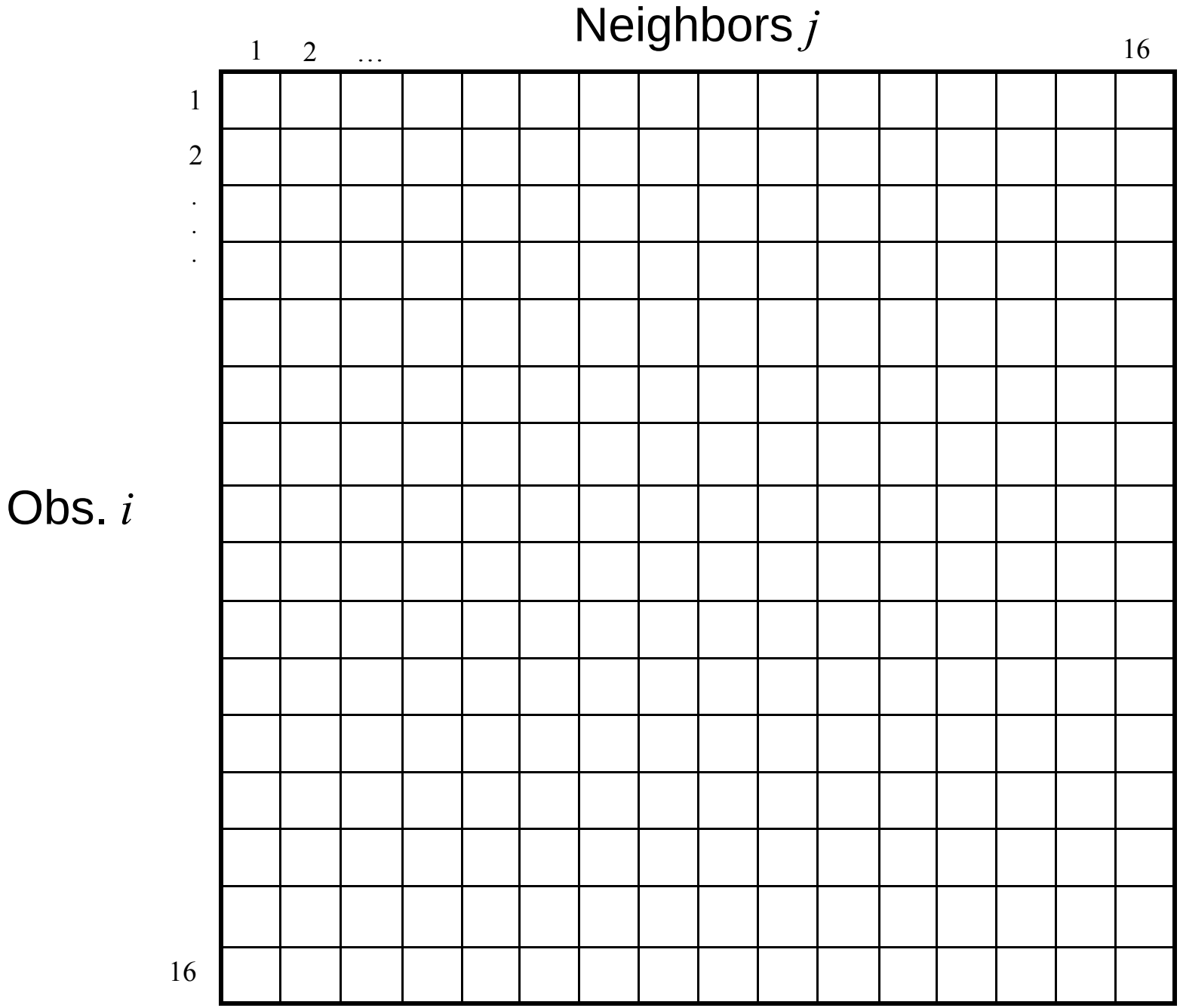


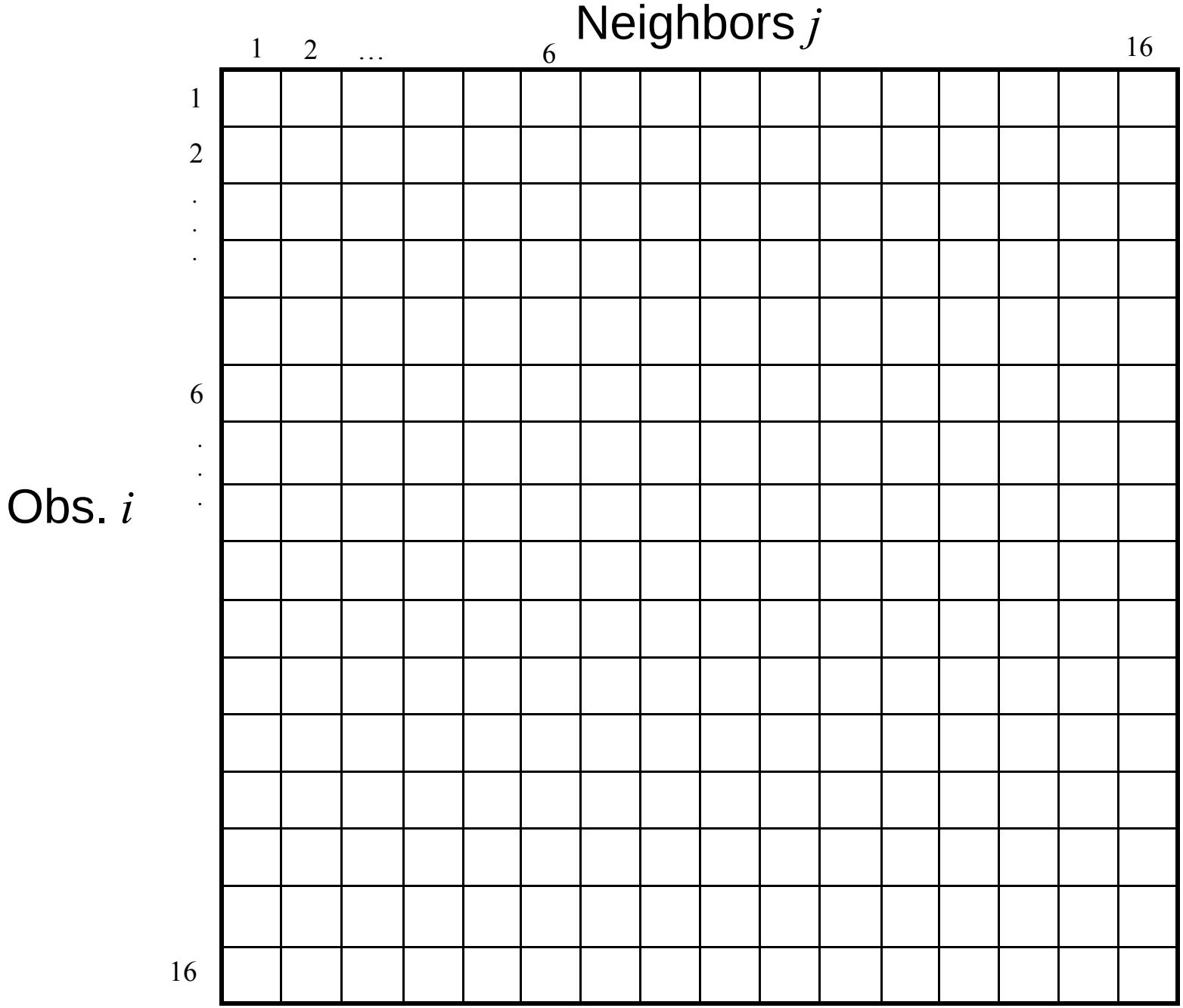
Bishop

Under a (1st order) “queen” criterion

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16

Let’s shift our thinking from the map to a matrix

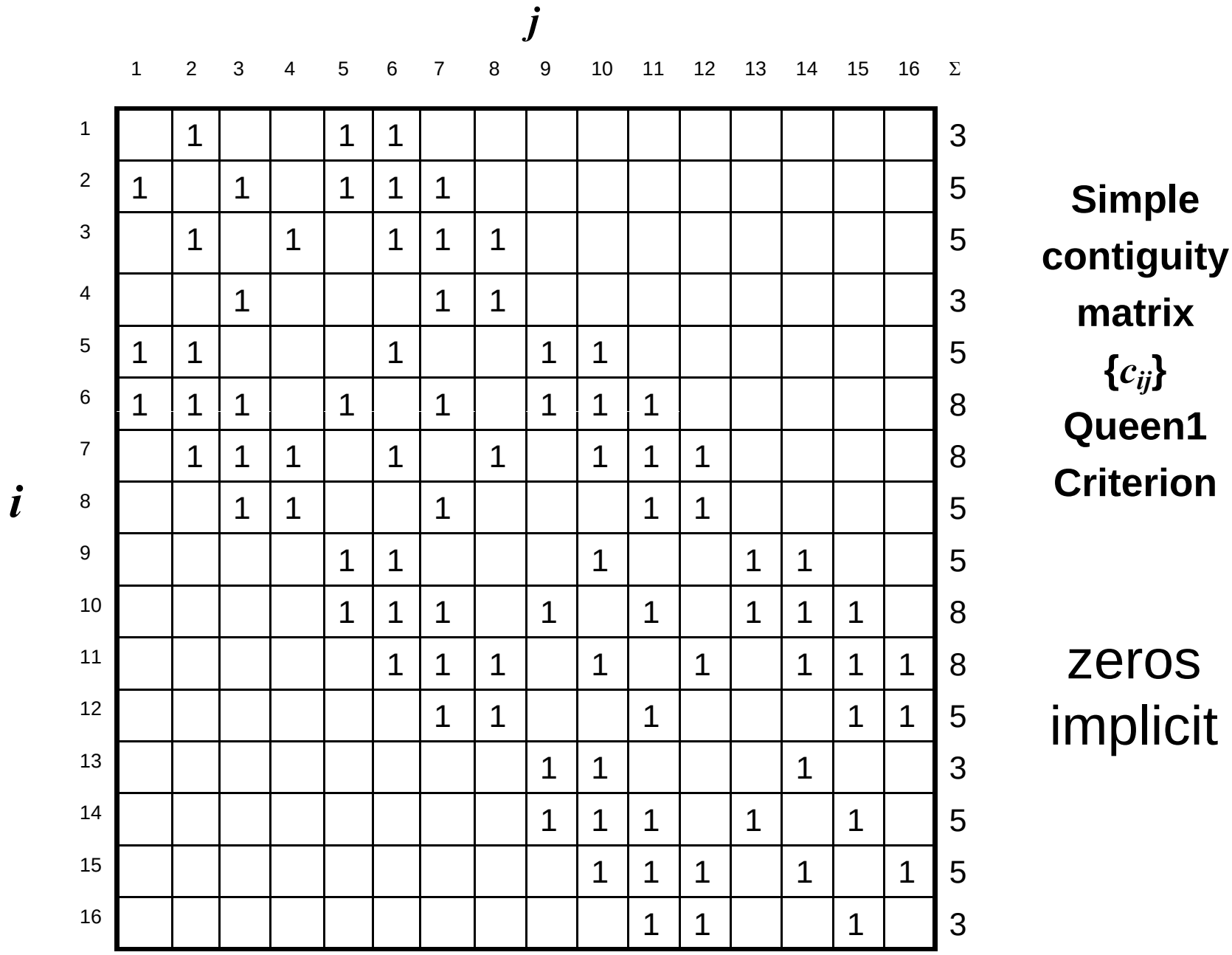

Neighbors j
1
2
…
16
Obs. i
1
2
16


Neighbors j
1
2
…
6
16
Obs. i
1
2
6
16

Now let's be a little more precise (the literature is not in agreement on these matters)

- "Contiguity matrix" – a general term that identifies neighbors with 1 and non-neighbors with 0
- "Weights matrix" – We will almost always reserve this term to refer to a row-standardized contiguity matrix, with weights:

  $0 \leq w_{ij} \leq 1$
- There a many varieties of weights matrices



Simple contiguity matrix $\{c_{ij}\}$ Queen1 Criterion

zeros implicit

i \ j	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Σ
1		1/3			1/3	1/3											1
2	1/5		1/5		1/5	1/5	1/5										1
3		1/5		1/5		1/5	1/5	1/5									1
4			1/3				1/3	1/3									1
5	1/5	1/5				1/5			1/5	1/5							1
6	1/8	1/8	1/8		1/8		1/8		1/8	1/8	1/8						1
7		1/8	1/8	1/8		1/8		1/8		1/8	1/8	1/8					1
8			1/5	1/5			1/5				1/5	1/5					1
9					1/5	1/5				1/5			1/5	1/5			1
10					1/8	1/8	1/8		1/8		1/8		1/8	1/8	1/8		1
11						1/8	1/8	1/8		1/8		1/8		1/8	1/8	1/8	1
12							1/5	1/5			1/5				1/5	1/5	1
13									1/3	1/3				1/3			1
14									1/5	1/5	1/5		1/5		1/5		1
15										1/5	1/5	1/5		1/5		1/5	1
16											1/3	1/3			1/3		1

Common row standardized weights matrix

$$w_{ij} = \frac{c_{ij}}{\sum_j c_{ij}}$$

zeros implicit

So, the elements of the weights matrix serve somewhat the role of an indicator variable in this equation. Nearby observations have non-zero weights; distant observations have zero weight

$$I = \left(\frac{n}{\left(\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \right)} \right) \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} (y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Now consider these y_i values, i = 1,…,16

1 7	2 6	3 4	4 5
5 4	6 5	7 4	8 4
9 5	10 6	11 3	12 4
13 3	14 4	15 1	16 2

Armed with this “map”, let’s now define what we mean by a “spatial lag”

For $i = 6$, the *spatial lag operator* $\boldsymbol{w}_{6j}\boldsymbol{y}_j$ is given by:

$$\boldsymbol{w}_{6j}\boldsymbol{y}_j = \sum_{j=1}^{j=16} w_{6j}y_j$$

$$= \frac{1}{8}7 + \frac{1}{8}6 + \frac{1}{8}4 + \frac{1}{8}4 + \frac{1}{8}4 + \frac{1}{8}5 + \frac{1}{8}6 + \frac{1}{8}3$$

$= 4.9$ (zeros not shown)

In words, how would we describe this value of 4.9?

Can you define the spatial lag for map area 16?

1 7	2 6	3 4	4 5
5 4	6 5	7 4	8 4
9 5	10 6	11 3	12 4
13 3	14 4	15 1	16 2

How many neighbors under a Q1 definition?

What are the weights here?

Recall the weights matrix.

i \ j	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Σ
1		1/3			1/3	1/3											1
2	1/5		1/5		1/5	1/5	1/5										1
3		1/5		1/5		1/5	1/5	1/5									1
4			1/3				1/3	1/3									1
5	1/5	1/5				1/5			1/5	1/5							1
6	1/8	1/8	1/8		1/8		1/8		1/8	1/8	1/8						1
7		1/8	1/8	1/8		1/8		1/8		1/8	1/8	1/8					1
8			1/5	1/5			1/5				1/5	1/5					1
9					1/5	1/5				1/5			1/5	1/5			1
10					1/8	1/8	1/8		1/8		1/8		1/8	1/8	1/8		1
11						1/8	1/8	1/8		1/8		1/8		1/8	1/8	1/8	1
12							1/5	1/5			1/5				1/5	1/5	1
13									1/3	1/3				1/3			1
14									1/5	1/5	1/5		1/5		1/5		1
15										1/5	1/5	1/5		1/5		1/5	1
16											1/3	1/3			1/3		1

Common row standardized weights matrix

$$w_{ij} = \frac{c_{ij}}{\sum_j c_{ij}}$$

zeros implicit

Can you define the spatial lag for map area 16?

1 7	2 6	3 4	4 5
5 4	6 5	7 4	8 4
9 5	10 6	11 3	12 4
13 3	14 4	15 1	16 2

How many neighbors under a Q1 definition?

What are the weights here?

Recall the weights matrix.

What's the spatial lag for area 16?

So, for $i = 16$, the *spatial lag operator* $\boldsymbol{w}_{16j}\boldsymbol{y}_j$ is given by:

$$w_{16\,j} = \sum_{j=1}^{j=16} w_{16\,j}\, y_j$$

$$= \frac{1}{3}3 + \frac{1}{3}4 + \frac{1}{3}1 \quad = \quad 2.7$$

(again, zeros not shown)

And how do we describe the value of 2.7?

The previous several slides have generated the w_{ij} elements based on adjacency.

There are lots of other options; e.g.,…

- Distance and inverse distance
- k nearest neighbors (knn)
- Cliff-Ord weights
- Weights matrix should be driven as much as possible by theory
- *GeoDa* allows us to create a knn weights matrix, but has difficulty using it
- Lots of options in R
- Can edit $\boldsymbol{W}$ in a general text editor

Feasible Range of Moran's I

- Function of n
- Function of the particular weights matrix used
- Function of the structure of the tesselation
- Minimum/maximum *theoretical* values generally just above $|\pm 1|$
- As a practical matter, the minimum *empirical* value for an irregular lattice is generally around -0.6

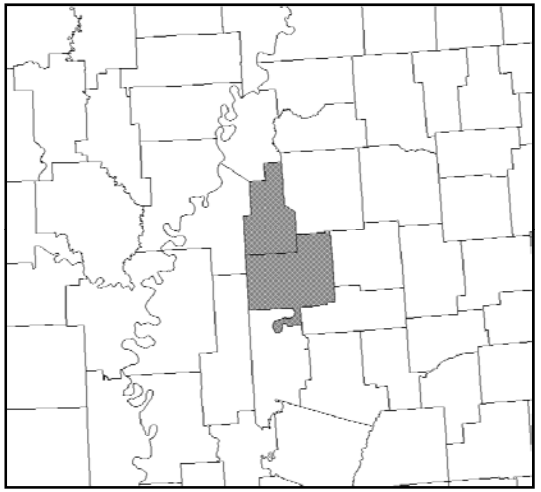
In general, the spatial lag is expressed (in matrix notation) as:

$$\boldsymbol{Wy} = \sum_{i=1}^{i=16} \sum_{j=1}^{j=16} w_{ij} y_j$$

where $\boldsymbol{W}$ is a (16 x 16) weights matrix and

$\boldsymbol{y}$ is a (16 x 1) column vector

Some interesting questions that might be addressed using spatial lag operator:



- Local tax rates

  ("spillover" in y?)
- Expenditures for police

  ("spillover" in x?)
- Demographic analysis: Are Quitman & Tallahatchie counties (two contiguous counties in the Mississippi Delta) really two separate observations?

  ("spillover" in ε ?)

We can simplify the expression for Moran's I using matrix algebra

$$I = \left(\frac{n}{\left(\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\right)} \right) \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Assume $\boldsymbol{W}$ row standardized and $z_i = y_i - \bar{y}$

$$I = \frac{z'Wz}{z'z}$$

Expected Value of Moran's I Under Hypothesis of No Spatial Autocorrelation

$$E(I) = -\frac{1}{n-1}$$

Variance of Moran's I Under Hypothesis of No Spatial Autocorrelation

- Theoretical variance: very messy
- Cliff & Ord (1973; 1991) derived the theoretical asymptotic moments of I (under two different assumptions regarding the DGP)
- Boots & Tiefelsdorf (1995) have derived the exact (small sample) moments of I, but, again, it's messy
- Anselin & Bera (1998:267) give the first two moments of I for OLS errors
- Again… *GeoDa* derives an empirical standard deviation using a permutation approach

If n is large…

$$Z = \frac{I - E(I)}{\sqrt{Var(I)}}$$

Global Geary's c

$$c = \left(\frac{n-1}{2(\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij})} \right) \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - y_j)^2}{\sum_{i=1}^{n}(y_i -$$

Can we deconstruct this?

Robert (Roy) Charles Geary
(1896-1983)

Expected Value of Geary's c Under Hypothesis of No Spatial Autocorrelation

$$E(c) = 1$$

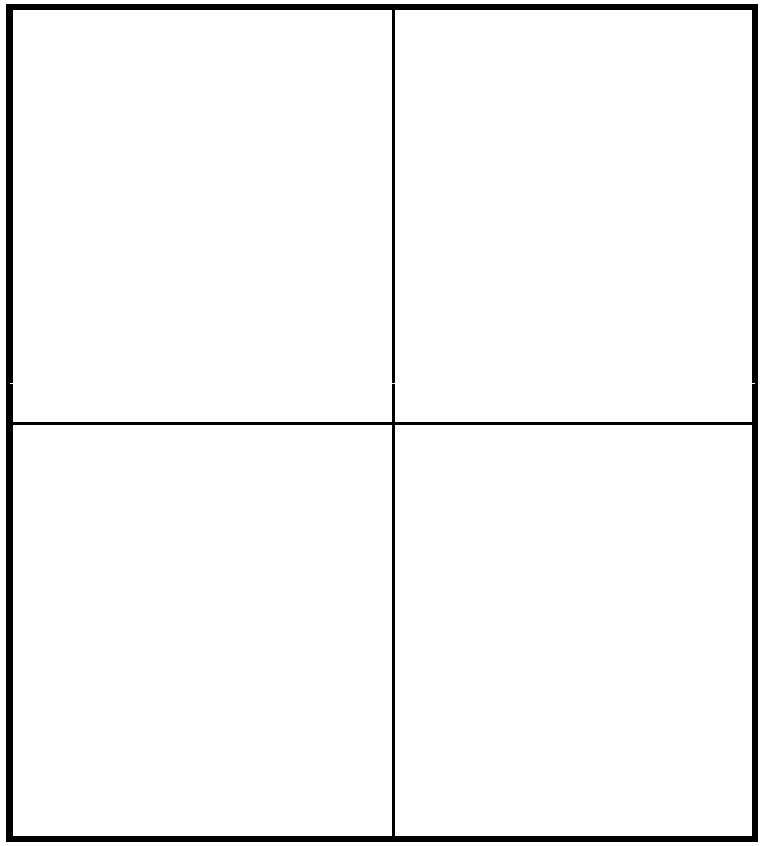
As with Moran's I, the variance of Geary's c under hypothesis of no spatial autocorrelation is messy

- But, Cliff & Ord (1973; 1981) derived the theoretical asymptotic moments of c
- *GeoDa* doesn't provide access to this test statistic
- As with Moran's I, under the null hypothesis of no spatial autocorrelation, c is asymptotically ~ N(0,1)

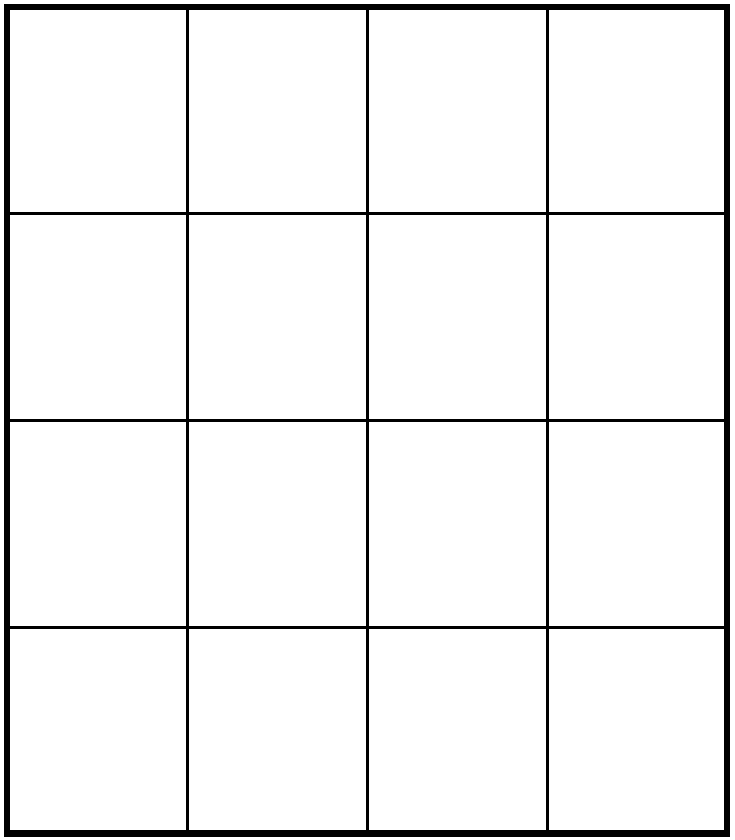
Thus, if n is large…

$$Z = \frac{c - E(c)}{\sqrt{Var(c)}}$$

Measures of spatial autocorrelation are scale dependent



Moran's I = -1.00

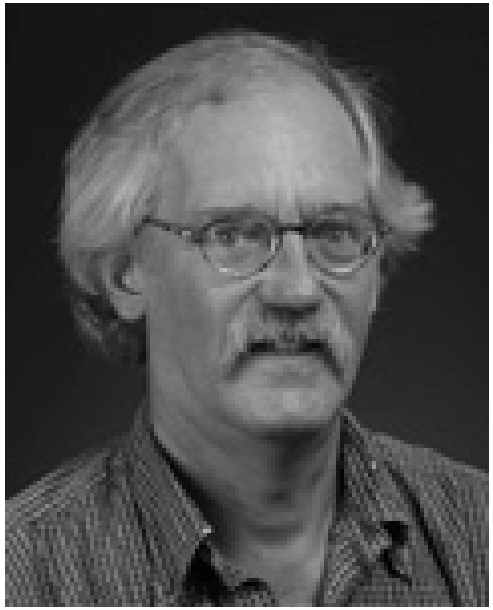


Moran's I = +0.33

LISA Statistics

Standard citation:

Anselin, Luc. 1995. "Local Indicators of Spatial Association – LISA." *Geographical Analysis* 27:93-115.



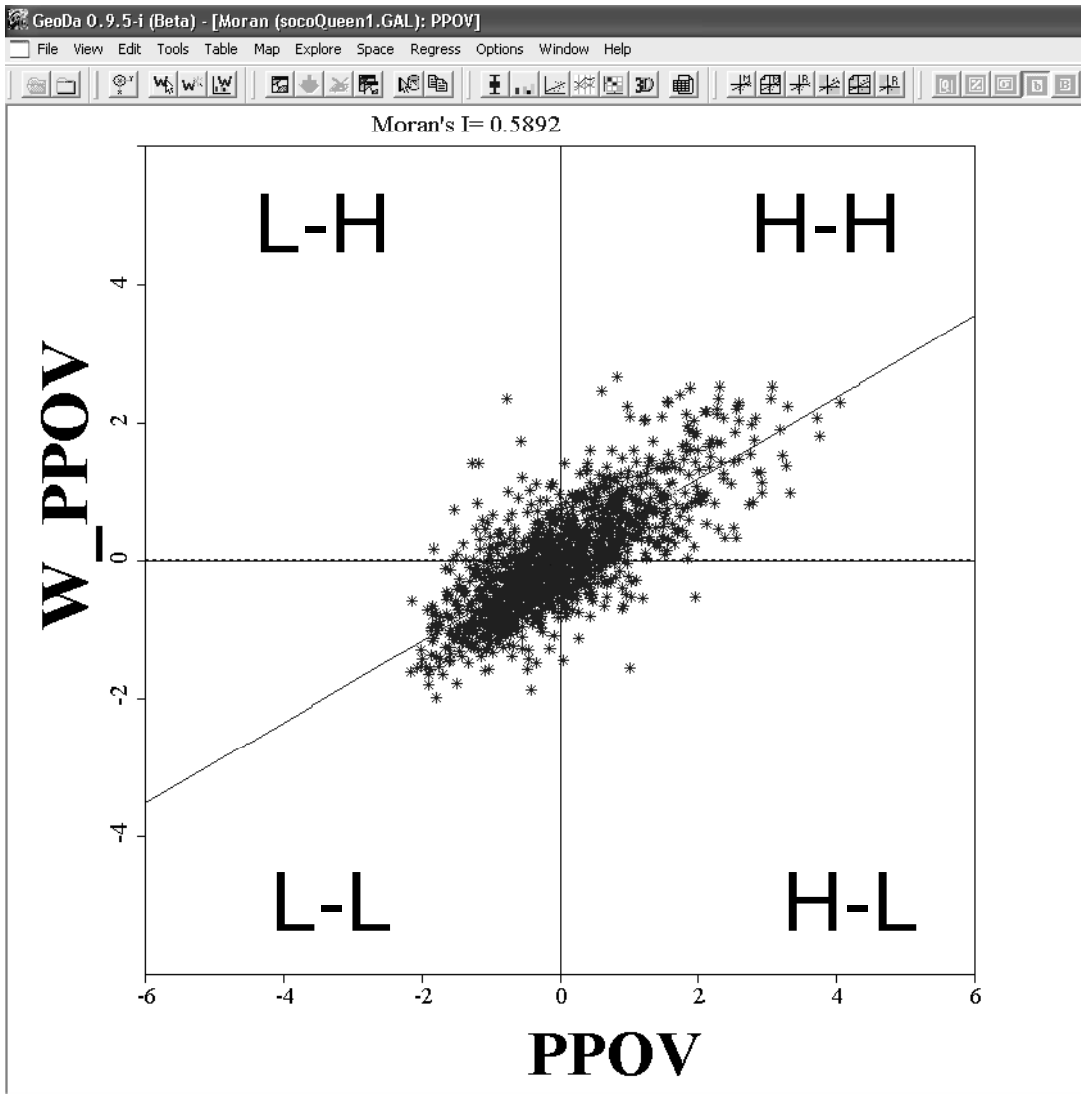
Anselin's Moran Scatterplot

Standard citation:

Anselin, Luc. 1996. "The Moran Scatterplot as an ESDA Tool to Assess Local Instability in Spatial Association." Pp. 111-125 in Fischer, Manfred, Henk J. Scholten, and David Unwin (eds.) *Spatial Analytical Perspectives on GIS: GISDATA 4* (London: Taylor & Francis).

Terrific ESDA tool

Moran Scatterplot of PPOV (1st Order Queen Weights)

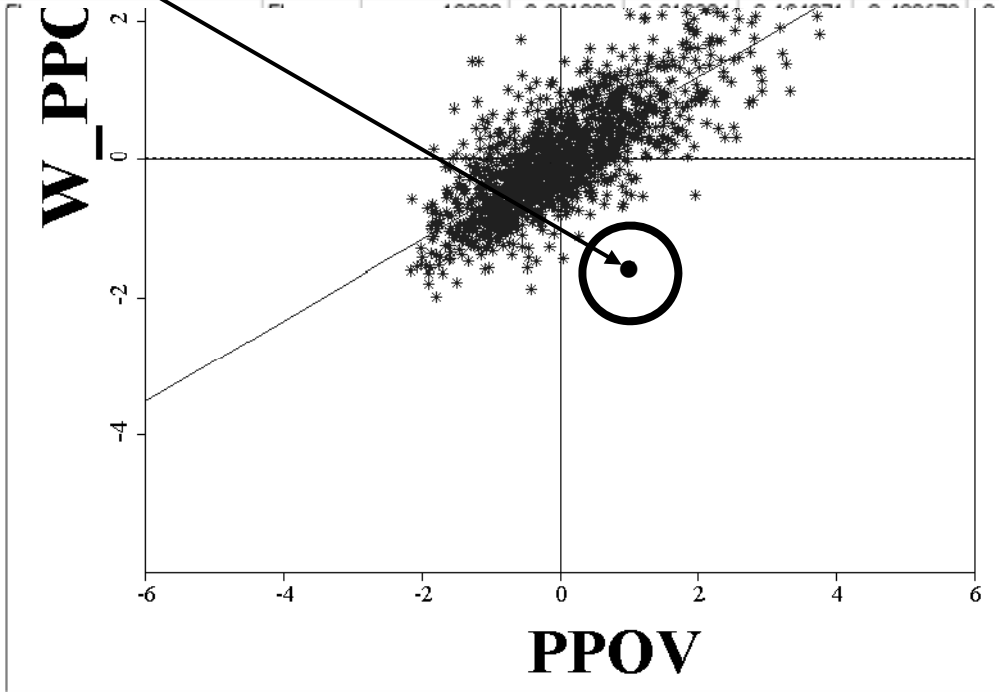


Moran Scatterplot of PPOV

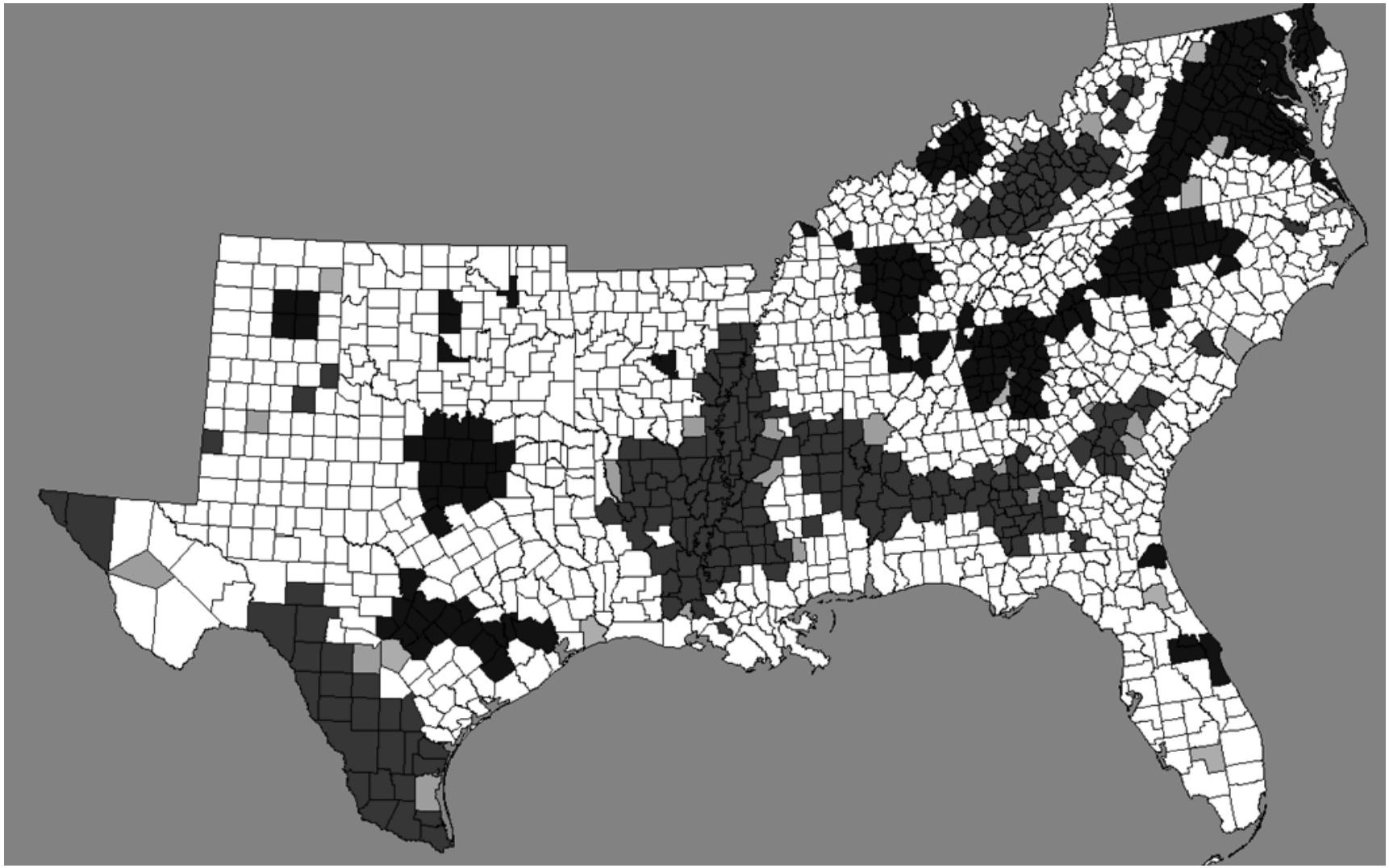
GeoDa 0.9.5-i (Beta) - [Table : south00]

File View Edit Tools Table Map Explore Space Regress Options Window Help

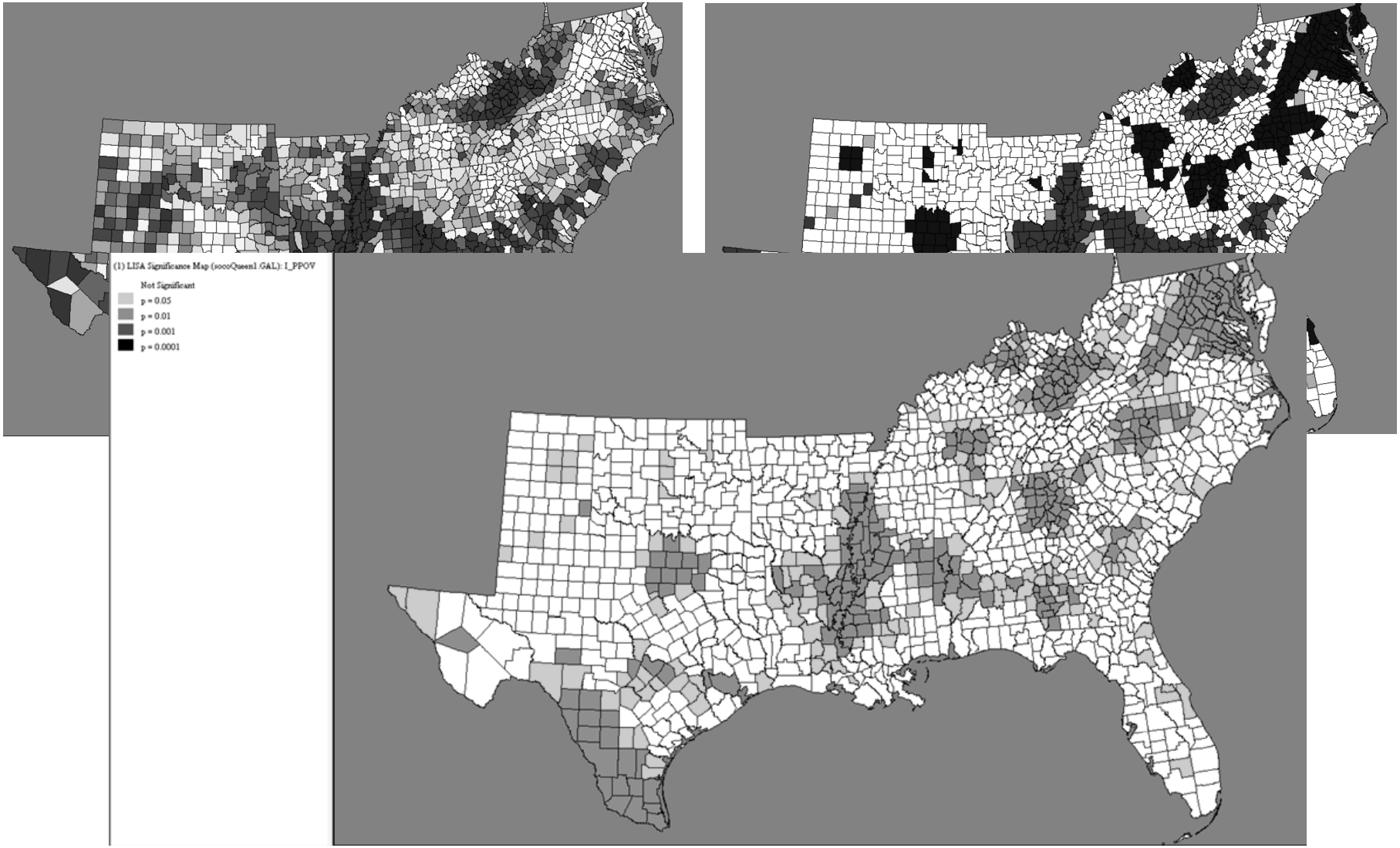
	CNTY_ST	STUSAB	FIPS	PPOV	PHSP	PFHH	PWKCO	PHSLS	PUNEM	PUD
139	Washington County AR	AR	5143	0.170803	0.081996	0.182608	0.865218	0.490111	0.078587	0.25
140	White County AR	AR	5145	0.184608	0.018819	0.170902	0.750297	0.597007	0.112925	0.23
141	Woodruff County AR	AR	5147	0.384615	0.007894	0.296703	0.745001	0.763816	0.079735	0.22
142	Yell County AR	AR	5149	0.208849	0.127300	0.181682	0.625210	0.727342	0.051796	0.20
143	Kent County DE	DE	10001	0.152381	0.032116	0.250566	0.793389	0.538092	0.055670	0.19
144	New Castle County DE	DE	10003	0.105865	0.052558	0.237606	0.855622	0.446308	0.051622	0.20
145	Sussex County DE	DE	10005	0.153209	0.044146	0.246042	0.764397	0.599651	0.048728	0.22
146	District of Columbia DC	DC	11001	0.317093	0.078581	0.500183	0.730463	0.424303	0.107963	0.26
147	Alachua County FL	FL	12001	0.198694	0.057319	0.294778	0.931430	0.287813	0.069809	0.35



LISA Map of PPOV
(1st Order Queen Weights)



LISA Map of PPOV
(1st Order Queen Weights)



Testing for spatial autocorrelation in your data is important

Unfortunately, identifying and quantifying the extent of spatial autocorrelation doesn't tell you what's *causing* it

It does alert you to the presence of Spatial "Effects" (or Spatial "Processes") at work in your data

Spatial dependence

Spatial heterogeneity

- Conceptually, these are very different processes and thus are modeled in very different ways
- Each precludes a straightforward application of standard econometric models
- Each is indicated by spatial autocorrelation

And when spatial autocorrelation in our data is indicated…

- At least one assumption of the standard linear regression model probably 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

Here’s where we pick things up tomorrow morning

Readings for today

- Anselin, Luc. 1996. "The Moran Scatterplot as an ESDA Tool to Assess Local Instability in Spatial Association." Pp. 111-125 in Fischer, Manfred, Henk J. Scholten, and David Unwin (eds.) *Spatial Analytical Perspectives on GIS: GISDATA 4* (London: Taylor & Francis).
- Tolnay, Stewart E., Glenn Dean, & E.M. Beck. 1996. "Vicarious Violence: Spatial Effects on Southern Lynchings, 1890-1919." *American Journal of Sociology* 102(3):788-815.
- Getis, Arthur. 2007. "Reflections on Spatial Autocorrelation." *Regional Science and Urban Economics* 37:491-496.
- Getis, Arthur. 2008. "A History of the Concept of Spatial Autocorrelation: A Geographer's Perspective." *Geographical Analysis* 40:297-309.
- Anselin, Luc. 2005. *Exploring Spatial Data with GeoDa: A Workbook*, (chapters 15-18).
- Anselin, Luc. 2005. *Spatial Regression Analysis in R: A Workbook,* (chapter 3).

Afternoon Lab

Spatial autocorrelation (using *GeoDa* & R)

Questions?