SPATIAL ANALYSIS AND REGRESSION

Measuring Spatial Autocorrelation

- We discussed a simple measure of spatial autocorrelation in the last lecture.
  - Moran's I is conceptually the correlation between observations and their neighbors.
- In this lecture we will discuss how to assess the significance of Moran's I using a **permutation test.**
- We will discuss **alternative measures** of **spatial autocorrelation**.
- We will discuss the difference between **global and local spatial statistics**.
- We dive into **spatial regression**.

To calculate Moran's I

- We need:
  - A variable of interest
  - A weights matrix
  - The spdep library

```
##I have pre-packaged the data into
a R data file.  This line will load
the file and create a spatial l data
frame called soco.
> load("/Users/Seth/Desktop/
soco.rda")
> soco_nbq<-poly2nb(soco) #Queen's
neighborhood
> soco_nbq_w<- nb2listw(soco_nbq)
#row standardization


>moran.mc(soco
  $PERPOV,listw=soco_nbq_w,nsim=999)
```

Data of the Day: Southern Counties Data (Voss 2012)

Variable	Description
CNTY_ST	County and state name
STUSAB	State abbreviation
FIPS	FIPS code
YCOORD	Y coordinate (meters)
XCOORD	X coordinate (meters)
SQYCORD	Y coordinate squared (for trend surface)
SQXCORD	X coordinate squared (for trend surface)
XYCOORD	X coordinate * Y coordinate (for trend surface)
PPOV	Proportion of children in poverty
PHSP	Proportion Hispanic
PFHH	Proportion female-headed households
PWKCO	Proportion work outside of county of residence
PHSLS	Proportion less than high school educated
PUNEM	Proportion unemployed
PUDEM	Proportion males underemployed, some work in 1999
PEXTR	Proportion employed in extractive industry
PPSRV	Proportion employed in professional services
PMSRV	Proportion employed in miscellaneous services
PNDMFG	Proportion employed in non-durable manufacturing
PNHSPW	Proportion non-Hispanic white
PMNRTY	Proportion minority (total - non-Hispanic white)
LO_POV	Logit transformation of proportion children in poverty
PFRN	Proportion foreign-born
PNAT	Proportion native-born
PBLK	Proportion African American/black, alone and including Hispanic
P65UP	Proportion 65 and older
PDSABL	Proportion disabled
METRO	Metro county
PERPOV	Persistent poverty, 1970-2000 (ERS)
OTMIG	Rate of out-migration
BINMIG	Rate of black in-migration from non-south
INCARC	Proportion males 18-64 in correctional institutions
BINCARC	Proportion black males 18-64 in correctional institutions
SQRTPPOV	Square root proportion children in poverty
SQRTUNEM	Square root proportion unemployed
SQRTPFHH	Square root proportion female-headed households
LOGHSPLS	Natural log proportion less than high school educated
PHSPLUS	Natural log proportion high school educated or more

How to assess the statistical significance of Moran's I

- We've seen how to calculate Moran's I using a simple linear regression.
  - Regress the variable on its lag.
- We do hypothesis tests on our regression models to see if they are significant…
- **How do you tell if Moran's I is statistically significant?**
  - There is some mathematical theory for this. The Moran's I is sort of asymptotically normal. The problem is that the weights matrix adds lots of complexity.
  - In general, people use simulation to assess the significance of Moran's I.
    - This is an alternative to making distributional assumptions (as we to in the f-test, t-test, LRT, etc.).

Moran's I : Significance

- The test is really simple....
  1. Keep the map units (polygons) constant.
  2. Randomly reassign values to the map units.
  3. Calculate Moran's I. Save the value.
  4. Return to step 2, repeat step 3, after the user specified number of iterations stop.
- The result is list of list of Moran's I values. These values represent observations of "randomness."
- Then you sort the list from lowest to highest.
- If our observation is near the beginning or end of the list we call it "a significant departure from randomness"
  - For example, if we run 999 permutations of the map plus our 1 "real" case we'd have 1000 moran's I values.
  - If our real case was the last element in the list we'd say the p-value was .001.

Assessing Moran's I

```
> MyMoran <- moran.mc(soco$PPOV,listw=soco_nbq_w,nsim=999)
> hist(MyMoran$res, breaks=50) #What is this??
> MyMoran

  Monte-Carlo simulation of Moran's I

data:  soco$PPOV
weights: soco_nbq_w
number of simulations + 1: 1000

statistic = 0.5893, observed rank = 1000, p-value = 0.001
alternative hypothesis: greater
```

Null Models for Spatial Data

- What is the null hypothesis for a randomization test?
  - The null hypothesis is a process in this case.
- Can you imagine other null models?
  - I think building null hypotheses from process models and using these to do statistical tests is a really interesting approach…
  - This isn't done often because building these process models can be difficult.
  - We'll explore how this can be done when we examine point pattern analysis.

Moran's I – Global vs Local.

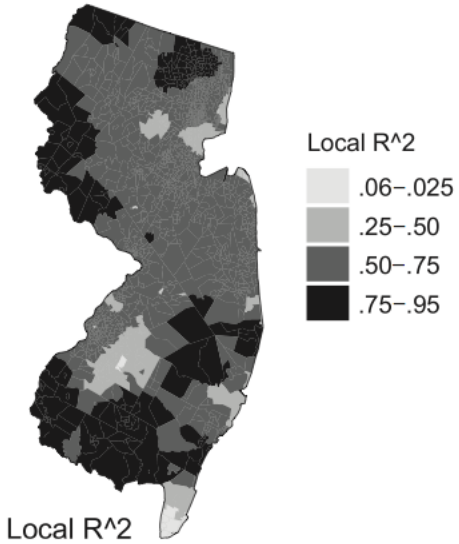
- Moran's I gives us a single number to describe spatial dependence over the entire study area.
  - It is the slope of the scatterplot (on the previous page)
- It is a "global" statistic because one number describes the entire map.
- Wouldn't it be nice to have a statistic that described "local" spatial dependence.
  - The relationship between each observation and its neighbors.
    - E.g. The relationship between Washington, DC and its neighbors.

Global vs. Local Analysis

- The desire to have statistic that describes local conditions exposes one of the fundamental methodological splits in spatial analysis:
  - Local Analysis (e.g. Local Moran's I)
    - Sensitive to local context.
    - Highlights heterogeneity
    - Conflicts with theory driven (deductive) approaches to science?
  - Global Analysis
    - One statistic to summarize the pattern in the whole study area/region
      - Clustering
      - Global dependence

Local vs Global Regression Models

Local Model
(different models for each observational unit)
Map shows r^2 for the local models



Global model

Table 3. Standardized Coefficients for Regression of PBT Density (Square Root)

Independent Variables	Model 1	Model 2	Model 3	Model 4
Constant	0.007	0.001	0.028	0.017
	(0.749)	(0.116)	(3.010)***	(1.637)
Percent Black	−0.021	−0.057		
	(−0.903)	(−2.031)*		
Percent Hispanic			0.142	0.097
			(5.335)***	(3.132)***
Population Density (natural log)	0.000	−0.036	−0.076	−0.076
	(−0.005)	(−1.476)	(−2.950)***	(−2.862)***
Percent Industrial	0.249	0.214	0.215	0.206
	(11.004)***	(9.032)***	(9.310)***	(8.673)***
Percent Living below the Poverty Line		0.097		0.037
		(3.188)***		(1.377)
Percent Manufacturing Employment		0.094		0.067
		(4.159)***		(2.679)**
N	1,933	1,933	1,933	1,933
Adjusted R^2	0.059	0.073	0.072	0.076

*$p<0.05$, **$p<0.01$, ***$p<0.005$; t-values appear in parentheses.

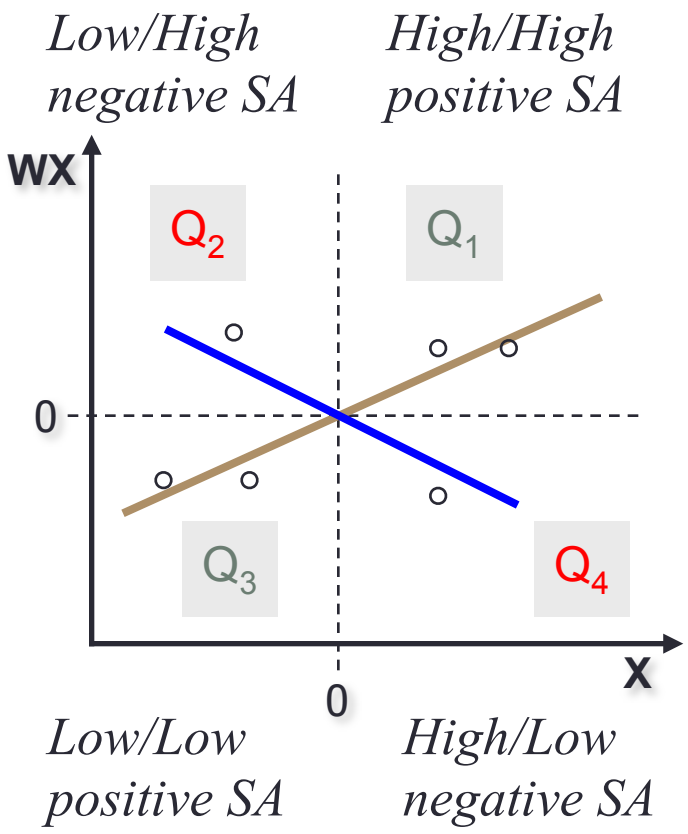
Mennis and Jordan (2006)

LISA Statistics

- Anselin (1995) developed “Local Indicators of Spatial Association” commonly called “LISA” statistics.
- In particular he developed a set of exploratory methods for a local version of the Moran’s I statistic.
- He developed the GEODA software to allow exploration of local statistics.
- The hub of these methods is a visualization called a “Moran Scatterplot”

Quadrants of Moran Scatterplot

Each quadrant corresponds to 1 of 4 different types of SA



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

Q_1 (values [+], nearby values [+]): **H-H**

Q_3 (values [-], nearby values [-]): **L-L**

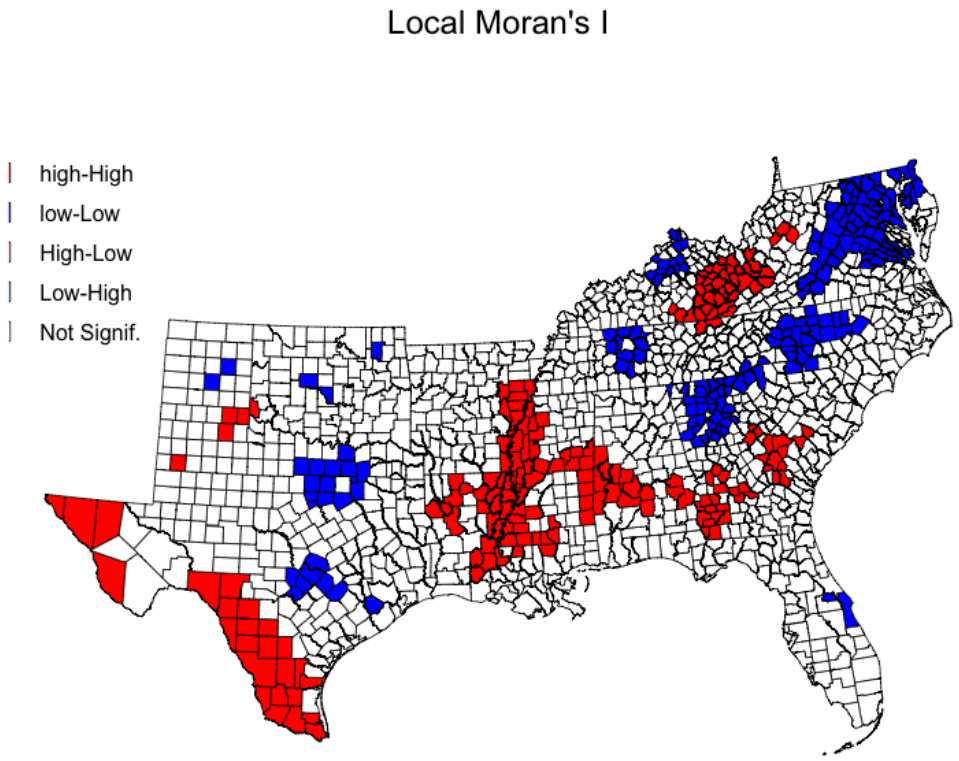
Locations of negative spatial association ("I'm different from my neighbors").

Q_2 (values [-], nearby values [+]): **L-H**

Q_4 (values [+], nearby values [-]): **H-L**

Mapping the Quadrants

THIS IS CALLED A "LISA" MAP



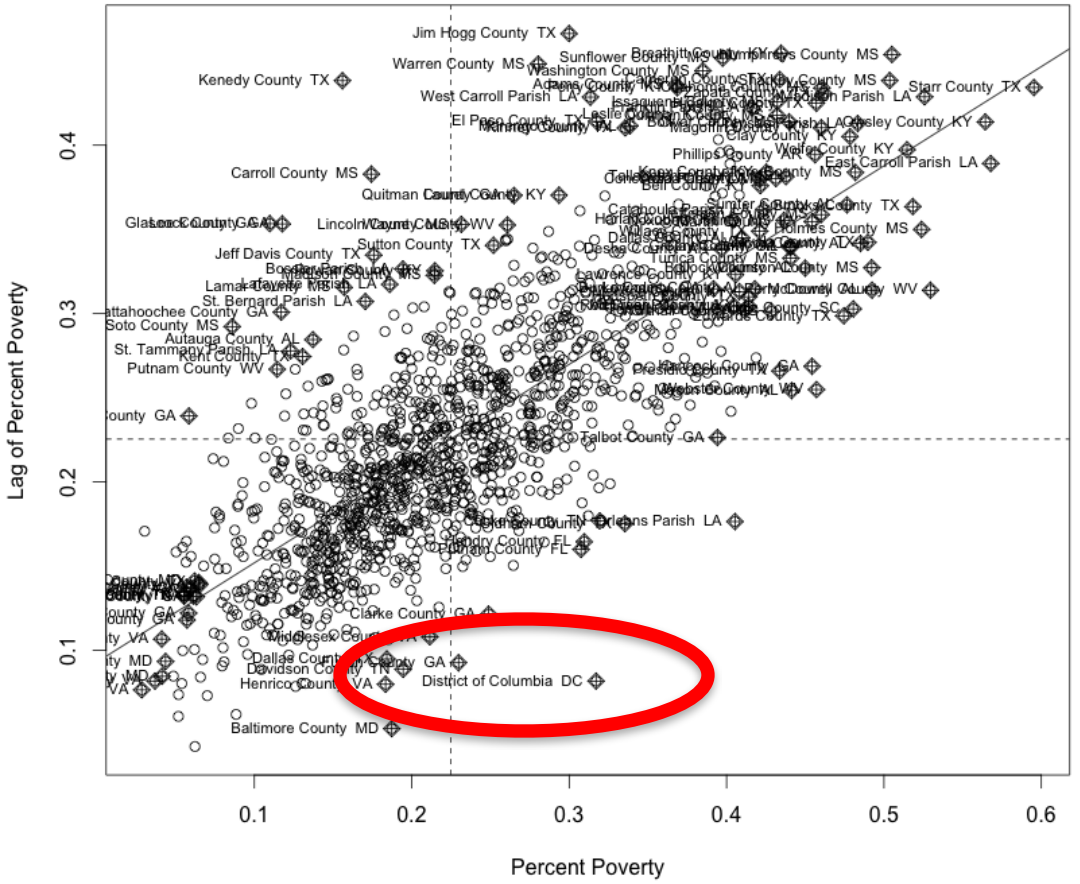
R Code to Draw LISA Map

```
library(classInt)
library(spdep)

##I used the save command to save an R-object of a spatial data frame.
load("/Users/Seth/Desktop/soco.rda")

##Create row-standardized Queens contiguity weights matrix.
soco_nbq<-poly2nb(soco) #Queen's neighborhood
soco_nbq_w<- nb2listw(soco_nbq)
locm <- localmoran(soco$PPOV, soco_nbq_w) #Calculate the local Morann
Summary(locm)

##Manually make a moran plot
##Need to standardize variables
soco$sPPOV <- scale(soco$PPOV) #save to a new column

#create a lagged variable
soco$lag_sPPOV <- lag.listw(soco_nbq_w, soco$sPPOV)

summary(soco$sPPOV)
summary(soco$lag_sPPOV)
plot(x= soco$sPPOV, y=soco$lag_sPPOV, main="Moran Scatterplot PPOV")
abline(h=0, v=0)
abline(lm(soco$lag_sPPOV ~ soco$sPPOV), lty=3, lwd=4, col="red")

##Check Out the outliers click on one or two and then hit escape (or click finish)
identify(soco$sPPOV, soco$lag_sPPOV, soco$CNTY_ST, cex=0.8)

##Identify the Moran plot quadrant for each observation
#This is some serious slicing and illustrate the power of the bracket...
soco$quad_sig <- NA
soco@data[(soco$sPPOV>= 0 & soco$lag_sPPOV>= 0) & (locm[,5]<= 0.05),"quad_sig"] <- 1
soco@data[(soco$sPPOV<= 0 & soco$lag_sPPOV<= 0) & (locm[,5]<= 0.05),"quad_sig"] <- 2
soco@data[(soco$sPPOV>= 0 & soco$lag_sPPOV<= 0) & (locm[,5]<= 0.05),"quad_sig"] <- 3
soco@data[(soco$sPPOV>= 0 & soco$lag_sPPOV<= 0) & (locm[,5]<= 0.05),"quad_sig"] <- 4
soco@data[(soco$sPPOV<= 0 & soco$lag_sPPOV>= 0) & (locm[,5]<= 0.05),"quad_sig"] <- 5 #WE ASSIGN A 5 TO ALL NON-SIGNIFICANT OBSERVATIONS

#####MAP THE RESULTS (courtesy of Paul Voss) #############
# Set the breaks for the thematic map classes
breaks <-seq(1,5,1)

# Set the corresponding labels for the thematic map classes
labels <- c("high-High", "low-Low", "High-Low", "Low-High", "Not Signif.")

#see ?findInterval - This is necessary for making a map
np <- findInterval(soco$quad_sig, breaks)

# Assign colors to each map class
colors <- c("red", "blue", "lightpink", "skyblue2", "white")
plot(soco, col=colors[np])  #colors[np] manually sets the color for each county
mtext("Local Moran's I", cex=1.5, side=3, line=1)
legend("topleft", legend=labels, fill=colors, bty="n")
```

Moran Plot in R

```
mp <- moran.plot (soco
  $PPOV, soco_nbq_w,
  labels=as.character
  (soco$CNTY_ST),
  xlab="Percent
  Poverty",
  ylab="Lag of Percent
  Poverty")
```

- **This plot is a mess! But is interesting.**
- **Why are some places labeled?**
  - **Uses a 2-sigma rule to assess significant outliers, labels these.**
- **You could make this plot manually? How?**

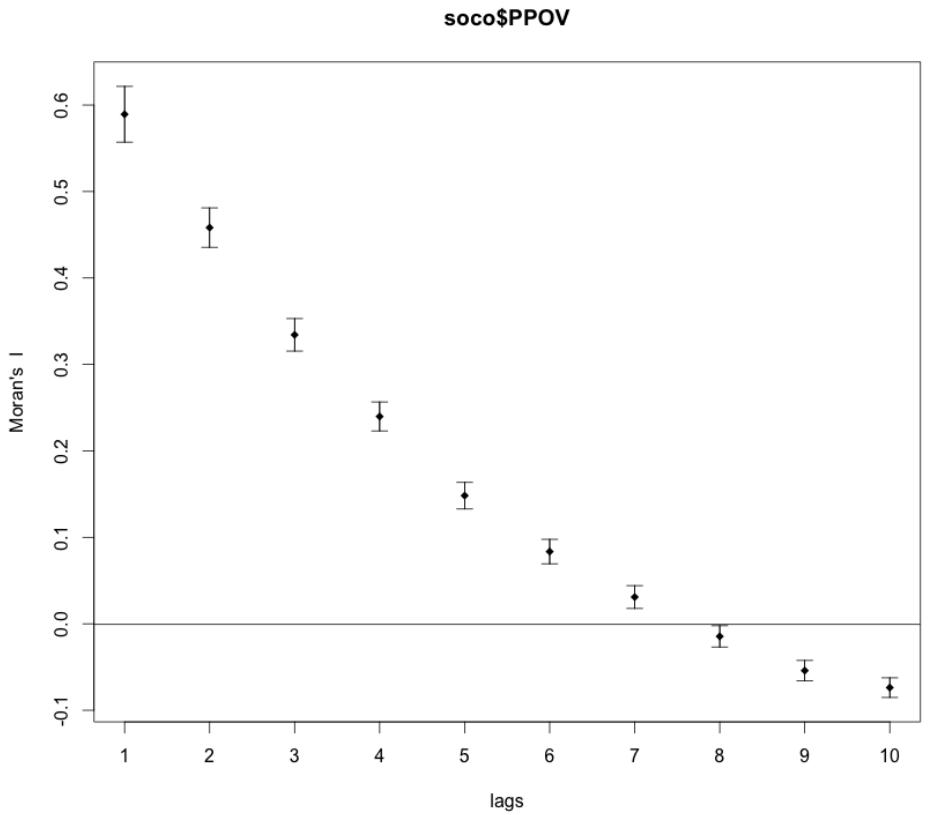


But what about the weight matrix…

- We discussed scale at length.
  - I argued that how you conceptualize and measure spatial questions has an impact on the results of an analysis.
  - Results are often scale dependent.
  - Scale dependence begs a host of questions about finding the “right” or best scale for an analysis.
  - We’ll explore two tools to help us gain insight into scale.
    - Correllogram
    - Variogram

Correlogram of spatial lags

```
> cor10<-sp.correlogram(soco_nbq, soco$PPOV, order=10, method="I")
> cor10
> plot(cor10)
```



Error bars are based on the SE
From randomization (permutation)

INTERPRETATION??

Why do a correllogram?

- Because scale matters…
- To assess sensitivity to scale of Moran's I to scale.
  - Tool is limited in that it will only calculate higher-order lags of a given nb object. Will not test various types of nb objects but one easily conduct such a test.

- Why not do a correllogram:
  - Because it doesn't tell you much about the spatial structure of the underlying data.
    - It tells you about spatial dependence as a function of lags.
    - It is conditional upon a weights matrix.

Extending the logic of a correlogram

- Conceptually the correlogram is looking at the covariance between location s_i and location s_j.
- The statistic looks at departures from the mean.
- We can demote some outcome Y at location s within region R:

$$\{Y(s), s \in R\}$$

- Where the expected value of Y at location s:

$$E\{Y(s)\} = \mu(s)$$

- And the variance is:

$$VAR\{Y(s)\} = \sigma^2(s)$$

Extending the logic of a correlogram

- A map is said to ***stationary*** if:

$$\mu(s) = \mu$$
$$\sigma^2(s) = \sigma^2$$

- The mean and the variance are constant on the map (i.e. constant within the region *R*).
- When these conditions are met the covariance between locations s_i, and s_j is simply a function of the distance separating the locations.

Spatial Structure

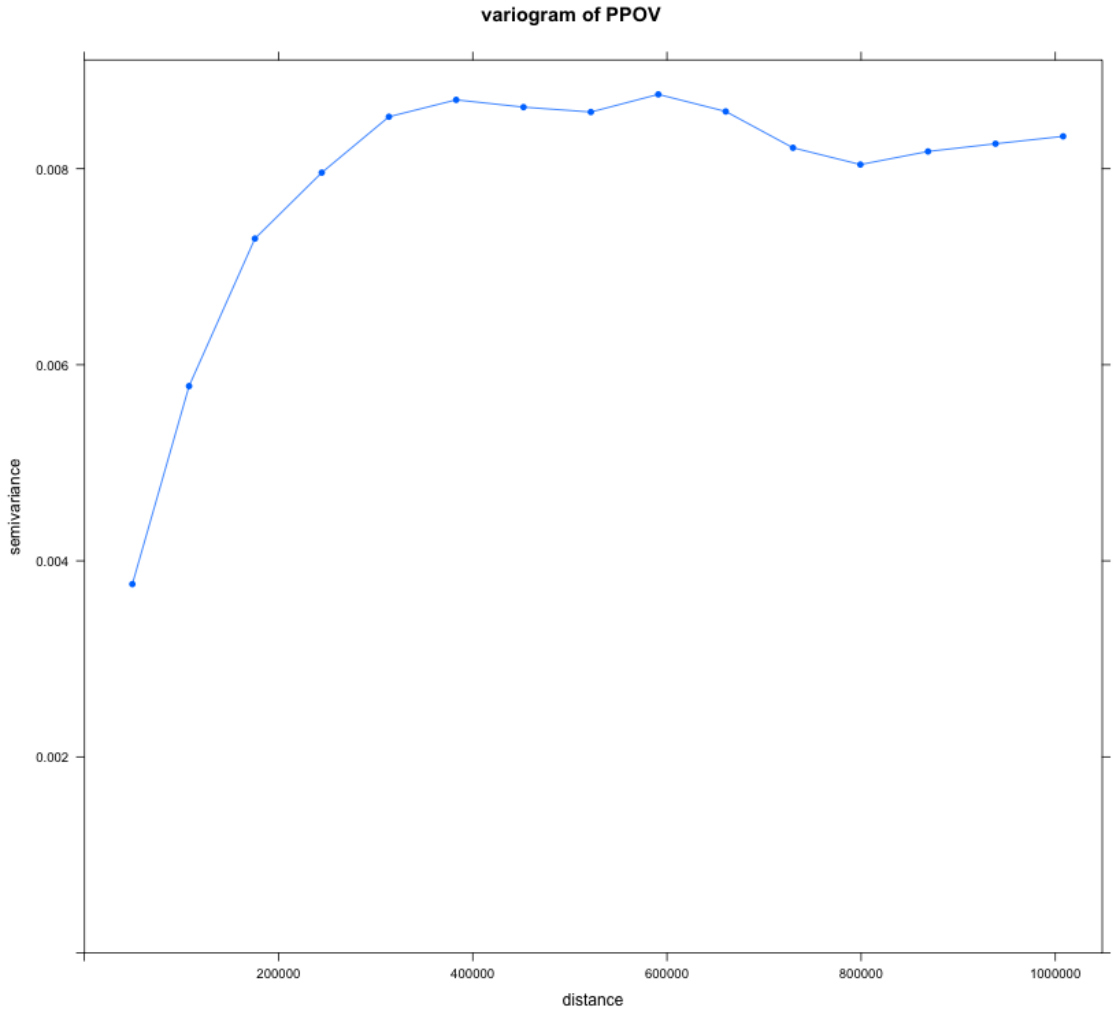
- In geostatistics spatial autocorrelation is conceptualized as a function of distance.
  - Geostats people see spatial dependence as a function NOT a statistic.
  - Measured using a semivariogram, usually just called a variogram (there is a technical difference between the two but for clarity I'll use the term "variogram" generally).
- We look at pairs of observations separated by a distance h.
  - This gives us a lot of data! Every observation has n-1 possible pairs.
    - For fun (if you have a fast computer) modify the code on the next slide to read "cloud=T"
- For each pair we know the distance between points.
- We measure the similarity of the point pairs (semivariance).
- We plot the semivariance for a range of distances, where "distance" is the separation of pairs of points.

Calculating semivariance

- N(h) denotes the set of pairs of observations separated by distance h.
  - |N(h)| is the number of pairs in the set.
  - h is usually an approximate distance implemented using a certain tolerance.
  - The “E” under the summation denotes all i,j pairs within the set N(h)

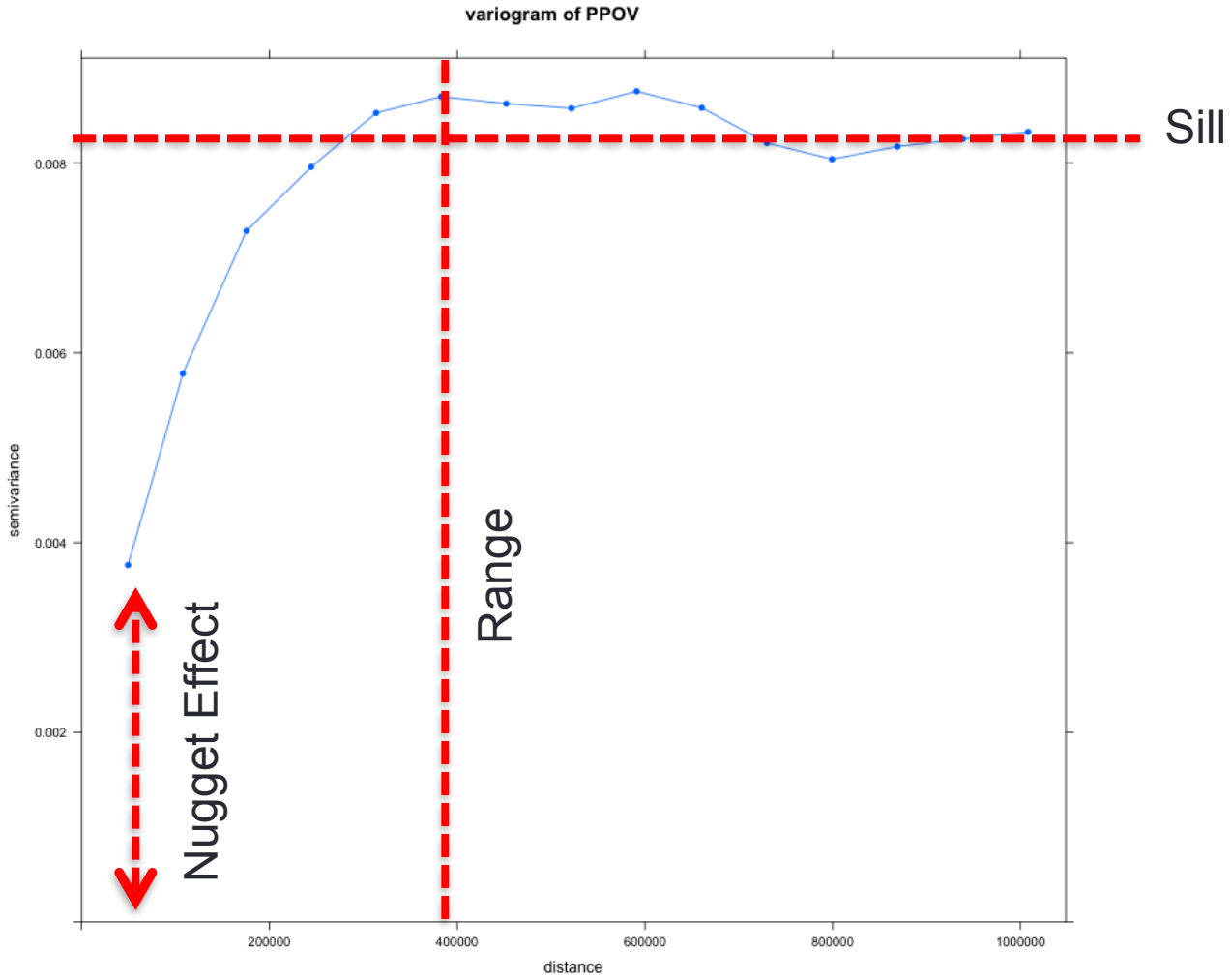
$$\hat{\gamma}(h) := \frac{1}{|N(h)|} \sum_{(i,j) \in N(h)} |x_i - x_j|^2$$

Reading a Variogram




```
library(gstat)
plot(variogram(soco$PPOV~1, locations=coordinates(soco), data=soco, cloud=F),type="b”)
```

Reading a Variogram




```
library(gstat)
plot(variogram(soco$PPOV~1, locations=coordinates(soco), data=soco, cloud=F),type="b")
```

Variogram Parts

- Range: The distance at which point pairs stop being similar to each other.
  - If we're sampling soil contamination.
  - We find that the nearby samples have similar levels of PCB.
  - We fit a variogram and find that the range is 10m
    - This means that samples separated by less than 10m will be similar. Once samples are more than 10m apart might as well be 10km apart.
  - Beyond the range samples are independent from each other.
- Sill: The background level of variance. Sort of a baseline for the entire study region.
- Nugget: Small scale discontinuity.

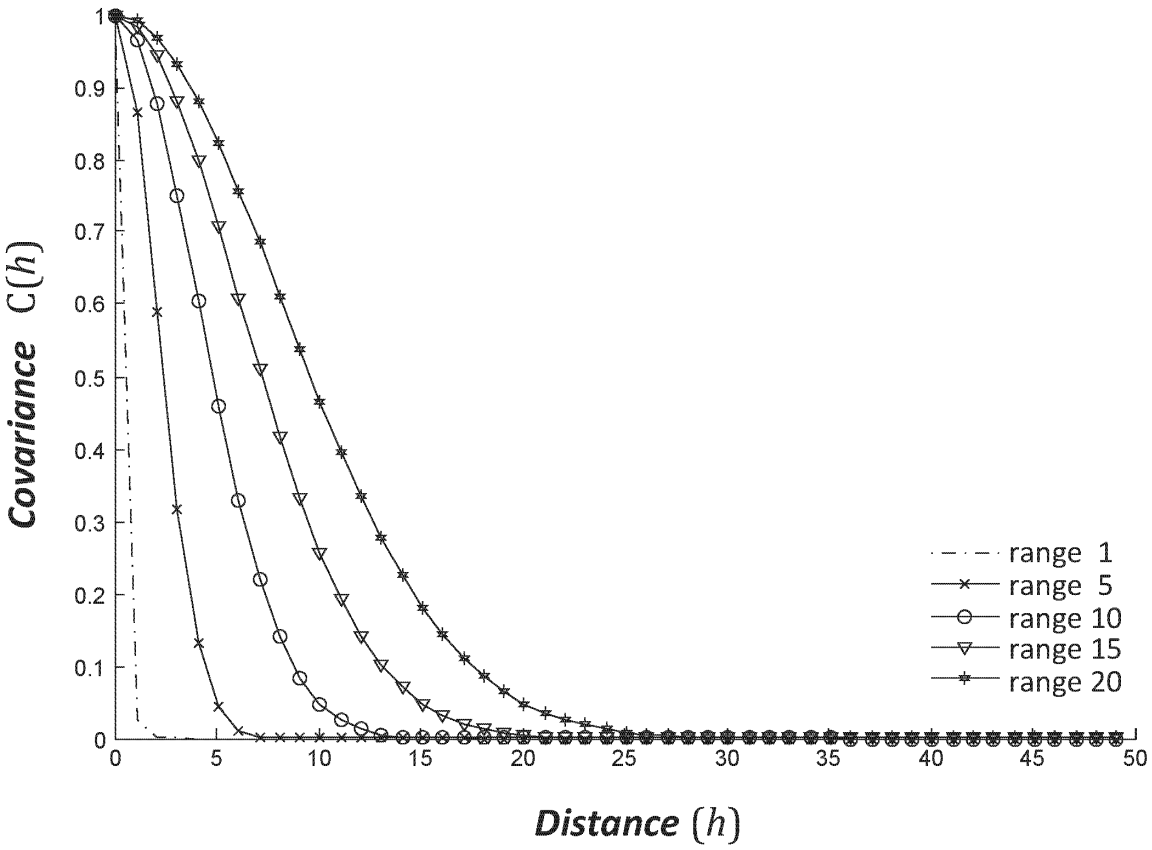
Variogram Parts

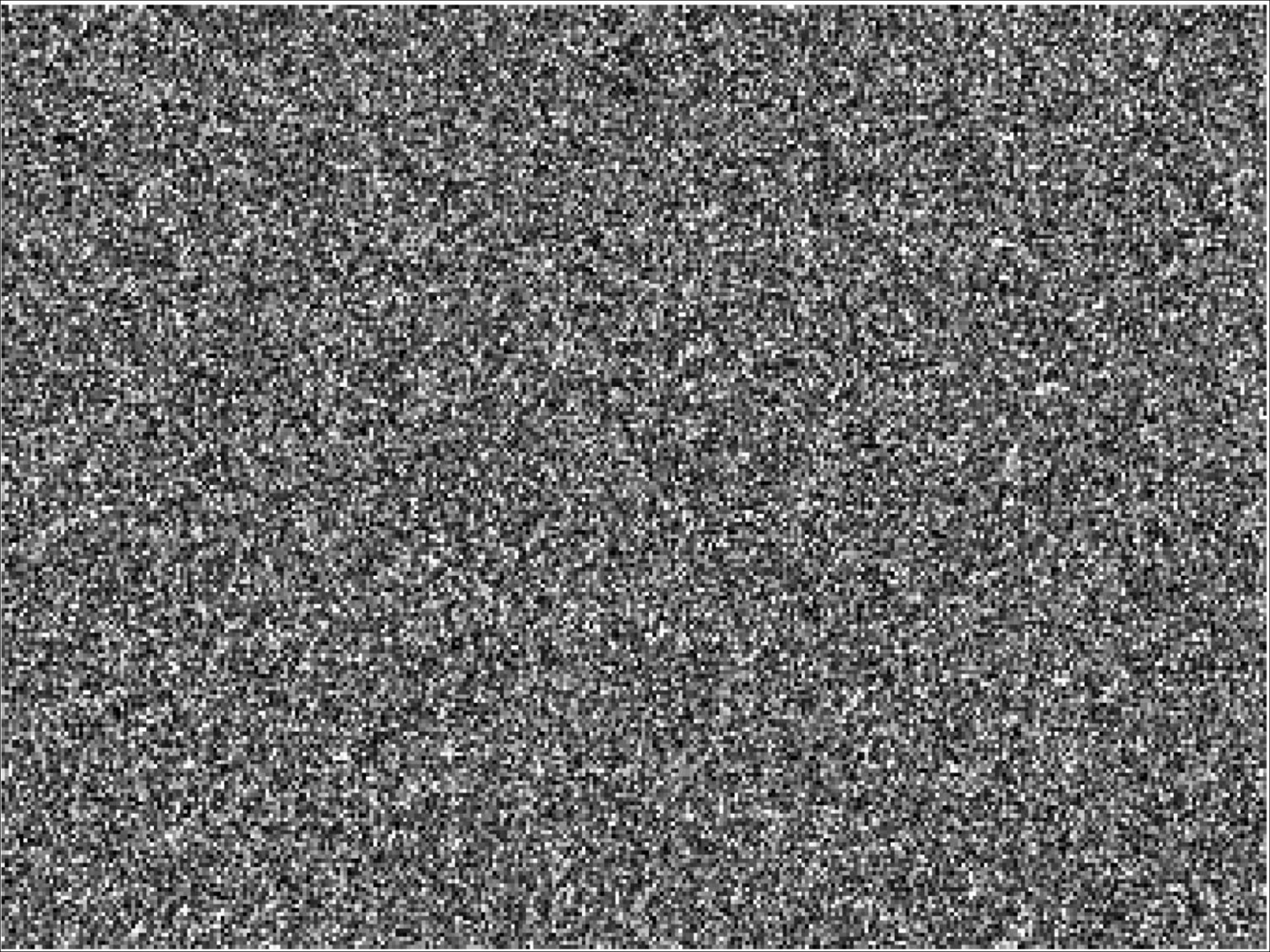
- Nugget: Small scale discontinuity.
  - In theory on the left a variogram should approach zero.
  - Measurements in the same sample location should be identical.
  - In practice this often isn't the case.
  - The nugget shows the amount of micro-scale variability/ heterogeneity.
- In practice there is an important distinction between a theoretical variogram and an empirical variogram.
  - An empirical variogram is used to fit a theoretical variogram for modeling purposes.

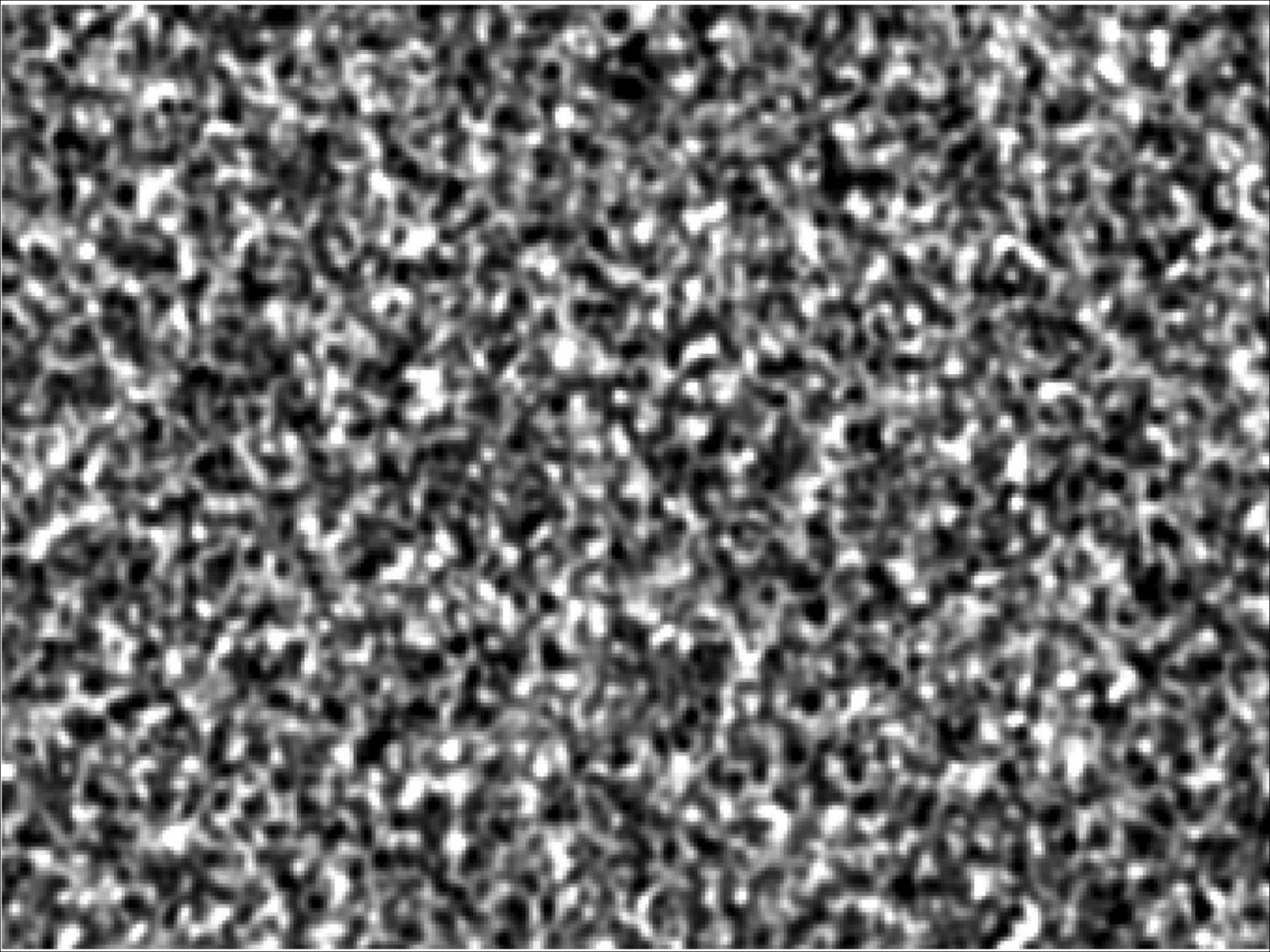
Using the Variogram

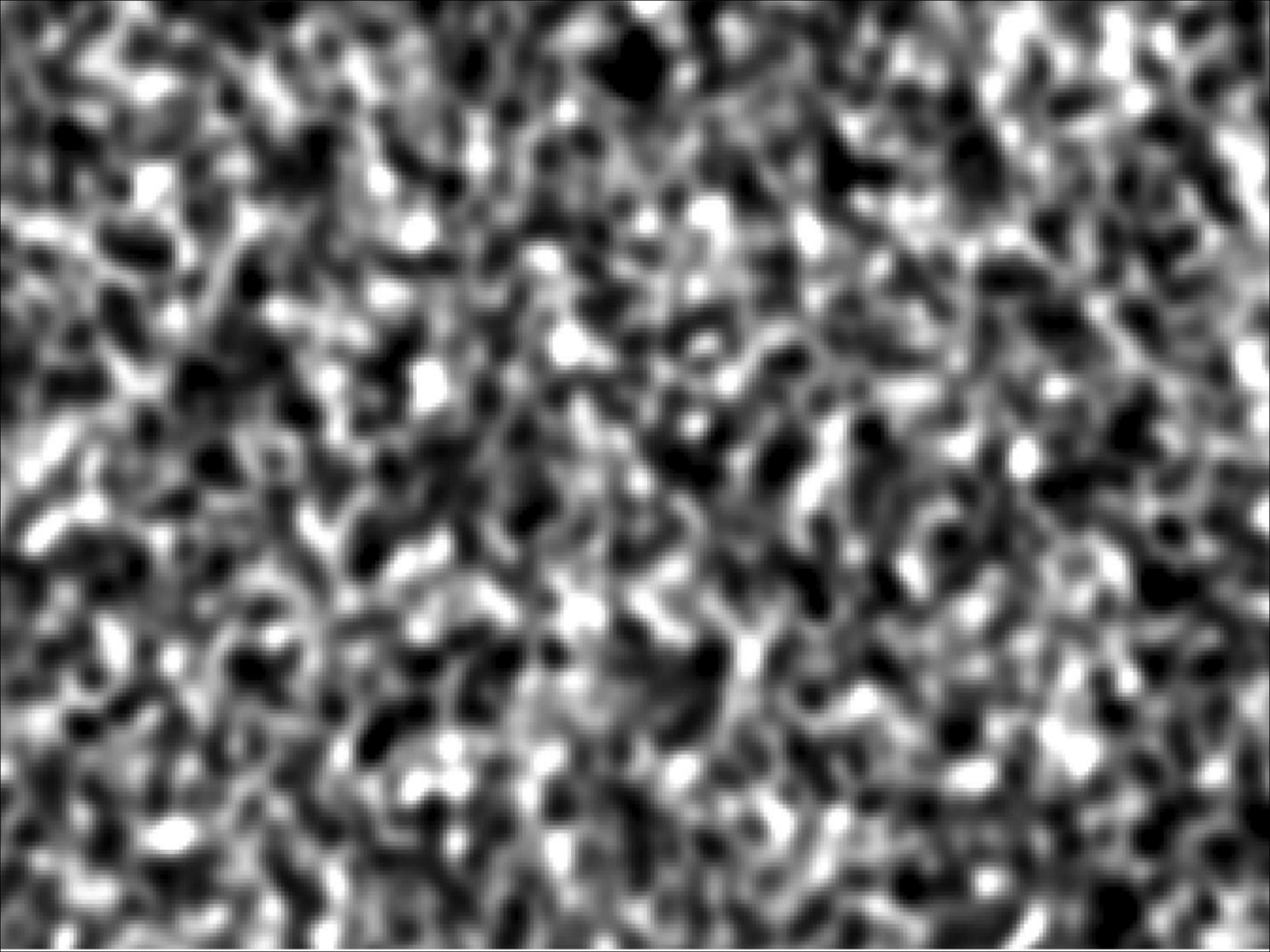
- Descriptive tool…
  - Illustrates the **spatial structure** of a variable.
  - Often used for prediction (kriging) – predicting values at unknown locations.
  - Examining correlation as a function of distance would be intuitive…
    - You can do:
      - sp.correlogram(soco_nbq, soco$PPOV, order=10, method="corr")
        - But this depends upon the weights matrix.

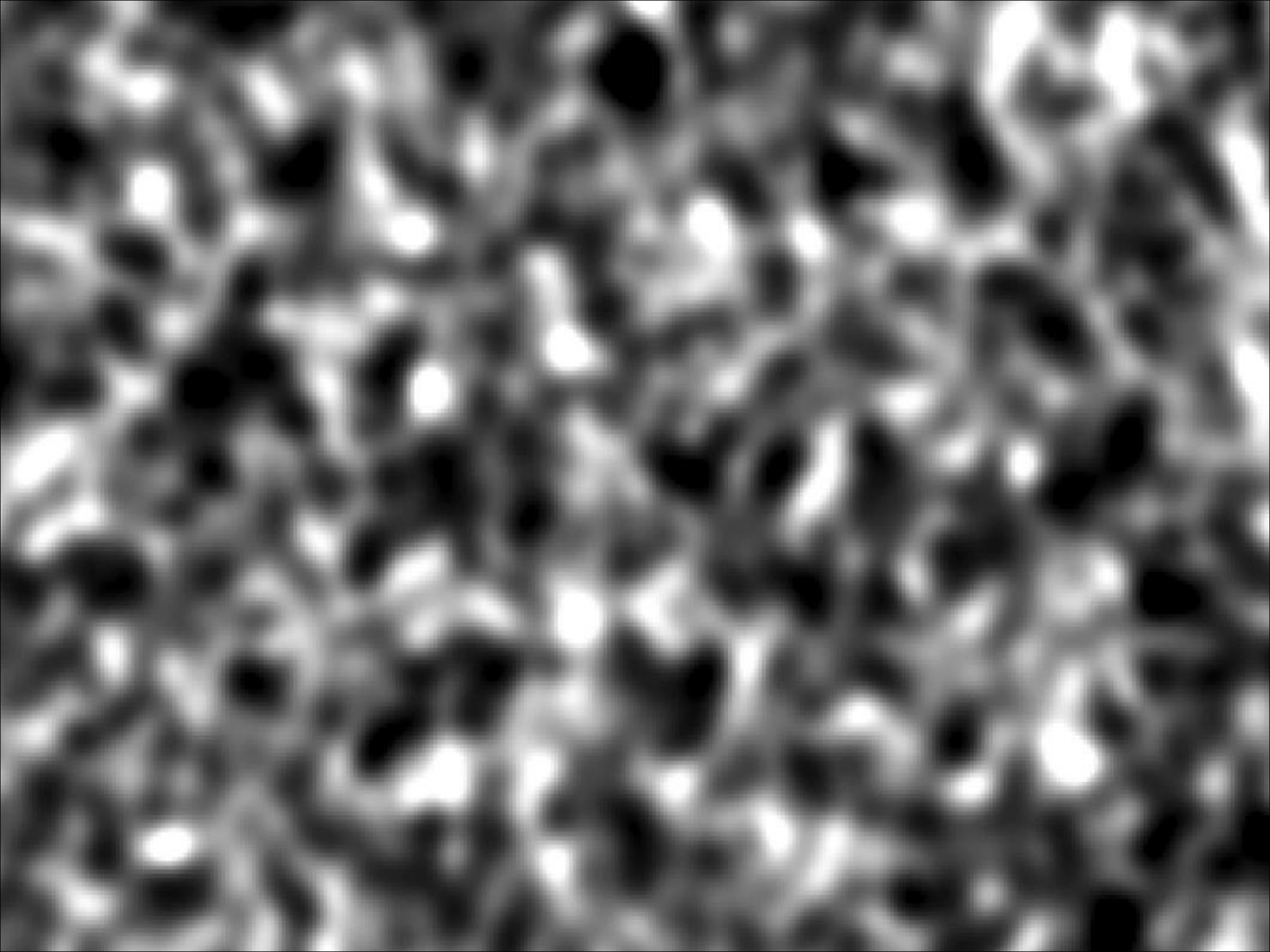
Covariance by distance

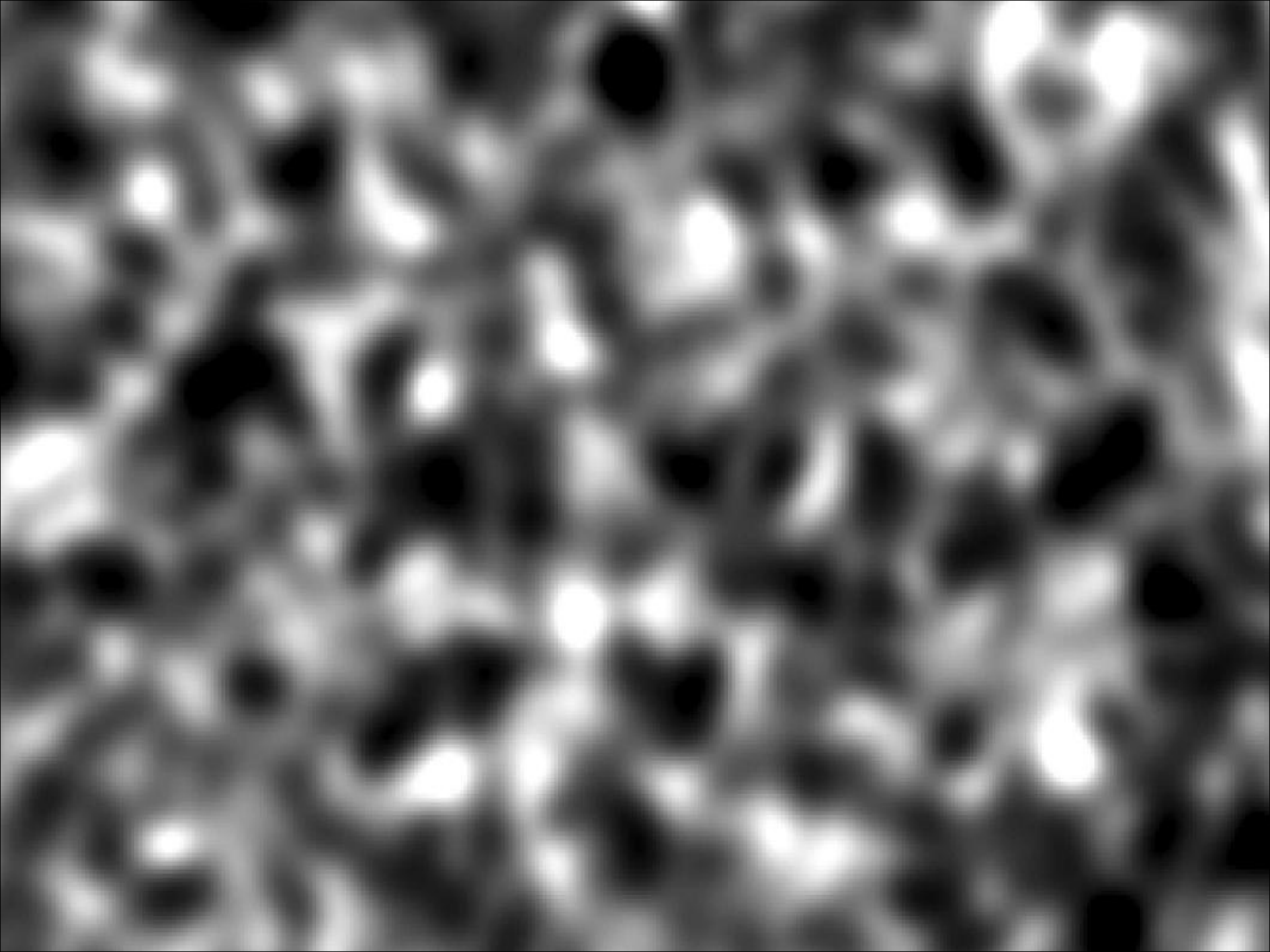






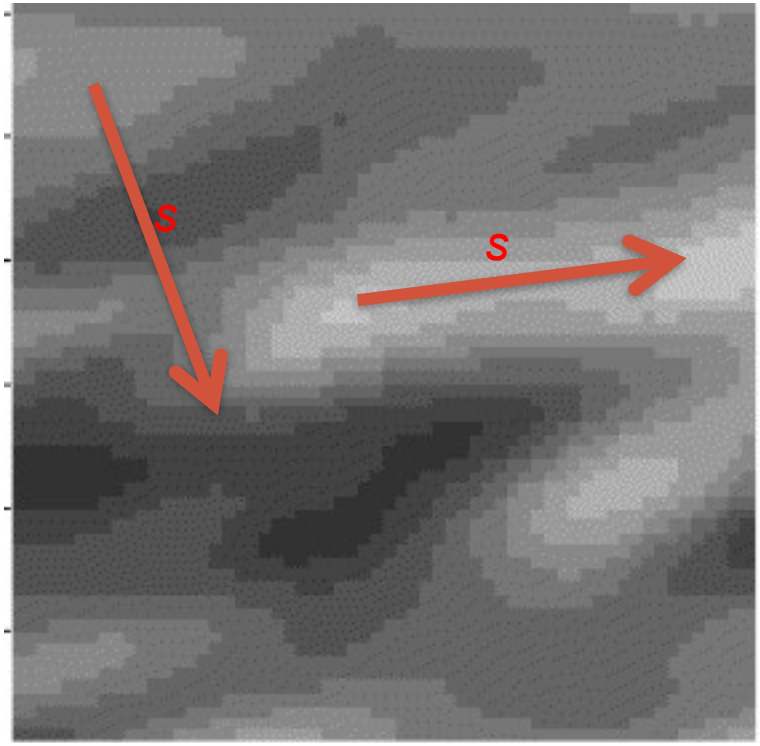






Autocovariance

- There are many possible autocovariance functions.
- In spatial situations these can get intense.
  - Imagine if the the correlation between two observations was not only a function of the distance separating them but also the direction.



HOLD UP!

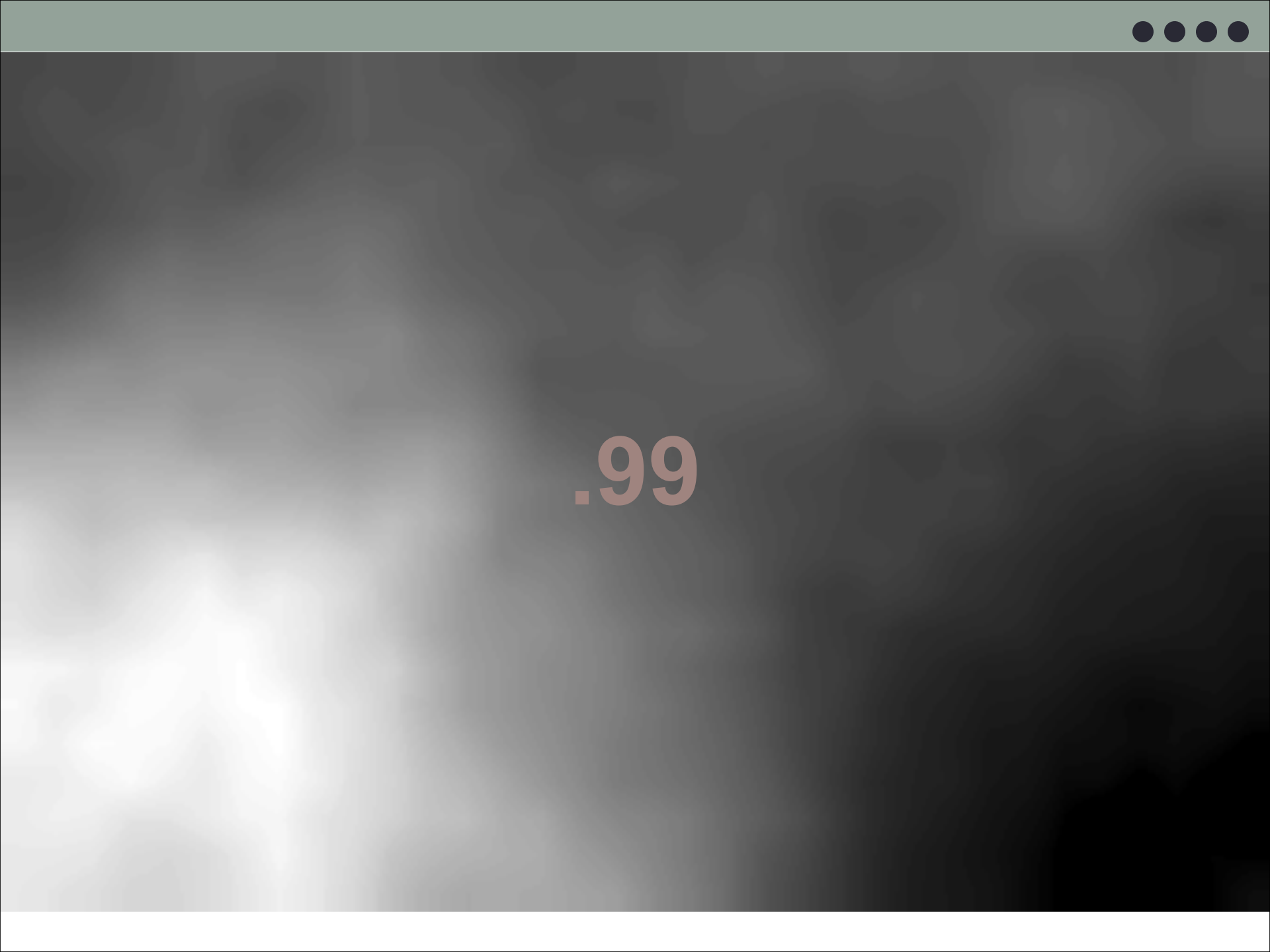
- We started talking about spatial regression.
- Then we discussed temporal autocorrealtion in the residuals.
- Now, we are talking about spatial issues again…
- Why!!!

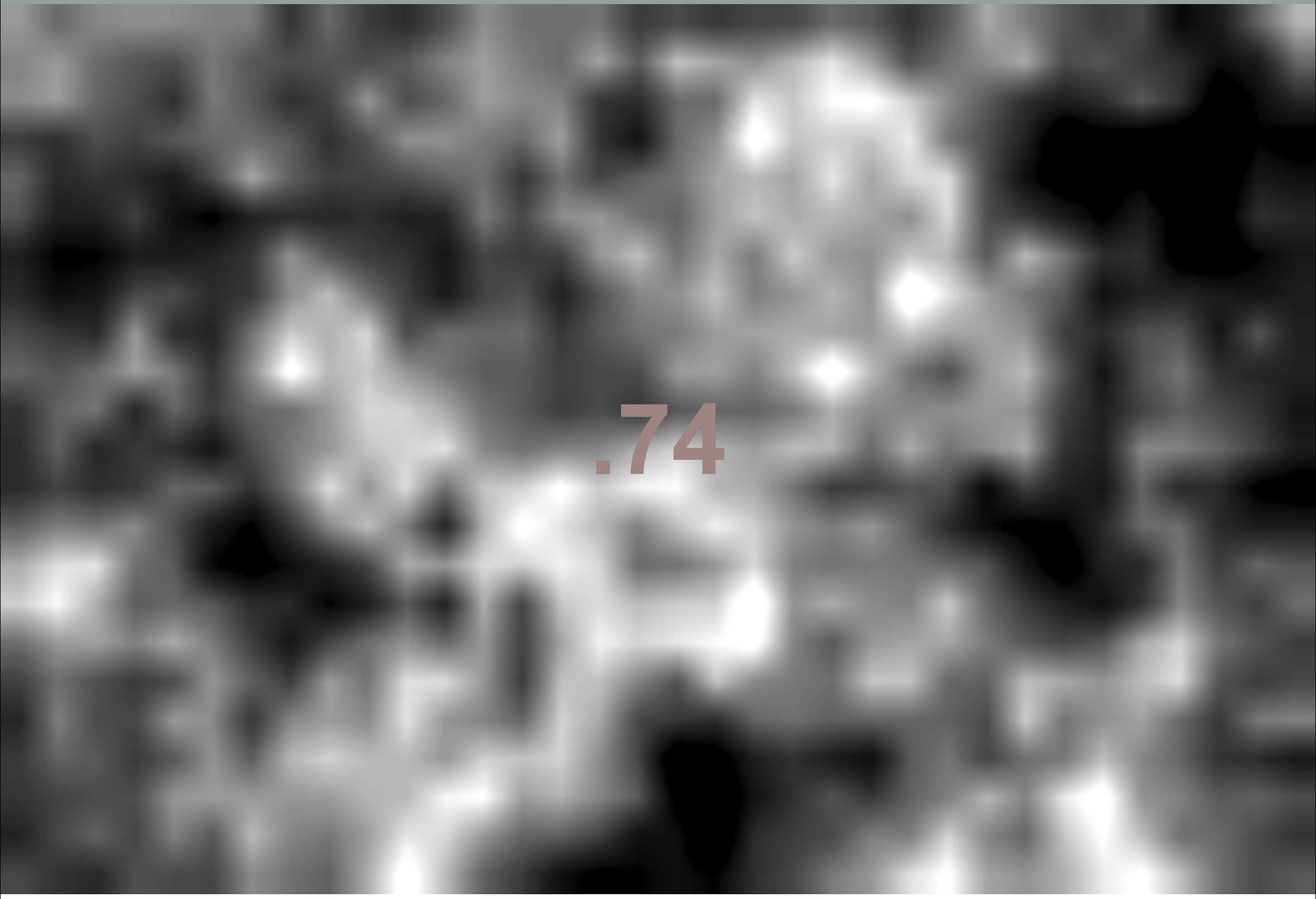
- Serial autocorrelation in regression residuals and spatial autocorrelation are the same problem (more or less).
- Some of the early innovations in spatial regression (in the 1960s) we by Durbin, who developed methods for dealing with temporal autocorrelation.

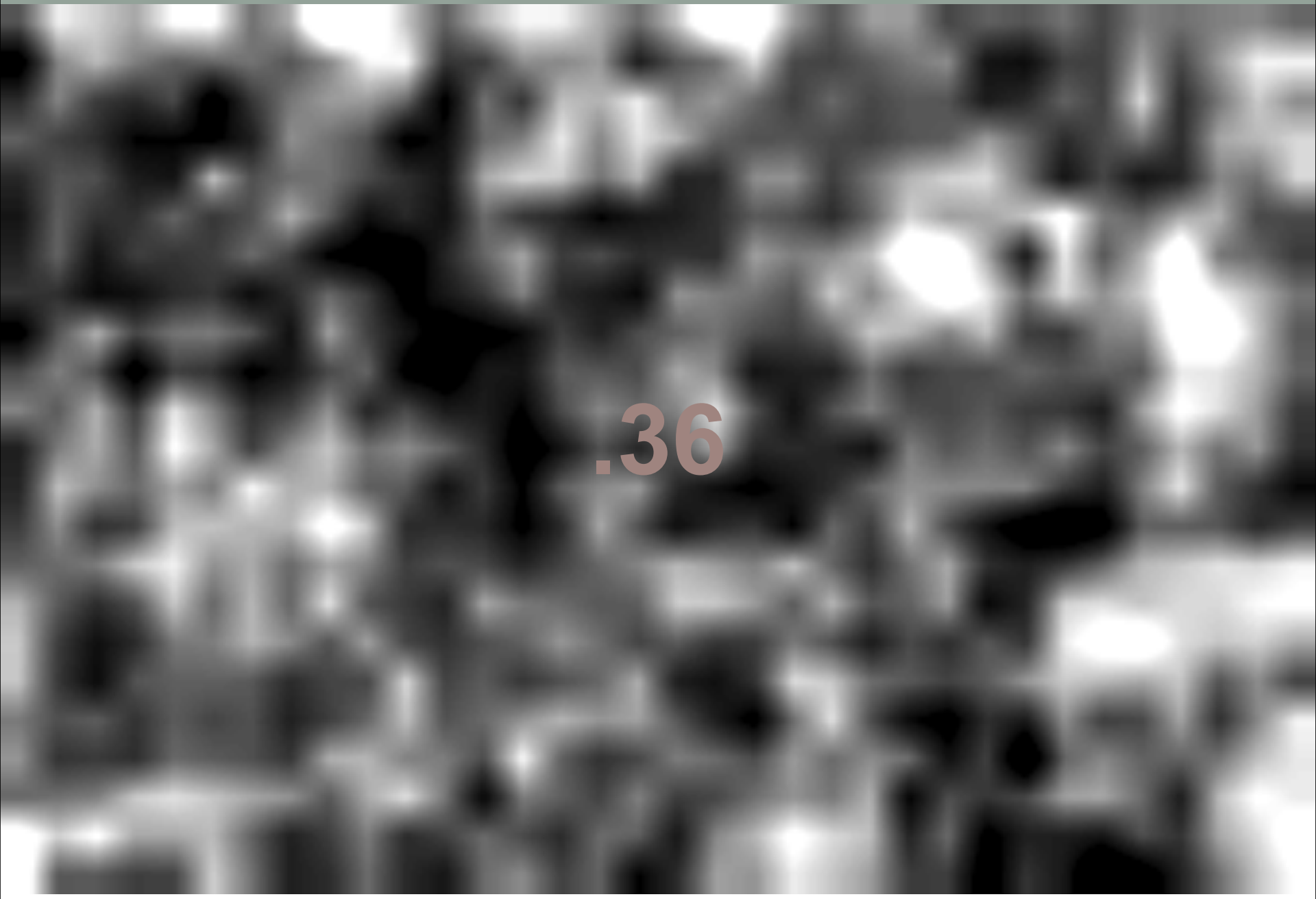
Moran's I: Is spatial covariance.

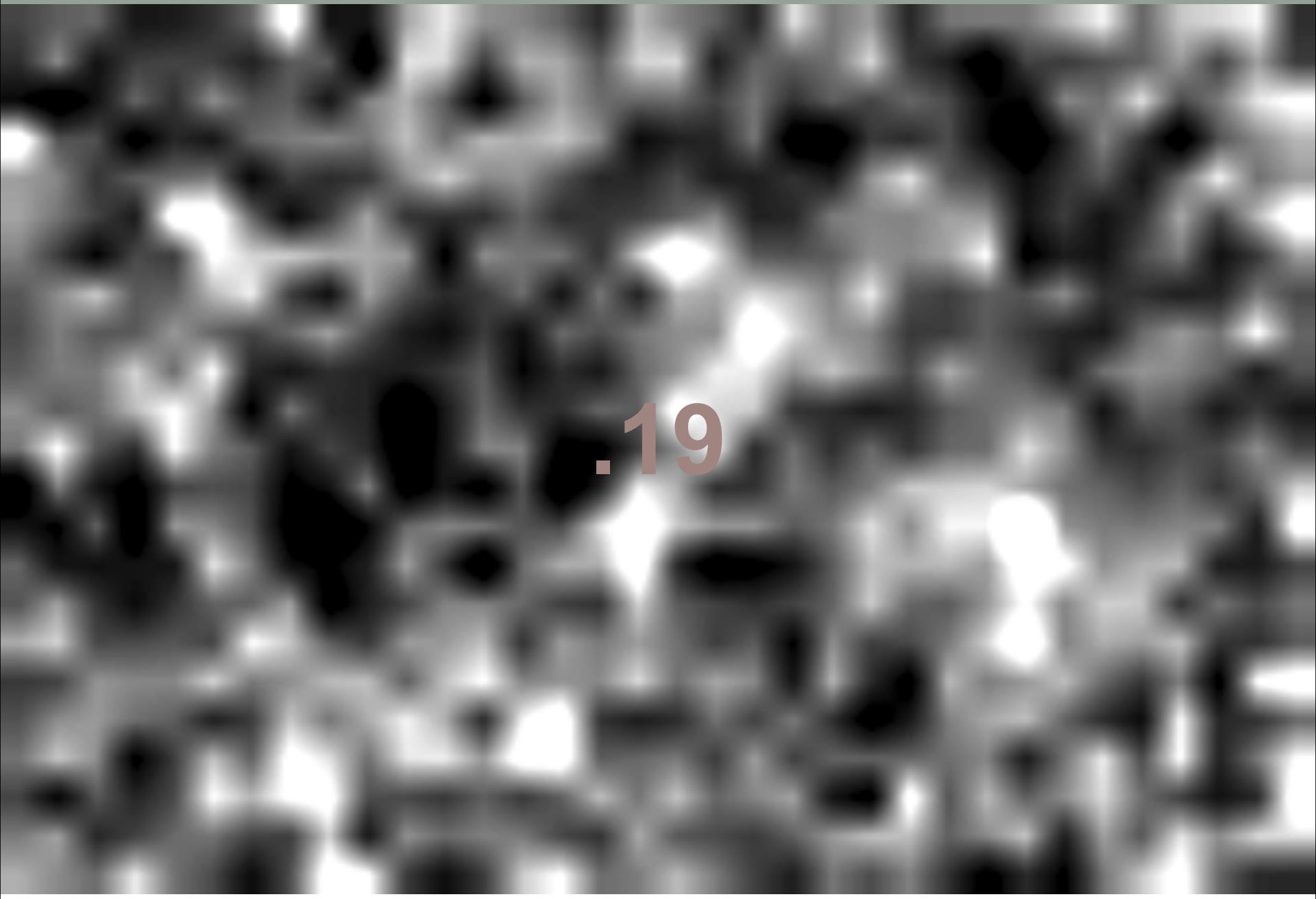
- It's the spatial version of autocorrelation….

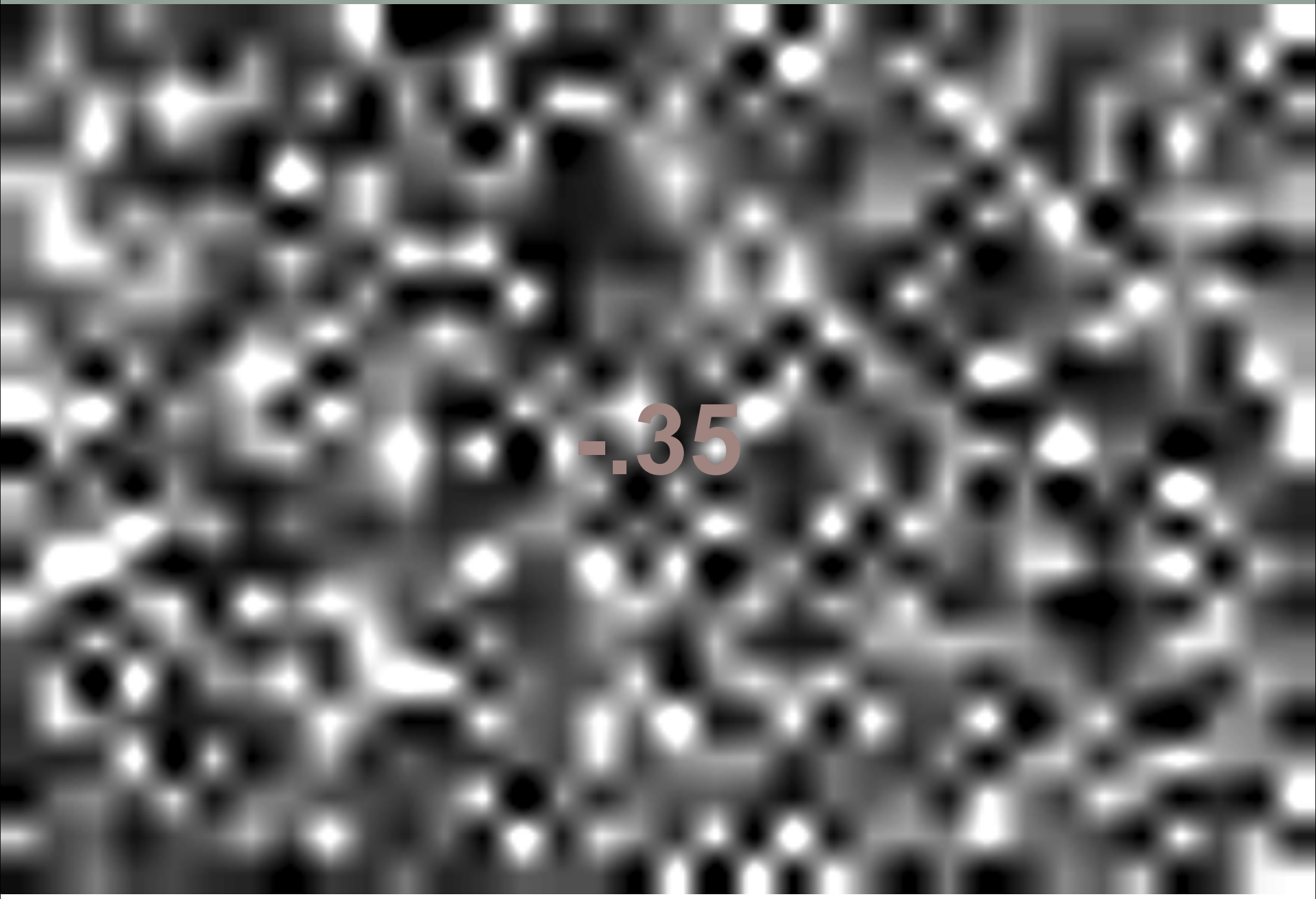
$$I = \frac{n \sum_{i}^{n} \sum_{j}^{n} w_{ij} \left(y_i - \bar{y} \right) \left(y_j - \bar{y} \right)}{\left(\sum_{i}^{n} \sum_{j}^{n} w_{ij} \right) \sum_{i}^{n} \left(y_i - \bar{y} \right)^2}$$

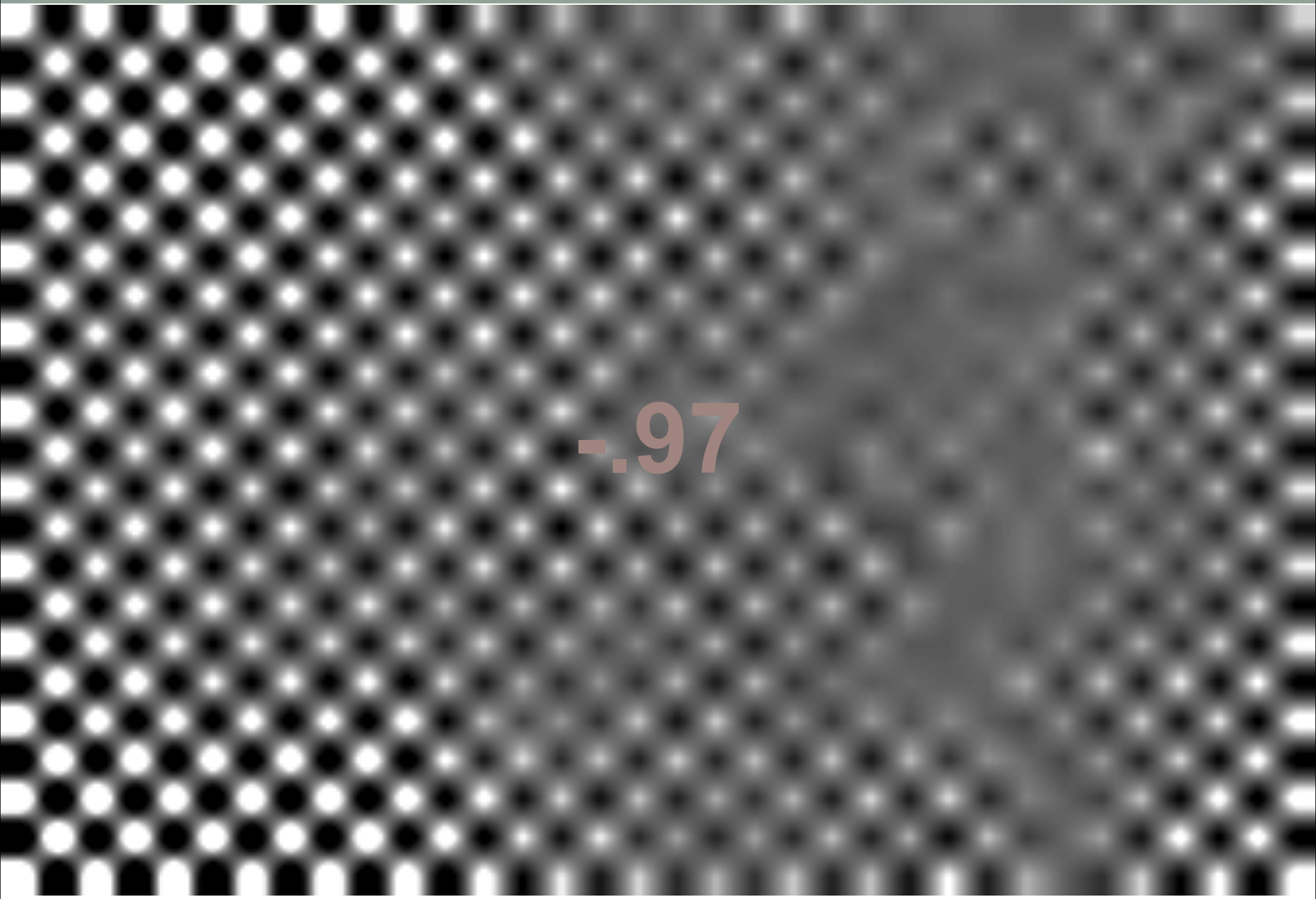

.99


.74


.36


.19


-.35


-.97

Spatial Dependence

- There is no single way to describe spatial dependence. Focusing on the spatial structure of a variable can give you more insight – but not answers…
- Moran's I is useful and intuitive but it depends on some a priori assumptions (a weights matrix).
- A variogram:
  - Can be used to describe spatial structure
  - To fit statistical models (geostatistics).

Spatial Analysis

- Often in statistical analysis spatial relationships are ignored
  - As we will show this weakens our ability generate meaningful inferences about the processes we study.
- The solution to this problem is to account for spatial relationships within regression models.
- Spatial regression models include spatial relationships within the model.

Why worry about spatial relationships?

- There are **substantive reasons**:
  - Perhaps an outcome Y is not entirely explained by the X's. Nearby values of Y might also have an effect.
  - For example the percent of trees killed by beetles might be due to stand characteristics AND the percent of trees in NEARBY stands that have been killed by beetles.

Why worry about spatial similarities?

- There are **statistical reasons**:
  - Violation of regression assumptions:
    - Residuals are correlated with each other.
  - If we ignore the spatial relationships in our data:
    - Our estimated regression coefficients are biased and/or have inflated variance.
    - Which leads to incorrect inference…
  - If spatial effects are present, and you don't account for them, your model is not accurate!

Two types of spatial models

- Spatial Lag Model
  - **Space is of substantive** interest.
  - Modeled as a simultaneous equation, the outcome at location ***a*** depends (in part) on the outcome at location ***b***.
- Spatial Error Model
  - **Space is a nuisance** which affects our ability to accurate estimate the impacts of the predictor variables on the outcome.
  - Modeled with “space” in the error term.
- Which model is “best.”
  - It is best to distinguish between these models on theoretical grounds…
  - There are some statistical tests to aid in this process.

Model Types

- Spatial lag

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

The spatial part

- Spatial error

$$y = X\beta + \epsilon$$

$$\epsilon = \lambda W \epsilon + \mu$$

The spatial part

Spatial Lag Model

- Incorporates spatial effects by including a spatially lagged dependent variable as an additional predictor

$$y = \rho W y + \beta_1 X_1 + \beta_2 X_2 + \cdots + \epsilon$$

- Where:

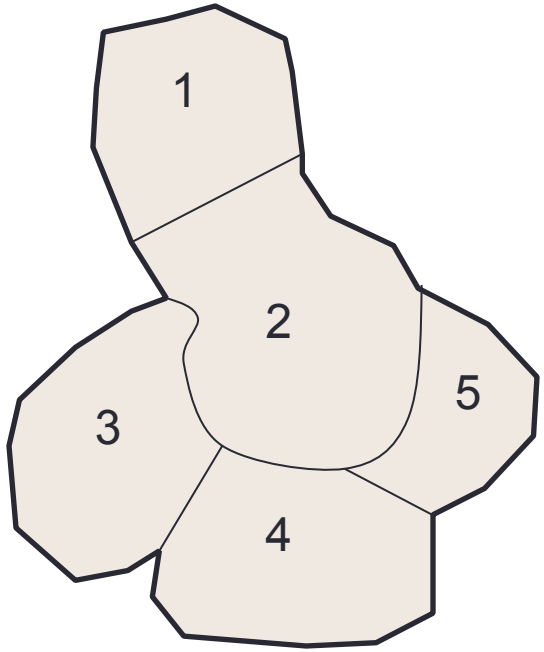
$$W y_i = \sum_{j=1}^{n} w_{ij} y_{(j, j \neq i)}$$

- If there is no spatial dependence, and y does not depend on neighboring y values, $\rho = 0$

How do we calculate that spatial lag term?

- The average of all neighbors

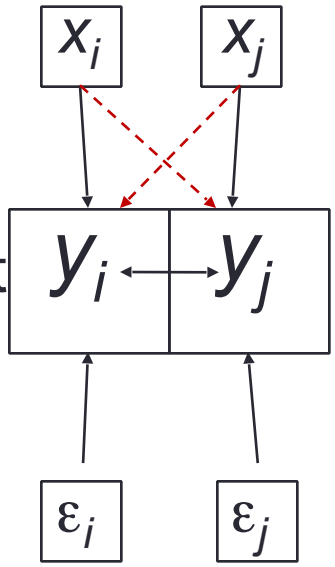
$$W = \begin{matrix} 0 & 1 & 0 & 0 & 0 \\ .25 & 0 & .25 & .25 & .25 \\ 0 & .5 & 0 & .5 & 0 \\ 0 & .33 & .33 & 0 & .33 \\ 0 & .5 & 0 & .5 & 0 \end{matrix}$$



Area	y	*Wy*
1	5	(1*7)=7
2	7	(.25*5)+(.25*9)+(.25*12)+(.
3	9	(.5*7)+(.5*12)=9.5
4	12	(.33*7)+(.33*9)+(.33*11)=8.91
5	11	(.5*7)+(.5*12)=9.5

The simple spatial lag model

- Incorporates spatial dependence by adding a "spatially lagged" variable (y) on the right-hand side of the regression equation
- Treats spatial correlation as a process or effect of interest
  - The values of y in one area are directly influenced by the values of y found in neighboring areas
  - Depends on how to we define neighborhood
- Positive spatial lag provides evidence that the y's in adjacent areas covary



OLS

```
> soco_OLS <- lm(PPOV ~ PFHH + PUNEM + PBLK + P65UP, data=soco)
> summary(soco_OLS)

Call:
lm(formula = PPOV ~ PFHH + PUNEM + PBLK + P65UP, data = soco)

Residuals:
      Min        1Q    Median        3Q       Max
-0.265701 -0.037108 -0.007851  0.032944  0.287326

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.055440   0.008913  -6.220 6.58e-10 ***
PFHH         0.457478   0.049409   9.259  < 2e-16 ***
PUNEM        2.181089   0.073337  29.741  < 2e-16 ***
PBLK        -0.059190   0.018781  -3.152  0.00166 **
P65UP        0.380227   0.042763   8.892  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0576 on 1382 degrees of freedom
Multiple R-squared: 0.6025, Adjusted R-squared: 0.6014
F-statistic: 523.8 on 4 and 1382 DF,  p-value: < 2.2e-16
```

Moran's I on the residuals

DO NOT USE MORAN.MC ON REDIUALS

```
> lm.morantest(soco_OLS, soco_nbq_w)

        Global Moran's I for regression residuals

data:
model: lm(formula = PPOV ~ PFHH + PUNEM + PBLK + P65UP,
data = soco)
weights: soco_nbq_w

Moran I statistic standard deviate = 22.5496, p-value <
2.2e-16
alternative hypothesis: greater
sample estimates:
Observed Moran's I        Expectation           Variance
      0.3608763174      -0.0020576309       0.0002590461
```

Spatial Lag Model (truncated output)

- Seems plausible that poverty might “spillover”.
- A spatial lag model will capture this spillover effect.

```
> soco_LAG <- lagsarlm(PPOV ~ PFHH + PUNEM + PBLK + P65UP, data=soco, soco_nbq_w)
> summary(soco_LAG)

Call:lagsarlm(formula = PPOV ~ PFHH + PUNEM + PBLK + P65UP, data = soco,     listw = soco_nbq_w)

Residuals:
       Min         1Q     Median         3Q        Max
-0.2457653 -0.0284360 -0.0028762  0.0262169  0.2374894

Type: lag
Coefficients: (asymptotic standard errors)
             Estimate Std. Error  z value  Pr(>|z|)
(Intercept) -0.100260   0.007375 -13.5946 < 2.2e-16
PFHH         0.429404   0.040246  10.6695 < 2.2e-16
PUNEM        1.354637   0.065959  20.5374 < 2.2e-16
PBLK        -0.069046   0.015335  -4.5025 6.716e-06
P65UP        0.291192   0.035210   8.2701 2.220e-16

Rho: 0.51719, LR test value: 491.48, p-value: < 2.22e-16
```

Reading these is not straight forward

Spillover effects

- The interconnected nature of spatial lag models makes it difficult to read the output from lagsarlm().
- A unit change in one predictor variable cascades throughout the system.
- To interpret the model we have to estimate the net effect of a unit change on Y in two places:
  - The location experiencing the change (**Direct Impacts**).
  - On the rest of the map (**Indirect Impacts**).
  - Together these constitute the **total impacts** of a change.

Interpretation of coefficients in a Spatial Lag model.

- Spatial Lag Models model encode spatial dependence.
- Coefficients do not have the same interpretation as in OLS.
- The β's reflect the direct impact of X on Y in a specific location.
- However, we also need to account for the indirect impact of x_i on y_i, from the influence yi exerts on its neighbors yj , which in turn feeds back into yi .
- Since each area has a different set of neighbors, the impact of a hypothetical change in xi will depend on which unit is being changed.
- A unit change in X may cause more than a Beta change in Y.
  - I think of this as similar to electricity, changes in one paces move through the weights matrix. I've heard econometricians refer to this as an equilibrium effect.

Example http://rpubs.com/Seth/LagEquibEff

- Fit a OLS and a spatial model, compare them:

```
> load("soco.rda")
> soco_LAG <- lagsarlm(PPOV ~ PFHH + PUNEM + PBLK + P65UP,
data = soco, soco_nbq_w)
> lm1 <- lm(PPOV ~ PFHH + PUNEM + PBLK + P65UP, data =
soco)
> anova(soco_LAG, lm1) #LRT Test
```

- Results indicate a significant difference between the models - **evidence that the spatial lag is warranted**.

Spatial Lag Model (full output)

```
> summary(soco_LAG)

Call:lagsarlm(formula = PPOV ~ PFHH + PUNEM + PBLK + P65UP, data = soco,     listw = soco_nbq_w)

Residuals:
       Min         1Q     Median         3Q        Max
-0.2457653 -0.0284360 -0.0028762  0.0262169  0.2374894

Type: lag
Coefficients: (asymptotic standard errors)
             Estimate Std. Error  z value  Pr(>|z|)
(Intercept) -0.100260   0.007375 -13.5946 < 2.2e-16
PFHH         0.429404   0.040246  10.6695 < 2.2e-16
PUNEM        1.354637   0.065959  20.5374 < 2.2e-16
PBLK        -0.069046   0.015335  -4.5025 6.716e-06
P65UP        0.291192   0.035210   8.2701 2.220e-16

Rho: 0.51719, LR test value: 491.48, p-value: < 2.22e-16
Asymptotic standard error: 0.02118
    z-value: 24.419, p-value: < 2.22e-16
Wald statistic: 596.28, p-value: < 2.22e-16

Log likelihood: 2238.884 for lag model
ML residual variance (sigma squared): 0.0021925, (sigma: 0.046825)
Number of observations: 1387
Number of parameters estimated: 7
AIC: -4463.8, (AIC for lm: -3974.3)
LM test for residual autocorrelation
test value: 4.8497, p-value: 0.027651
```

These are our coefficients.

These are tests for the significance of the spatial term

These are fit statistics.

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

A closer look at the output

- The coefficients are useless (but we'll fix that)
- The second block tells us about the spatial parameters.
  - `Rho: 0.51719` – this is the spatial effect.
  - `LR test value: 491.48, p-value: < 2.22e-16`

  - Is equivalent to a "leave one out" test of the spatial term:
    - `> anova(soco_LAG, soco_OLS) #LRT Test`

	Model	df	AIC	logLik	Test	L.Ratio	p-value
soco_LAG	1	7	-4463.8	2238.9	1		
soco_OLS	2	6	-3974.3	1993.1	2	491.48	0

- The Z-value on the next line is a test of the significance of the spatial term (Rho)
  - `z-value: 24.419, p-value: < 2.22e-16`
  - `Its computed:`
    - `Rho/"Asymptotic standard error" = .51719/.02118 = 24.419`

Interpreting Coefficients in a Spatial Model (from LeSage 2014)

- Its hard to directly interpret the coefficients due to spatial spillovers.
- Spillovers = Impacts passing through neighboring regions and back to the region itself.
- The magnitude of this type of feedback will depend upon
  - The position of the region in space (or in general in the connectivity structure)
  - The degree of connectivity among regions governed by the weight matrix W used in the model,
  - The parameter ρ measuring the strength of spatial dependence
  - The magnitude of the coefficient estimates for β and θ.

Understanding the lag…

- One way to understand the lag is to ask a question like:
  - **What would happen to poverty in the Southeast if the unemployment rate rose from 6% to 75% in Jefferson County, Alabama?**
- We can answer this question by modifying the data and then examining how the changes affect the predicted values.

```
#copy the data frame (so we don't mess up the
original)
soco.new <- soco

#Change the unemployment rate
soco.new@data[soco.new@data$CNTY_ST=="Jefferson
County  AL", "PUNEM"] <- .75
```

notice the @data

Predict using new data

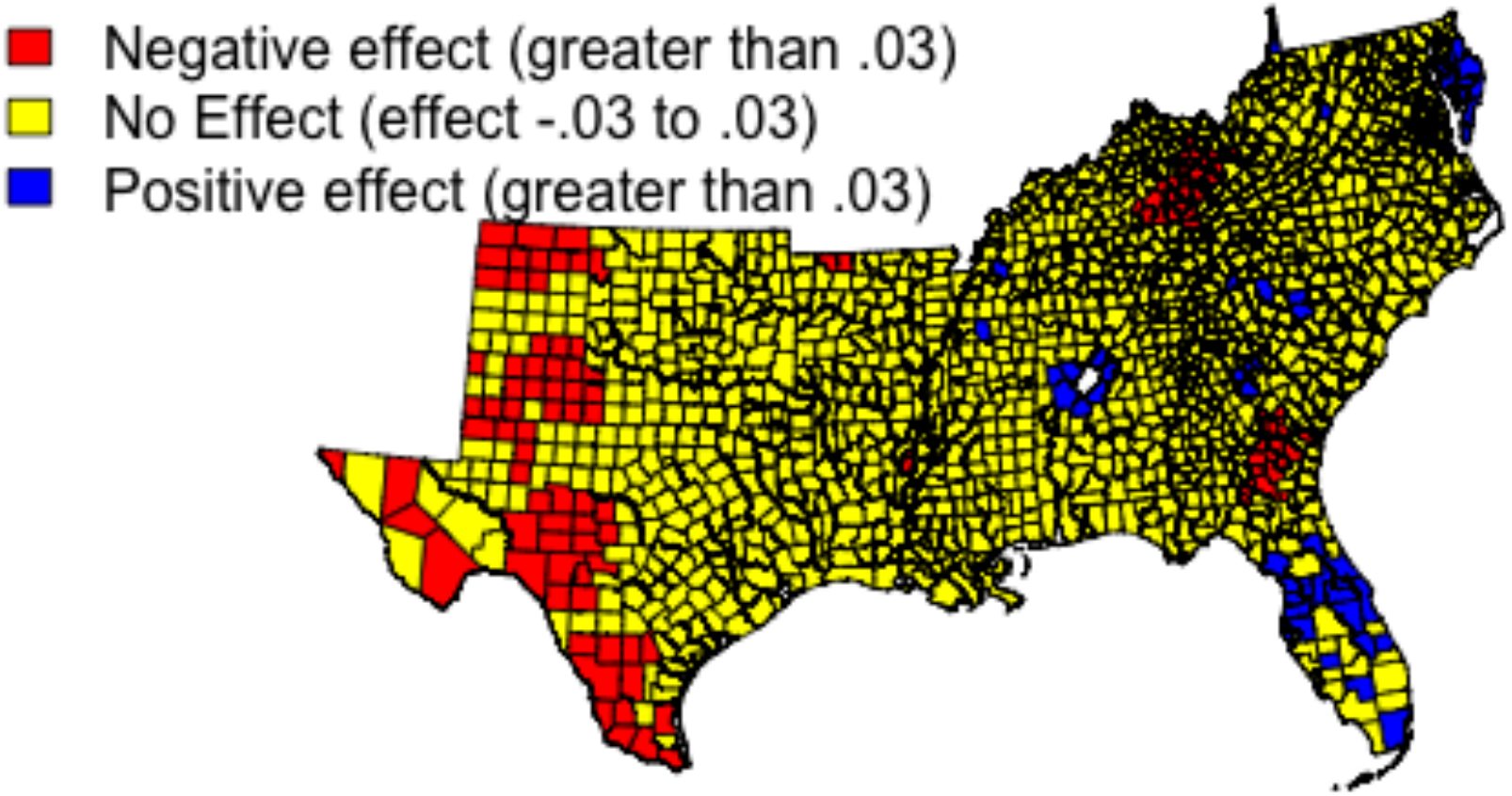
```
#The original predicted values
orig.pred <-
as.data.frame(predict(soco_LAG))

#The predicted values with the new
#unemployment rate in Alabama
new.pred <- as.data.frame(predict(soco_LAG,
newdata=soco.new, listw=soco_nbq_w))

#the difference between the predicted values
jCoAL_effect <- new.pred$fit-orig.pred$fit
```

Lag transmits changes…

```
#create a data frame to hold the county names and the effect of the changes to
Jefferson Cty, AL.
> el<- data.frame(name=soco$CNTY_ST, diff_in_pred_PPOV=jCoAL_effect)

#Also save the results into the spatial data frame so that we can map them.
> soco.new$jee <- el$diff_in_pred_PPOV

#sort counties by absolute value of the change in predicted PPOV
> el <- el[rev(order(abs(el$diff_in_pred_PPOV))),]
> el[1:10,] #show the top 10 counties

                    name      diff_in_pred_PPOV
37     Jefferson County  AL         0.99079188
4           Bibb County  AL         0.11636820
58     St. Clair County  AL         0.11165219
5         Blount County  AL         0.10881281
64        Walker County  AL         0.10738479
59        Shelby County  AL         0.10515699
63    Tuscaloosa County  AL         0.09640505
1050     El Paso County  TX        -0.08628607
1197      Sutton County  TX        -0.08107505
437          Lee County  KY        -0.07284785
```

KY is far away from AL!

Effects of a change in Jefferson County, AL (set PUNEM = .75) on predicted values in a spatial lag model



Readable coefficients?

```
> impacts(soco_LAG,listw=soco_nbq_w) #takes
a minute or two
Impact measures (lag, exact):
           Direct    Indirect       Total
PFHH    0.45695128  0.43243253  0.8893838
PUNEM   1.44153990  1.36419083  2.8057307
PBLK   -0.07347529 -0.06953281 -0.1430081
P65UP   0.30987229  0.29324539  0.6031177
```

Average Change - within area

Average Change - outside area

Average Change - Overall

Using "impacts" from LeSage 2008

- **Average direct effect:** averaged over all n regions/observations providing a summary measure of the impact arising from changes in the ith observation of variable r. For example, if region i increases the number of commuters who use public transportation, what will be the average impact on the commuting times in region i ? This measure will take into account feedback effects that arise from the change in the ith region's public transportation usage on commuting times of neighboring regions in the system of spatially dependent regions.
- The **Average Total effect= Average Direct effect + Average Indirect** effect. This scalar summary measure has two interpretations. Interpretation 1), if all regions raise public transportation usage, what will be the average total impact on commuting times of the typical region ? This total effect will include both the average direct impact plus the average indirect impact. Interpretation 2) measures the total cumulative impact arising from one region j raising its public transportation usage on commuting times of all other regions (on average).
- The **Average Indirect effect= Average Total effect - Average Direct** effect by definition. As an example, this effect could be used to measure the impact of all other regions raising their public transportation usage on the commuting times of an individual region, again averaged over all regions.

A closer look at the output

- There is one more interesting bit at the bottom of the lagsarlm() output:

```
LM test for residual autocorrelation
test value: 4.8497, p-value: 0.027651
```

- This LM test is telling us that we still have residual autocorrelation - we'll run a different a model in a bit…

- This could be due to other problems (like heteroscedacity).
  - Remember the BP test tries to model the variance in as a function of the predictors, a significant BP test indicates heteroscadacity i(which is bad).
  - The BP test tells us that our error term is heteroscedacitic.
  - In spatial models this is often due to varying population sizes in spatial units.

```
> bptest.sarlm(soco_LAG)

        studentized Breusch-Pagan test

data:
BP = 50.4553, df = 4, p-value = 2.901e-10
```

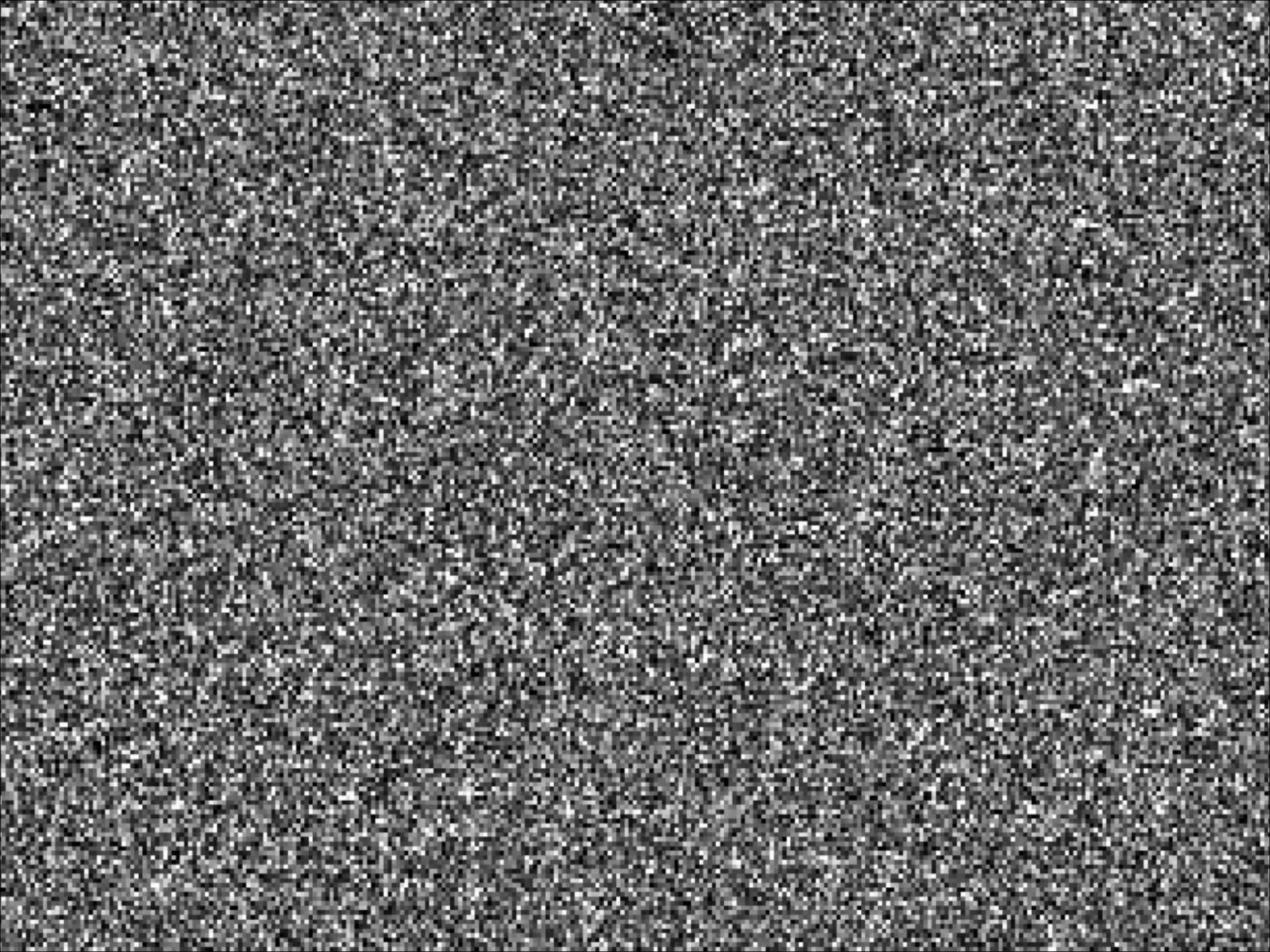
Spatial Lag Model Diagnostics

- Diagnostics of a spatial model are largely similar to OLS.
  - Residuals should be homoscedastic (BP-Test)
  - Residuals should have a mean of 0.
  - Residuals should be normally and independent.
- There are a few added wrinkles:
  - Circular logic makes it hard to understand the coefficients.
    - Maps of the “system” can help.
    - The impacts() function helps a lot!
  - We should check to see if the spatial term is significant (z-test and Wald test in the block with spatial output).
  - We should check to see if a spatial model is better than a non-spatial model (LR test).
  - We should check to see if there is significant residual autocorrelation (LM test)

RESIDUAL AUTOCORRELATION

Autocorrelation

- Auto or “self” correlation occurs when a variable observed at one location (in space or time) is similar to itself at a different location.
- Autocorrelation is what makes geography interesting.



Autocorrelated Residuals

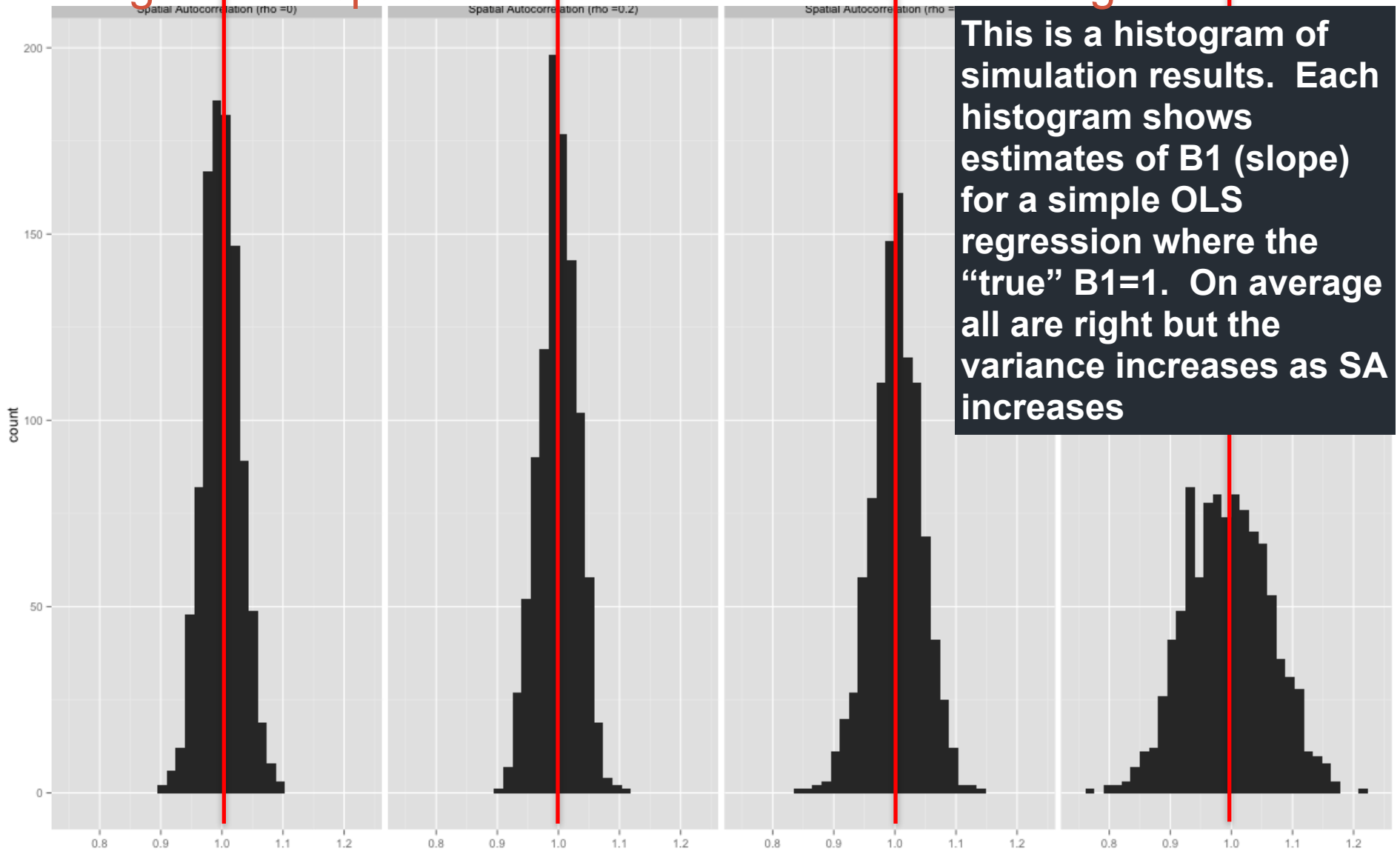
- In the previous spatial lag model we had autocorrelation in our residuals.
  - This inflates the variance in our estimated model coefficients.
- Autocorrelation makes pretty maps but bad regressions!!!
- Autocorrelation is not just a spatial problem, its also a temporal problem…

Why Are AC Errors Bad??

- When errors are autocorrelated the standard errors for our regression coefficients are underestimated.
  - This makes variables look like significant predictors (when they are not). It increases Type I (false positive) errors in inference about regression coefficients.
  - Confidence intervals for models coefficients get messed up.
  - Prediction gets messed up…
- HOWEVER, the estimated regression coefficients not biased.
  - But not having a good sense of the SE makes them hard to interpret and use.

Spatial dependence in the error.

Magnitude of spatial autocorrelation increases left to right



This is a histogram of simulation results. Each histogram shows estimates of B1 (slope) for a simple OLS regression where the "true" B1=1. On average all are right but the variance increases as SA increases

Autocorrelation

- For example, consider the following regression model:

$$Y_t = \beta_0 + \beta_1 X_t + \epsilon_t$$

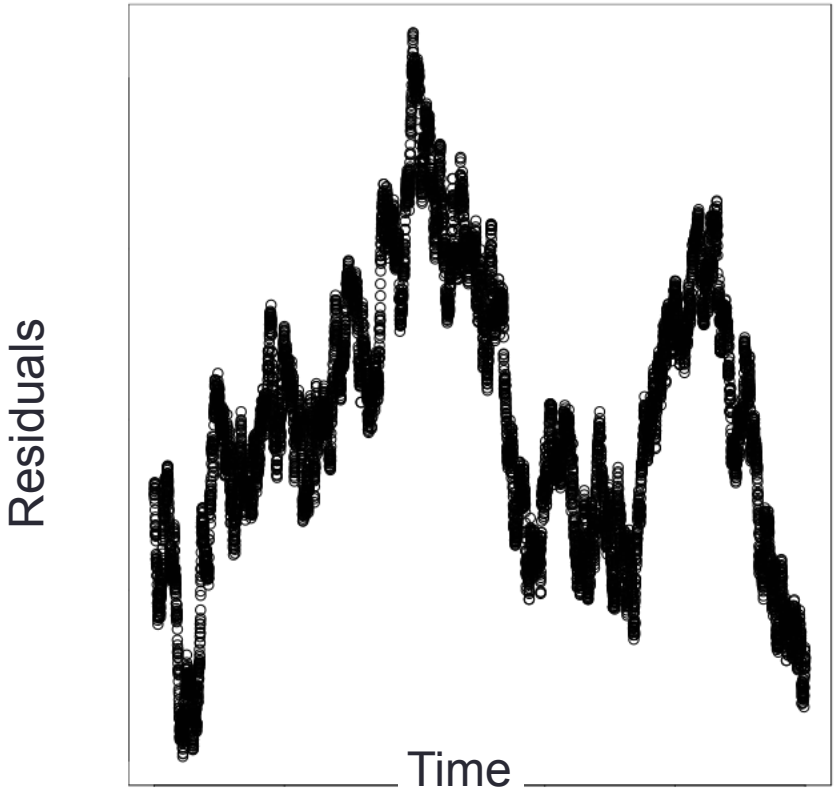
- Where Y_t is the predicted response at time t and ε_t is the error.
- Autocorrelation (AC) in the error term would mean that there errors in our model were systematic. A simple way to represent this:

$$\epsilon_t = \epsilon_{t-1} + u_t$$ **This is BAD!!!**

- The error at time t is a function of the previous error and a **random normal variable** u_t with a mean of 0 and a SD = σ.

Autocorrelated Residuals

This is BAD!



Residuals at time t are related to the residuals at time t+1

In the following examples we'll talk about time - which is similar to space only 1 dimensional. Its a line, not a map…

Error is always (partially) random

- When errors (residuals) are autocorrelated:
  - Error is partially systematic (it depends on what is happening nearby) and partially random.
  - The random part, in discussions of regression models, follows a normal distribution.

Correlated errors

- We can add an ***autocorrelation parameter ρ (rho)*** to the previous formulation of autocorrelated errors.

$$\epsilon_t = \rho\epsilon_{t-1} + u_t$$

- Rho measures the amount of autocorrelation.
- If rho=0 then we have a normal regression model with independent residuals that have a mean of zero.

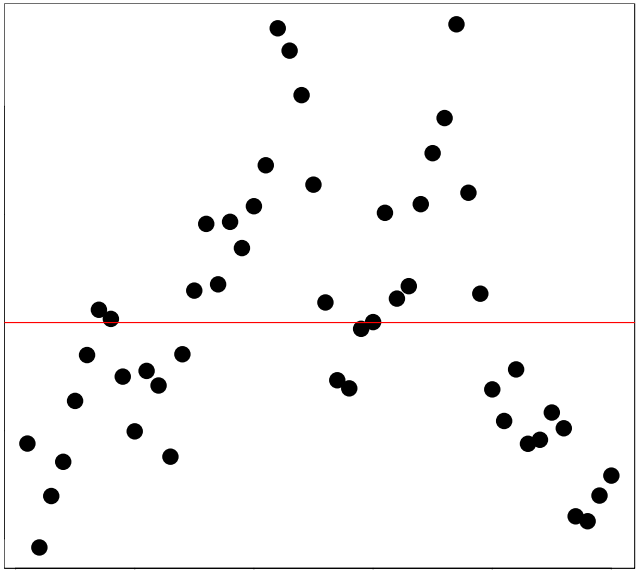
$$Y_t = \beta_0 + \beta_1 X_t + \epsilon_t$$

- If rho is non-zero we have a pattern in our residuals.

Durbin-Watson Test

- Durbin developed a hypothesis test to test hypotheses about rho. This is useful because it is sometimes hard to definitively interpret the residual plot.
- Might be nice to have a statistical test for autocorrelation in the residuals.

Are these residuals Autocorrelated?



Durbin-Watson Test

- The Durbin-Watson (DW) Test tests the null hypothesis that:

$$H_0 : \rho = 0$$

- The DW Test is calculated by fitting a model, taking the residuals, and then calculating:

$$D = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

Durbin-Watson Test

- If D is big it means that e_t and e_{t-1} are very different from each other. This is taken as evidence in support of the null hypothesis.

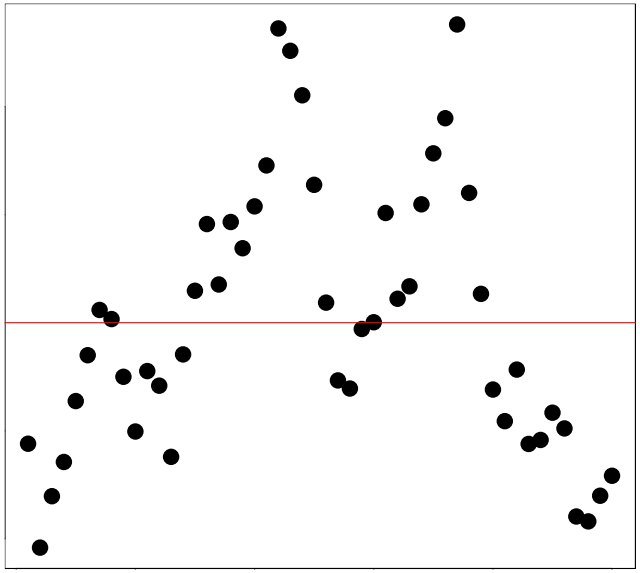
$$D = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

- If D is small it means that the adjacent errors are similar.
- Figuring out thresholds for D is complicated, D can be inconclusive, but R does some work for us (package lmtest)…

Durbin-Watson Test

```
> library(lmtest)
> lm1 <-lm(Y~X)
> dwtest(lm1)

        Durbin-Watson test

data:  lm1
DW = 0.1775, p-value < 2.2e-16
alternative hypothesis: true
autocorrelation is greater than 0
```

Are these residuals Autocorrelated?



Cochrane-Orcutt Procedure

- The Cochrane-Orcutt Procedure is designed to correct for autocorrelation in the residuals.
- This was our model for autocorrelated residuals.

$$\epsilon_t = \rho\epsilon_{t-1} + u_t$$

- We can view this as a regression equation.
- Where rho is the slope of the regression.
- The first step in the Cochrane-Orcutt Procedure is to **estimate the autocorrelation parameter via this regression.**
- We predict error based on the previous error. The strength of this relationship is our estimate of rho,

Cochrane-Orcutt Procedure

- We then remove the autocorrelation:

$$Y_t' = Y_t - \rho Y_{t-1}$$

$$X_t' = X_t - \rho X_{t-1}$$

$$\beta_0' = \beta_0(1-\rho)$$

$$\beta_1' = \beta_1$$

- And fit a new regression model:

$$Y_t' = \beta_0' + \beta_1' X_t' + \epsilon_t$$

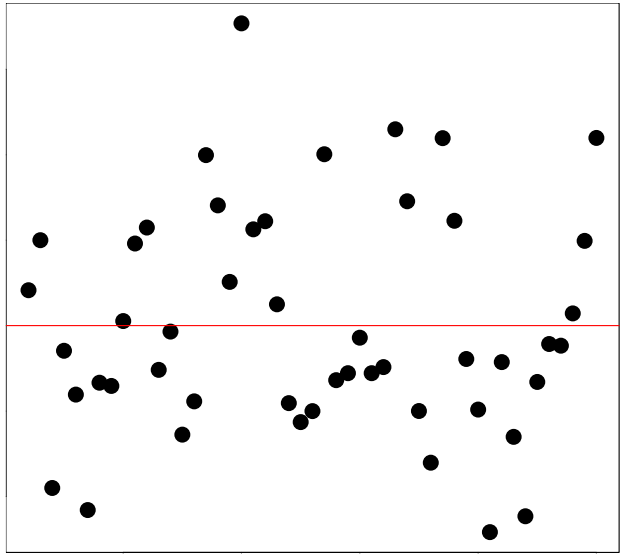
Cochrane-Orcutt Procedure

- The we check this model with the Durbin-Watson test. If we finds residual autocorrelation we repeat the process.
- Generally after a few iterations the the autocorrelation is gone.
- The final autocorrelation parameter (rho) tells us about the magnitude of the serial autocorrelation.

Cochrane-Orcutt Procedure

```
> lm1 <-lm(Y~X)
> lm2<-cochrane.orcutt(lm1)
> lm2
$Cochrane.Orcutt


Call:
lm(formula = YB ~ XB - 1)


Residuals:
    Min      1Q  Median      3Q     Max
-120.81  -44.27  -14.67   50.03  176.89


Coefficients:
             Estimate Std. Error t value Pr(>|t|)
XB(Intercept)  155.741    288.192   0.540    0.591
XBX             53.811      7.441   7.231 3.65e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 66.64 on 47 degrees of freedom
Multiple R-squared: 0.9052,   Adjusted R-squared: 0.9012
F-statistic: 224.5 on 2 and 47 DF,  p-value: < 2.2e-16
```

Are these residuals Autocorrelated?



$rho
[1] 0.9095355

```
$number.interaction
[1] 6
```

```
plot(y=lm2$Cochrane.Orcutt$residuals, x=2:50)
abline(h=0, col="red")
```

SPATIAL ERROR MODELS

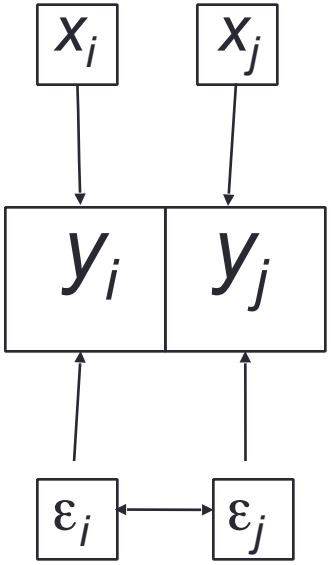
Applying the concept of autocorrelated errors to a spatial model.

More spatial…

- Spatial autocorrelation may be due to a misspecified model:
  - "…autocorrelation is often produced spuriously by model misspecification" (McMillen, 2003)
  - "What habitually underlies models with residual autocorrelation is a problem of misspecification of the equation due, in general, to the omission of relevant variables on the right-hand side" (López-Bazo and Fingleton, 2004)
- The spatial lag model is used when there is some substantive spatial process of interest.
- Often spatial data will have autocorrelated errors even when the there is no reason to suspect a spatial process.
  - Omitted variables (are a big problem).
  - Mismatch between data scale and process scale.
- In these cases we might consider a spatial error model

The spatial error model

- Examines spatial autocorrelation between the residuals of adjacent areas
- Treats spatial correlation primarily as a nuisance
  - Disregards the idea that spatial correlation may reflect some meaningful process
- Positive spatial error may reflect:
  - A misspecified model (particularly a omitted variable that is spatially clustered)
  - Incorrect spatial unit of aggregation



Spatial autocorrelation in residuals
Spatial error model

- Incorporates spatial effects through error term

$$y = X\beta + \epsilon$$

- Where:

$$\epsilon = \lambda W \epsilon + \mu$$

Spatial lag of residuals

Random error

- If there is no spatial correlation between the errors, then $\lambda = 0$ and we have a plain old regression model.

Cochrane-Orcutt Procedure

- Remember our friend, the Cochrane-Orcutt procedure.
- It is designed to correct for autocorrelation in the residuals.
- This was our model for autocorrelated residuals.

$$\epsilon_t = \rho\epsilon_{t-1} + u_t$$

- We can view this as a regression equation.
- Where rho is the slope of the regression.
- We predict error based on the previous error. The strength of this relationship is our estimate of rho (the autocorrelation parameter).
- This is a commonly used step in the estimation of spatial error models (e.g. Keleigian and Prucha 2099).

Running a spatial error model

```
> soco_err<-errorsarlm(PPOV ~ PFHH + PUNEM + PBLK + P65UP, data=soco, listw=soco_nbq_w)
> summary(soco_err)

Call:errorsarlm(formula = PPOV ~ PFHH + PUNEM + PBLK + P65UP, data = soco,     listw =
  soco_nbq_w)

Type: error
Coefficients: (asymptotic standard errors)
             Estimate Std. Error z value  Pr(>|z|)
(Intercept) 0.0070874  0.0095400  0.7429    0.4575
PFHH        0.3649742  0.0405868  8.9924 < 2.2e-16
PUNEM       1.1472020  0.0711971 16.1131 < 2.2e-16
PBLK        0.0849301  0.0206069  4.1214 3.765e-05
P65UP       0.3516854  0.0431300  8.1541 4.441e-16

Lambda: 0.74655, LR test value: 526.26, p-value: < 2.22e-16
Asymptotic standard error: 0.021739
    z-value: 34.342, p-value: < 2.22e-16
Wald statistic: 1179.3, p-value: < 2.22e-16

Log likelihood: 2256.274 for error model
ML residual variance (sigma squared): 0.0019718, (sigma: 0.044405)
Number of observations: 1387
Number of parameters estimated: 7
AIC: -4498.5, (AIC for lm: -3974.3)
```

These are our coefficients.

These are tests for the significance of the spatial error term

These are fit statistics.

Hypothesis tests for spatial models

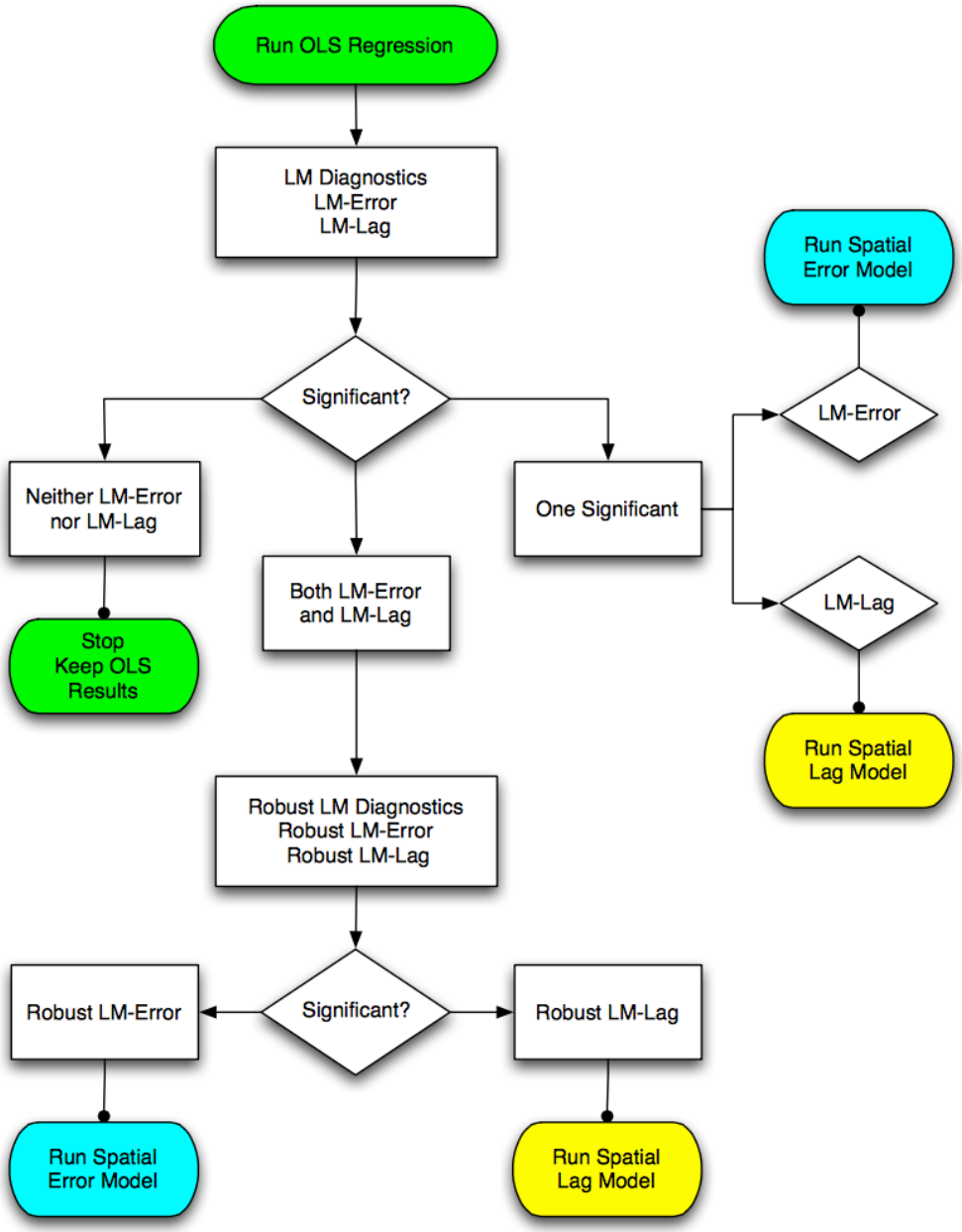
- Wald Test – “Wald statistic”
  - Like the t-test. Have to estimate a full and reduced model. Hypothesis test on the difference between a model with and without the lag/error term. If significant the spatial term improves the model.
- Likelihood Ratio Test – “LR test”
  - Equivalent to anova(model1, model 2). Tests the signifigance of the spatial term. If significant the spatial term improves the model.
- Lagrange Multiplier Test – “LM test for residual autocorrelation”
  - Default output in lag model only.
  - Useful for model selection (spatial error, spatial lag, or non-spatial OLS).

Two Model Selection Methods

- There is a forward “Anselin Style” model selection strategy:
  - Fit an OLS model
  - Test for spatial autocorrelation. Base model selection on diagnostic tests.
  - If autocoreelation is found use the spatial lag or error model (as indicated by theory or model selection strategies).
- A spatial Durbin model.
  - Spatial model fitting strategies are described in detail by Florax et. al (2003)

Anselin Style Model Selection

- Which of the two spatial models should we fit?
- Anselin provides a decision tree based on Lagrange Multiplier (LM) tests.
  - Theory of LM tests is complicated, involves the derivatives of the likelihood function (I think?).


Run OLS Regression
LM Diagnostics
LM-Error
LM-Lag
Significant?
Neither LM-Error
nor LM-Lag
Stop
Keep OLS
Results
Both LM-Error
and LM-Lag
One Significant
LM-Error
Run Spatial
Error Model
LM-Lag
Run Spatial
Lag Model
Robust LM Diagnostics
Robust LM-Error
Robust LM-Lag
Significant?
Robust LM-Error
Robust LM-Lag
Run Spatial
Error Model
Run Spatial
Lag Model

Spatial Regression Analysis (recipe inspired by Ward)

- Map the data, especially the dependent variable (outcome variable, Y).
- Check the dependent variable for spatial autocorrelation.
  - Calculate Moran's I.
  - Make the LISA map.
  - See if these suggest missing variables. You should have variables that (largely) explain the observed patterns in Y. For example, regional clusters might be a reason to add a regional dummy variable.
- Spatial regression.
  - Use theory, personal preference, or model selection criteria (like the Anselin Decision Tree) to choose a model.
- Fully explain results. If you fit a lag model be sure to explore direct and indirect effects.

Determining the type of dependence

```
> summary(lm.LMtests(soco_OLS, listw=soco_nbq_w, test="all"))
  Lagrange multiplier diagnostics for spatial dependence
data:
model: lm(formula = PPOV ~ PFHH + PUNEM + PBLK + P65UP, data =
  soco)
weights: soco_nbq_w

       statistic parameter   p.value
LMerr    497.333         1 < 2.2e-16 ***
LMlag    571.276         1 < 2.2e-16 ***
RLMerr    68.174         1 < 2.2e-16 ***
RLMlag   142.116         1 < 2.2e-16 ***
SARMA    639.449         2 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual Autocorrelation

```
> moran.test(resid(soco_LAG), soco_nbq_w, randomisation=FALSE)

          Moran's I test under normality

Moran I statistic standard deviate = 1.7276, p-value = 0.04203
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
     0.0271772045     -0.0007215007      0.0002607895

> moran.test(resid(soco_err), soco_nbq_w, randomisation=FALSE)

          Moran's I test under normality

Moran I statistic standard deviate = -3.7067, p-value = 0.9999
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
    -0.0605816480     -0.0007215007      0.0002607895
```

Spatial Durbin Model

- A spatial Durbin Model includes lags all the X variables and the Y (as specified in LeSage 2009).
- Hard to interpret.
  - Logic is highly circular.
    - The X's affect the Y's at multiple locations.
- Run it by specifying `type="mixed"` in `lagsarlm()`

$$Y = \rho W y + \beta_1 X_1 + ... + \beta_n X_n + \theta_1 W X_1 + ... + \theta_n W X_n + \epsilon$$

Spatial Durbin Model Output

```
> socDb <- lagsarlm(PPOV ~ PFHH + PUNEM + PBLK + P65UP, data=soco, listw=soco_nbq_w, type="mixed”)
> summary(socDb)

Call:
lagsarlm(formula = PPOV ~ PFHH + PUNEM + PBLK + P65UP, data = soco,
    listw = soco_nbq_w, type = "mixed")

Residuals:
       Min         1Q     Median         3Q        Max
-0.1633684 -0.0276632 -0.0032547  0.0233727  0.2235147

Type: mixed
Coefficients: (asymptotic standard errors)
             Estimate Std. Error z value  Pr(>|z|)
(Intercept) -0.052496   0.010990 -4.7769 1.780e-06
PFHH         0.357979   0.040384  8.8644 < 2.2e-16
PUNEM        1.082361   0.071536 15.1304 < 2.2e-16
PBLK         0.111305   0.021657  5.1394 2.757e-07
P65UP        0.325358   0.043521  7.4758 7.683e-14
lag.PFHH    -0.173054   0.072126 -2.3993   0.01643
lag.PUNEM    0.285779   0.127931  2.2338   0.02549
lag.PBLK    -0.161000   0.029298 -5.4952 3.903e-08
lag.P65UP   -0.151350   0.066045 -2.2916   0.02193

Rho: 0.60048, LR test value: 377, p-value: < 2.22e-16
Asymptotic standard error: 0.027692
    z-value: 21.684, p-value: < 2.22e-16
Wald statistic: 470.2, p-value: < 2.22e-16

Log likelihood: 2321.321 for mixed model
ML residual variance (sigma squared): 0.0019018, (sigma: 0.04361)
Number of observations: 1387
Number of parameters estimated: 11
AIC: -4620.6, (AIC for lm: -4245.6)
LM test for residual autocorrelation
test value: 29.729, p-value: 4.9686e-08
```

Evaluating Spatial Durbin Model

- The Durbin model can be compared to a spatial error/lag model using `anova()` because the models are nested.

```
> anova(soco_err, socDb)
         Model df      AIC logLik Test L.Ratio p-value
soco_err     1  7 -4498.5 2256.3    1
socDb        2 11 -4620.6 2321.3    2  130.09       0
```

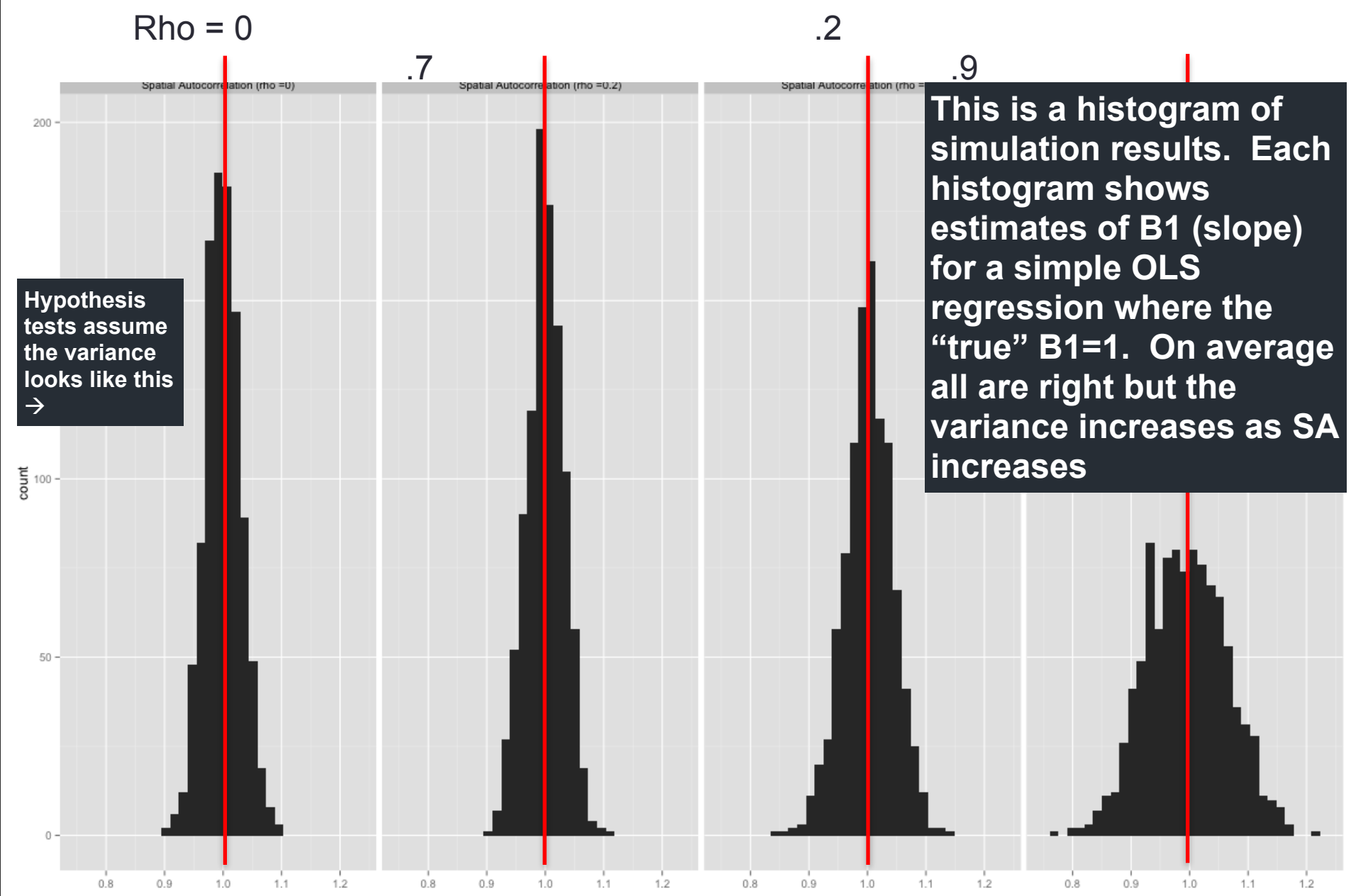
- This LRT indicates that the Durbin model exlains signifcantly more variation than the error model.
  - However, its really hard to interpret…

Why do we need spatial models…

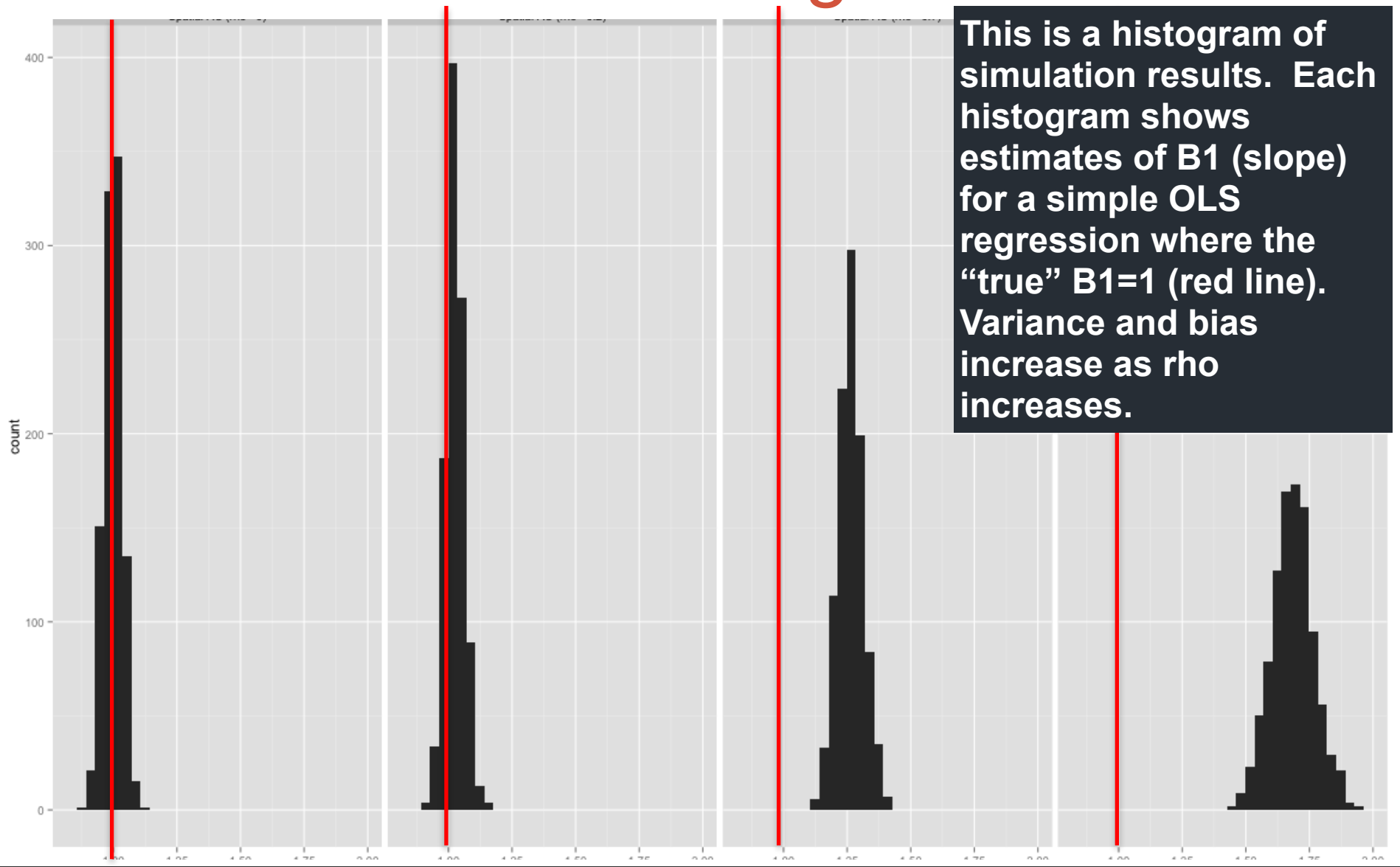
- Spatial autocorrelation does nasty things to regression coefficients.
- If there is a spatial lag process going on (neighbors influencing neighbors) and we fit an OLS model our coefficients will be **biased and inefficient**.
- If there is a spatial error process OLS coefficients are **unbiased but inefficient**, which means over many replications, on average they will work out to be correct but spatial autocorrelation increases the run-to-run variance.
  - In an individual study it high SE mean you are more likely to be far from the "truth" (but you'd never know it….)

Using simulation to gain insight into the consequences of spatial autocorrelation.

- SEE http://rpubs.com/Seth/spatial_autocorrelation_and_OLS
- Do we really need spatial models? What are the consequences of not using them?
- To answer this question:
  - Create a data set consisting of an X and a Y with a known level of spatial dependence.
  - Place the spatial dependence either in the error (errors are random but correlated with neighbors) or in the Y itself (the value of your neighbors partially determines your value).
  - During the simulation of Y set the beta (slope) coefficient to 1. The “true” relationship between Y and X is Y= a + bx + e, where b=1.
  - Estimate the relationship between Y and X using OLS (lm() in R).
  - Save the the estimate of b, if its different from 1, it is incorrect…
  - Repeat the process 1000 times for each level of spatial autocorrelation and each type of spatial process (error, lag).


Rho = 0
.2
.7
.9
Spatial Autocorrelation (rho =0)
Spatial Autocorrelation (rho =0.2)
Hypothesis tests assume the variance looks like this →
This is a histogram of simulation results. Each histogram shows estimates of B1 (slope) for a simple OLS regression where the "true" B1=1. On average all are right but the variance increases as SA increases
count
200
100
50
0
0.8
0.9
1.0
1.1
1.2

Rho increases left to right.



This is a histogram of simulation results. Each histogram shows estimates of B1 (slope) for a simple OLS regression where the "true" B1=1 (red line). Variance and bias increase as rho increases.

EXPLORING SPATIAL HETEROGENEITY

Spatial Heterogeneity

- Regression in general, and the spatial error and lag model in particular, assume that fitted regression model applies to entire study area.
- Often we are interested in spatially varying relationships, that is we want to understand how relationships vary in space.
- One of the simplest ways to do this is fit a model with regional dummy variables.
- Other, more sophisticated strategies include:
  - Spatial Regimes model (Chow Test)
  - Spatial Expansion Method
  - Geographically Weighted Regression

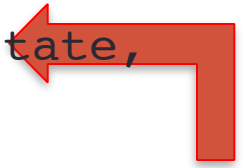
Spatial Regimes Model

- Used to assess the null hypothesis that model coefficients are the same across regions.
  - Sometimes referred to as “structural break”
  - When you cross the region boundary do relationships significantly change?
- Fit a model with and without a region dummy.
- Test if coefficients vary by region using the Chow Test.
- Can be run for spatial and non-spatial regression models.
- R code for Chow test is by Anselin (2007, not incorporated into a package)

OLS Regimes

```
>soco$state <- as.factor(soco$STUSAB)
>soco_OLS <-
       lm(PPOV ~ PFHH + PUNEM + PBLK + P65UP,
       data=soco)

>soco_OLS_R <-
       lm(PPOV ~ 0 + (PFHH + PUNEM + PBLK + P65UP) : state,
       data=soco)
```

Regression w/ o an intercept



State specific coefs for all variables

Regimes models

- Nice because they're a simple way to test for regional variability.
- You need to know the regions in advance.
- Can be applied to spatial autoregressive models
- If "pchow" is significant you have evidence of "regimes" different models for different regions.
- See Anselin, Luc. 2007. "Discrete Spatial Heterogeneity" & "Continuous Spatial Heterogeneity." *Spatial Regression Analysis in R: A Workbook*. For a spatial version of the chow test.

```
chow.test <- function(rest,unrest)
{
  er <- residuals(rest)
  eu <- residuals(unrest)
  er2 <- sum(er^2)
  eu2 <- sum(eu^2)
  k <- rest$rank
  n2k <- rest$df.residual - k
  c <- ((er2 - eu2)/k) / (eu2 / n2k)
  pc <- pf(c,k,n2k,lower.tail=FALSE)
  list(c,pc,k,n2k)
}
```

```
> chow.test(soco_OLS,
soco_OLS_R)
[[1]]
[1] 103.1003

[[2]]
[1] 1.607957e-92

[[3]]
[1] 5

[[4]]
[1] 1377
```

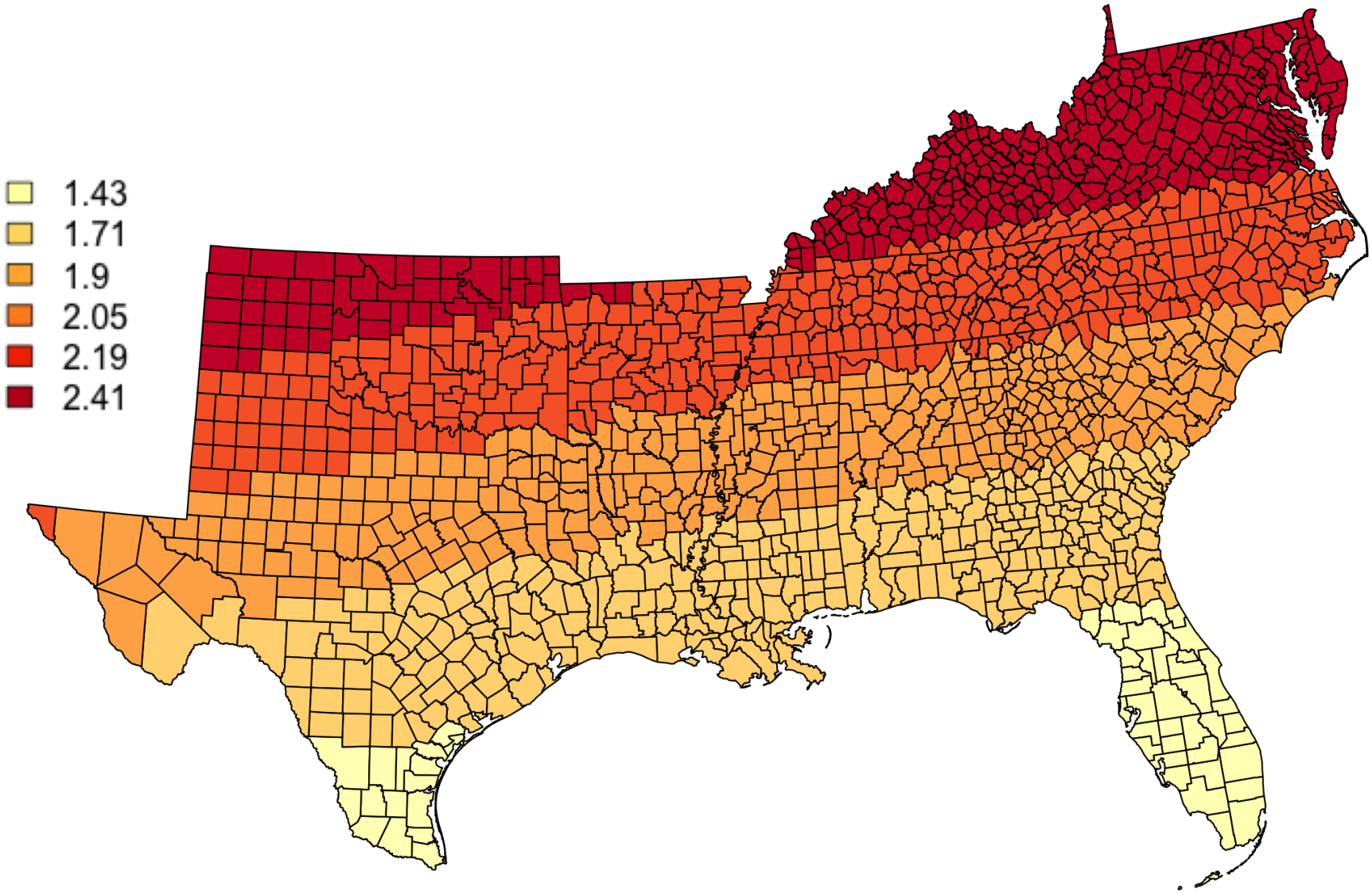
Spatial Expansion Method

- The chow test tells us if there are significant regional differences.
- The expansion method allows regression coefficients to vary as a function of the X and Y coordinates (method usually attributed to Casetti 1972, Jones + Casetti 1986).
- Estimating the expansion model is simple:
  - Add interactions between each variable and the X and Y coordinates.
  - Then combine the expanded coefficients to get local, place specific coefficients.
  - Shows how the influence of variables changes in space.

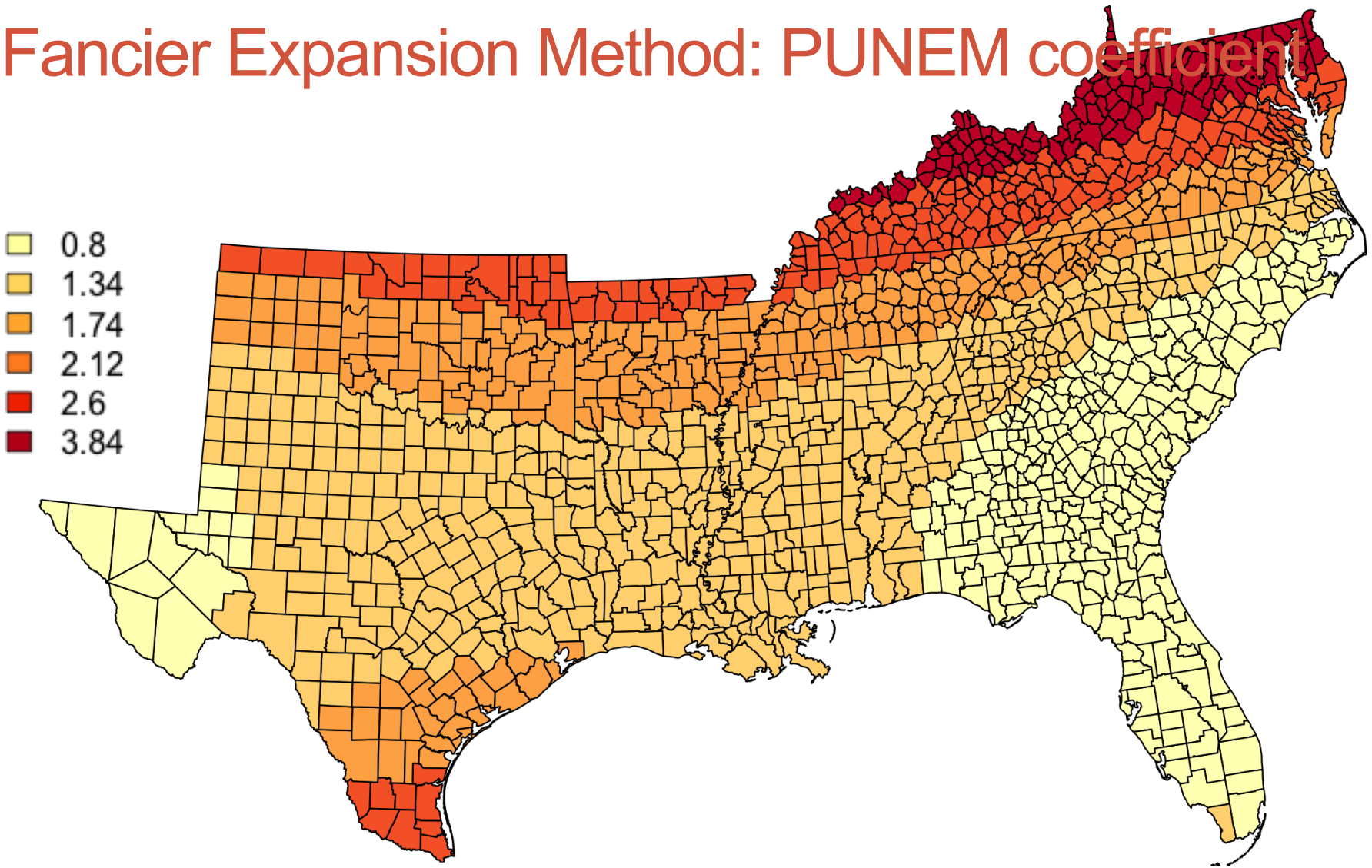
Expansion Method

```
##EXPANSION METHOD
soco_ex <- lm(PPOV ~ (PFHH + PUNEM + PBLK + P65UP) * (XCOORD + YCOORD), data=soco)
b <- soco_ex$coefficients

##This is the PUNEM coefficient for each county
soco$bPUNEM <- b[3] + b[10] * soco$XCOORD + b[11] * soco$YCOORD

#Map the influence on PUNEM
pal5 <- brewer.pal(6, "YlOrRd")
cats5 <- classIntervals(soco$bPUNEM, n = 5, style = "jenks")
colors5 <- findColours(cats5, pal5)
plot(soco, col=colors5)
legend("topleft", legend=round(cats5$brks,2), fill=pal5, bty="n")
mtext("Expansion Model: PUNEM", side=3, line=1)
```

Expansion Method: PUNEM coefficient



Fancier Expansion Method: PUNEM coefficient



- soco_ex2 <- lm(PPOV ~ (PFHH + PUNEM + PBLK + P65UP) * (XCOORD + YCOORD + **I(XCOORD^2) + I(YCOORD^2)**), data=soco)

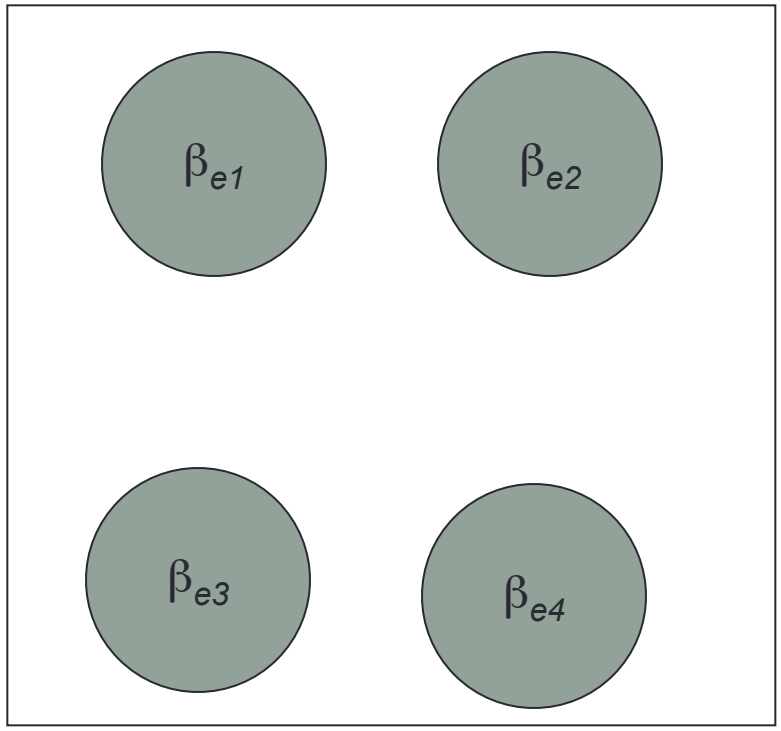
Geographically Weighted Regression

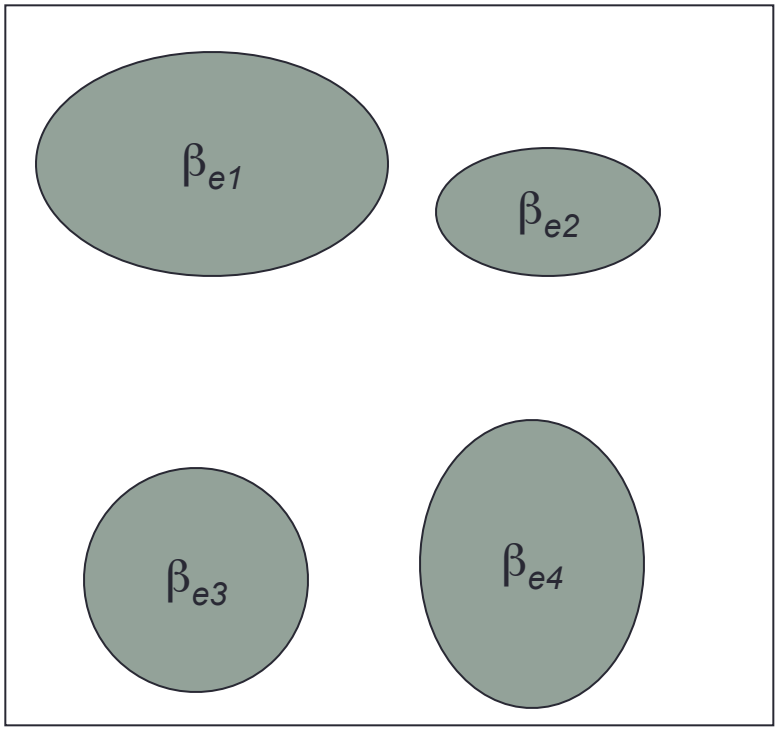
- When the same stimulus provokes a different response in different parts of the study region.
- Why do relationships vary spatially?
  - Nuisance variation (from sampling, population size, etc)
  - Relationships intrinsically different across space (Place matters!)

Local vs Global Models

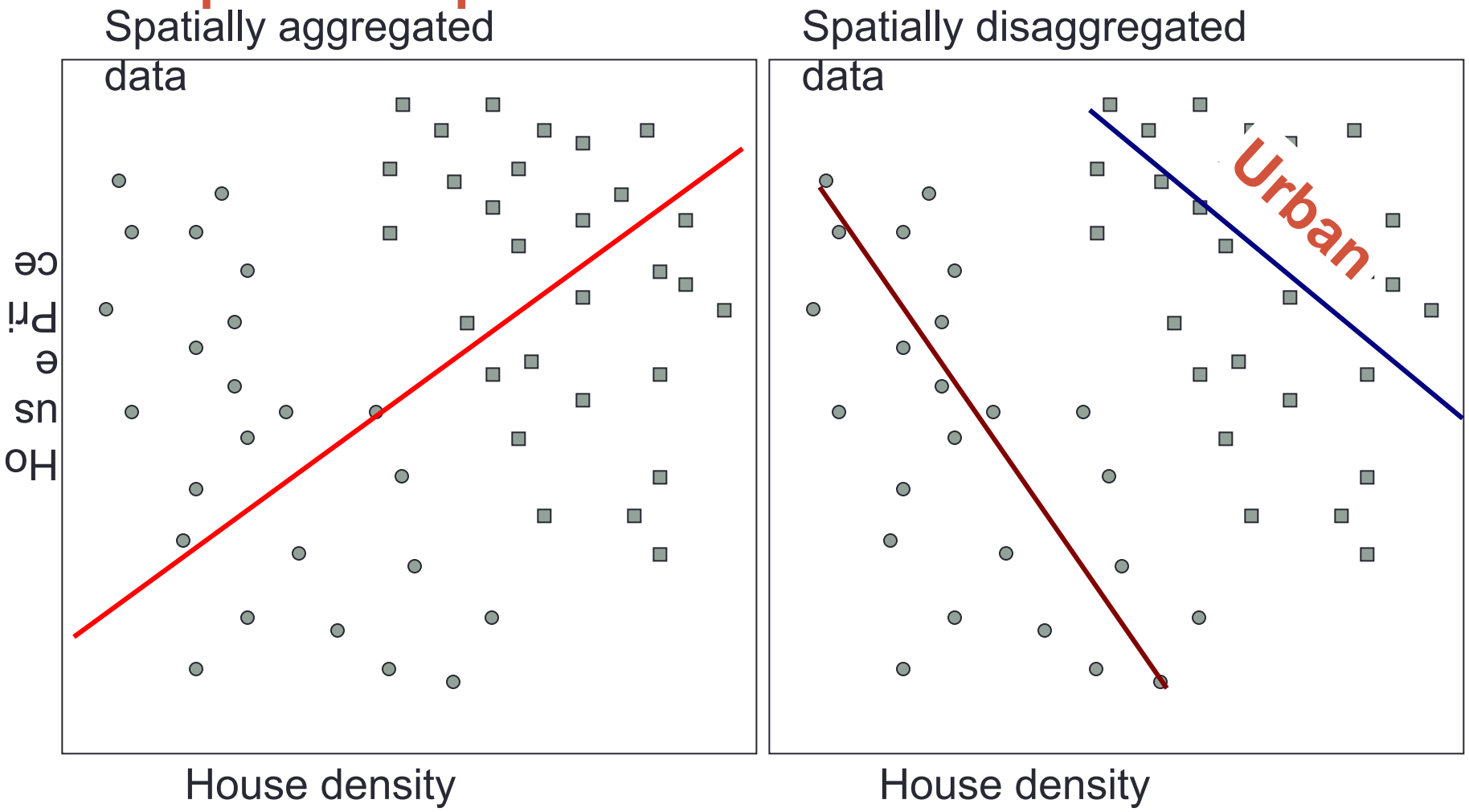
$$y_i = \beta_0 + \beta_1 x_{1i}$$

$$y_i = \beta_{i0} + \beta_{i1} x_{1i}$$





Simpson's paradox



Regression

- In a typical linear regression model applied to spatial data we assume coefficients (β) the same across all locations *i:*

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

- When applied to spatial data, this model suggests that the same stimulus provokes the same response in all parts of the study region…
  - The model allows for random error but not substantive, place based differences.

Geographically Weighted Regression

- Local statistical technique to analyze spatial variations in relationships
- Spatially varying relationships are assumed.
- Fit dozens of models that allows the relationships to vary over space
  - i.e., βs do not need to be the same everywhere
  - continuous surface of parameter values

GWR

- GWR is:

$$y_i = \beta_{i0} + \beta_{i1} x_{1i} + \beta_{i2} x_{2i} + \ldots + \beta_{in} x_{ni} + \varepsilon_i$$

  - where i is now the location at which estimates of the coefficients (β) are obtained

- Instead of remaining the same everywhere, βs now vary in terms of locations (*i*)

Calibration of GWR

- Coefficients are estimated using nearby observations
  - This means we create a weights matrix similar to what we did in spatial regression!
  - Local weights matrix vs. global weights matrix
- Create a matrix of weights specific to each location i such that observations nearer to i are given greater weight than observations further away.
  - To allow for this distance decay neighborhood is not created using adjacency or k nearest neighbors.
  - Instead a kernel is applied…
- The kernel can be a fixed shape or an adaptive shape.

Some weight functions

- Fixed discrete:

  w_{ij} = 1 if d_{ij} <= d (some distance)

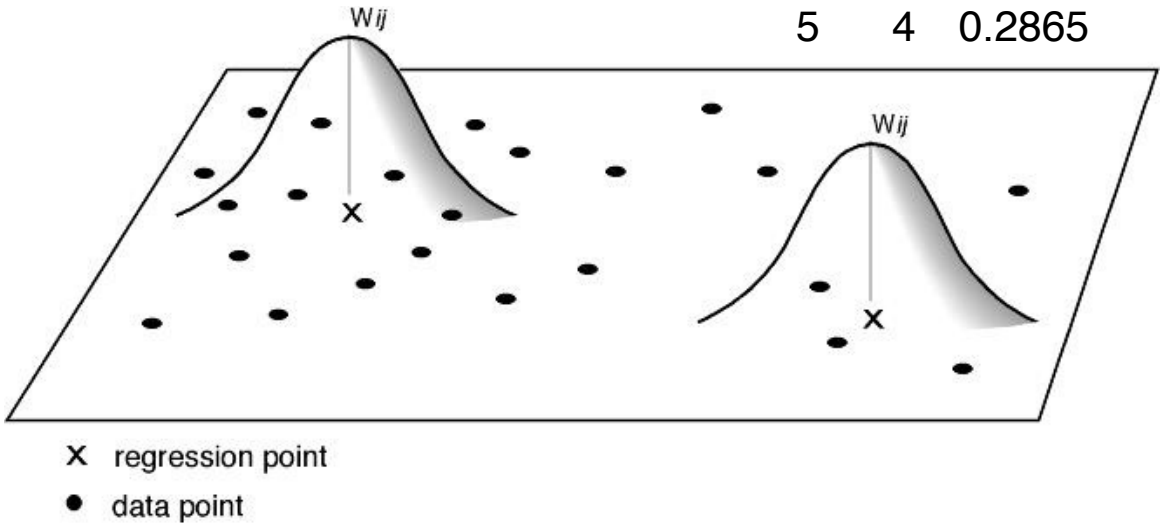
  w_{ij} = 0 otherwise

- Fixed continuous:

  $w_{ij} = \exp(-d_{ij}^2/h^2)$

  where h is known as the bandwidth and controls the degree of distance-decay

Fixed		Fixed continuous		
d(ij)	w(ij)	d(ij)	h	w(ij)
1	1	1	2	0.6065
2	1	2	2	0.3679
3	1	3	2	0.2231
4	0	4	2	0.1353
5	0	5	2	0.0821
d=3		1	4	0.7788
		2	4	0.6065
		3	4	0.4724
		4	4	0.3679
		5	4	0.2865



Problems of fixed schemes

- Might produce large estimate variances where data are sparse, while mask subtle local variations where data are dense
- In extreme condition, fixed schemes might not be able to calibrate in local areas where data are too sparse to satisfy the calibration requirements

Some weight functions

- Adaptive continuous:

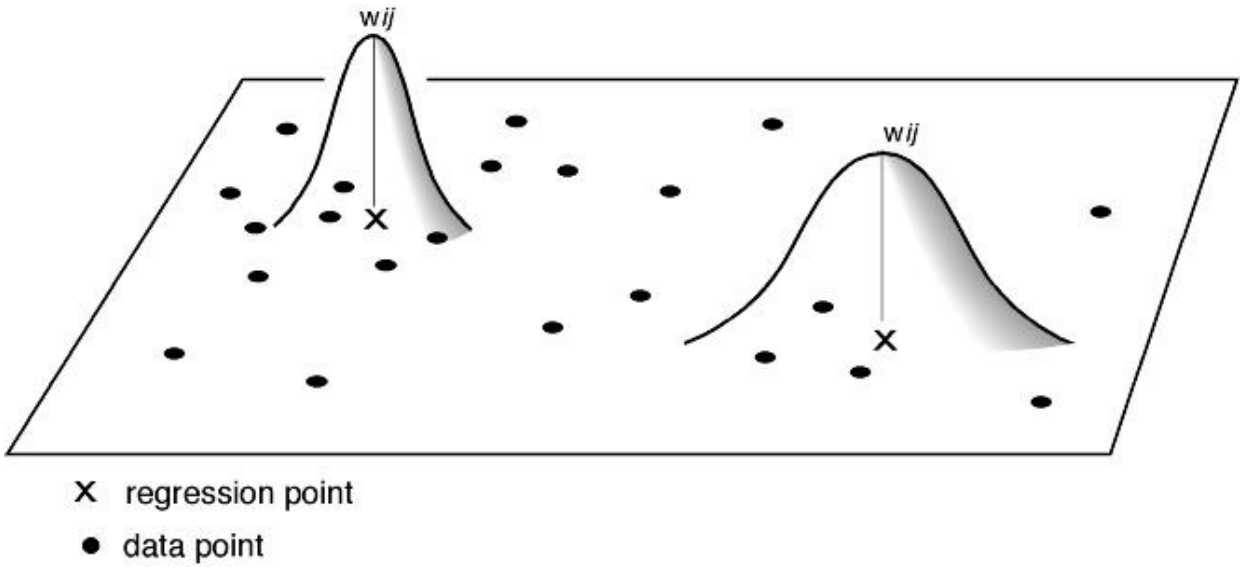
$w_{ij} = (1 - (d_{ij}/h)^2)^2$

if j is one of the n^{th} nearest neighbors of I OR

if $d_{ij} < h$

$w_{ij} = 0$ otherwise

if d(ij) < h			if j is one of 2 NNs			
d(ij)	h	w(ij)	d(ij)	h	NN	w(ij)
1	4	0.878	1	4	Y	0.878
2	4	0.562	2	4	Y	0.562
3	4	0.191	3	4	N	0.000
4	4	0.000	4	4	N	0.000
5	4	0.000	5	4	N	0.000



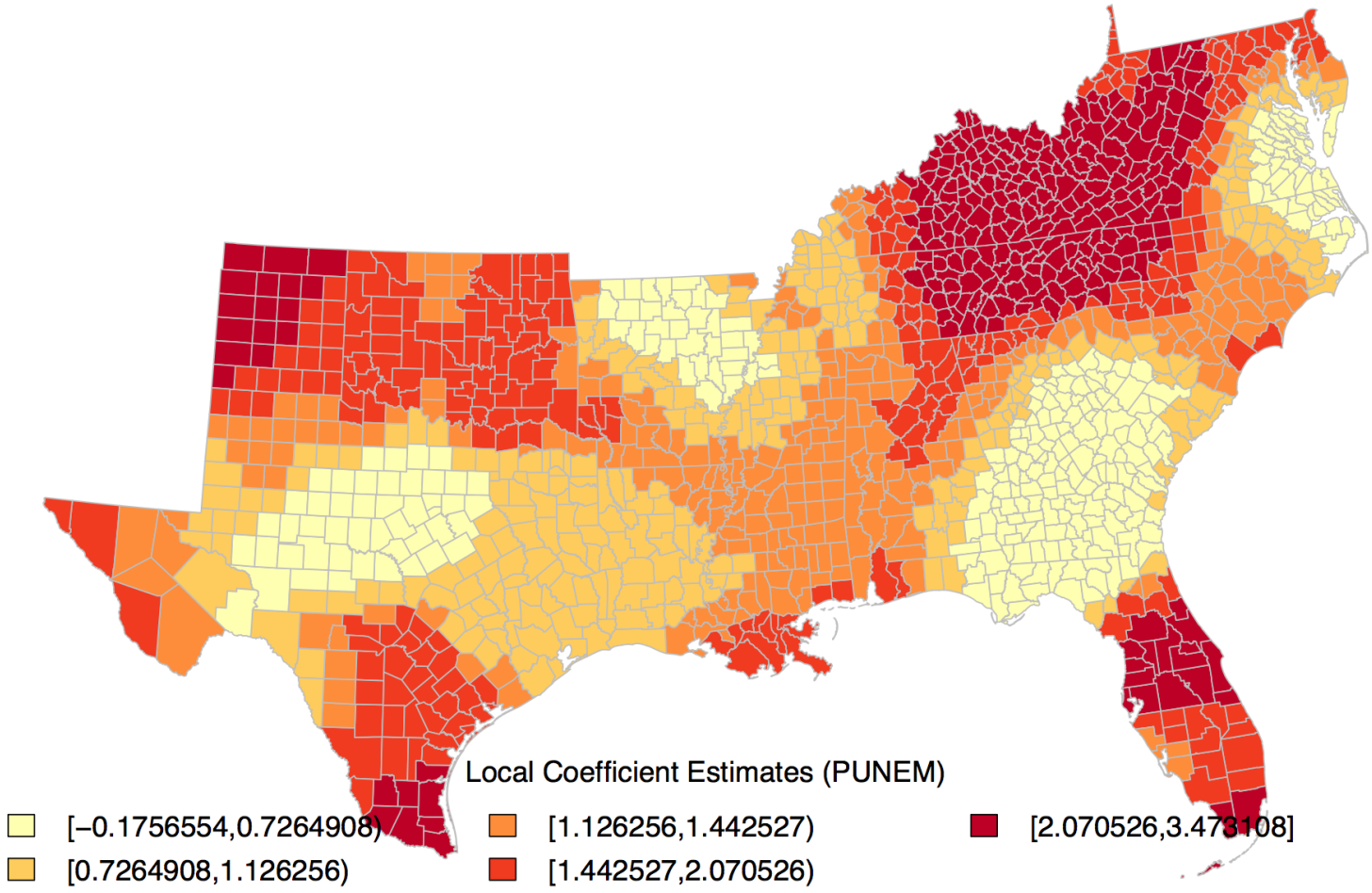
Adaptive weighting schemes

- Adaptive schemes adjust itself according to the density of data
  - Shorter bandwidths where data are dense and longer where sparse
  - Finding nearest neighbors are one of the often used approaches

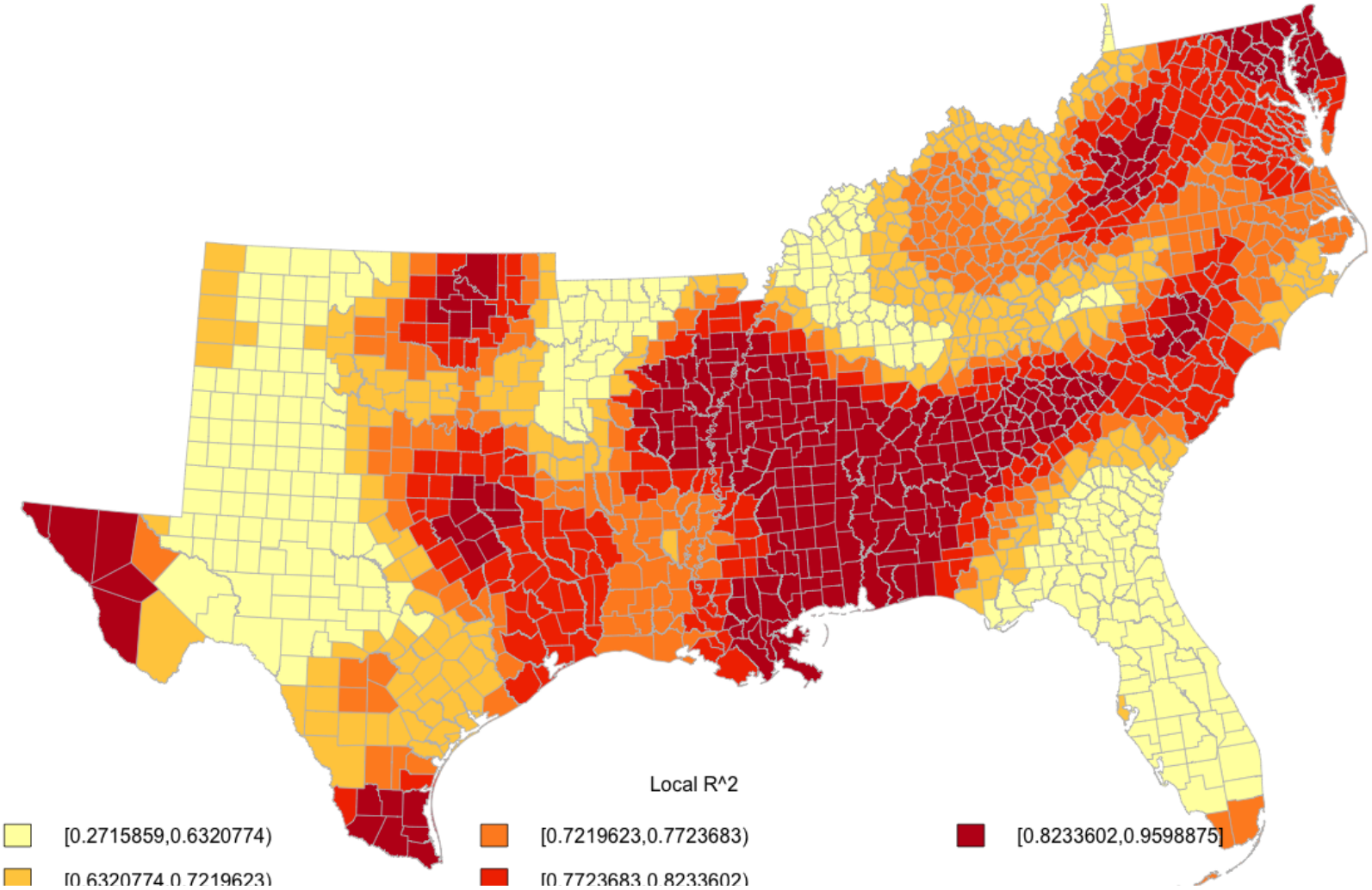
Bandwidth Selection

- Results are sensitive to the degree of distance-decay
- An optimal value of either h or n has to be obtained
- 3 methods to set the bandwidth
  - Manually
  - Automagically in R: Cross-validation (CV) minimization(gwr.sel())
    - The difference between observed value and the GWR calibrated value using the bandwidth or nearest neighbors
  - Akaike Information Criterion (AIC) minimization

GWR Coefficient Estimates (PUNEM)



GWR Local R-squared



Criticisms of GWR

- The usual regression diagnostics are ignored
  - Local multicollinearities
- Wheeler & Tiefelsdorf (2005) note:
  - “correlation of local regression coefficients potentially invalidates any interpretation of individual GWR parameter estimates and can facilitate misleading conclusions if the situation is not properly diagnosed.”
- Models should be checked for correlation among regression coefficients.

David Wheeler · Michael Tiefelsdorf

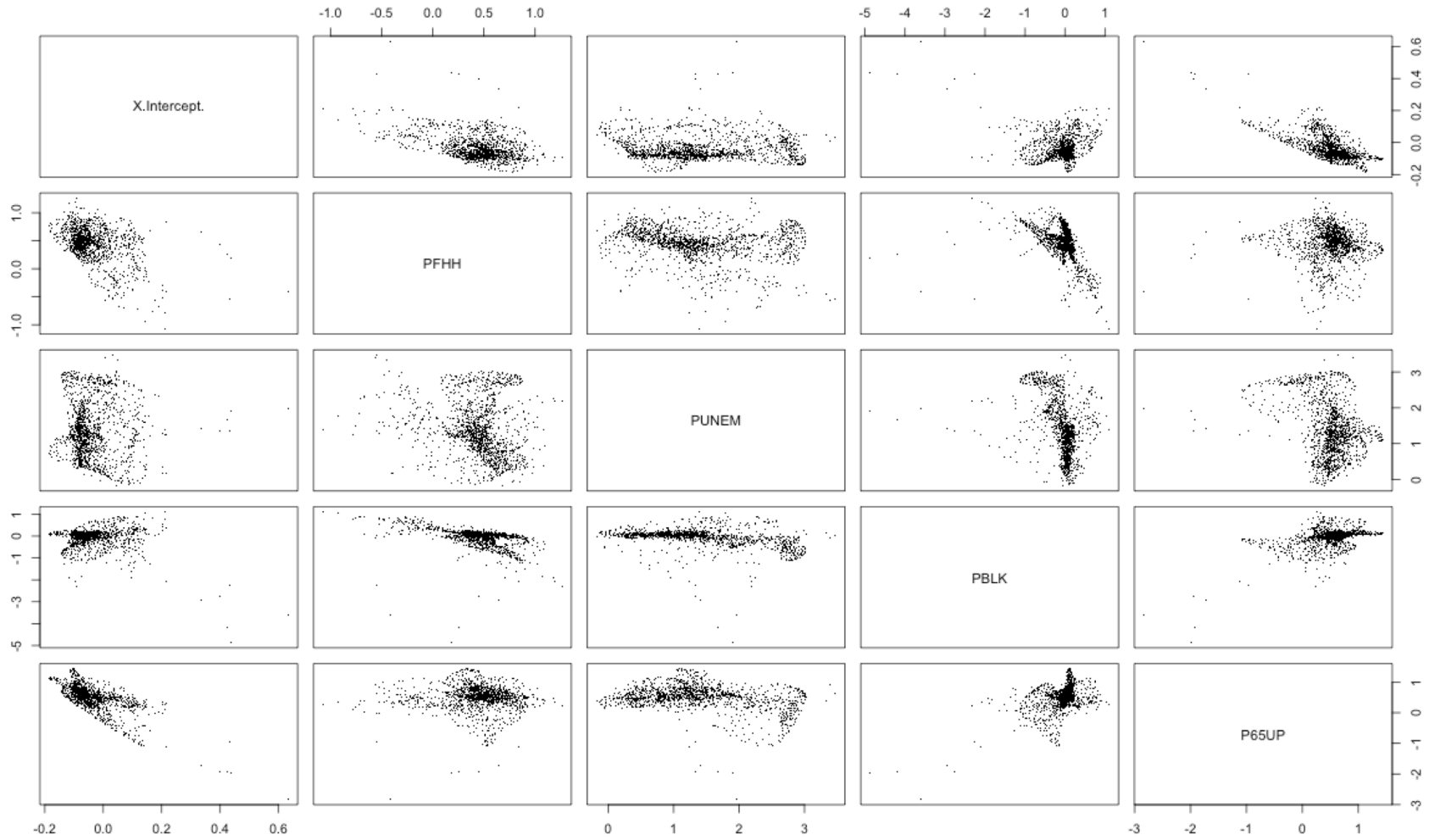
Multicollinearity and correlation among local regression coefficients in geographically weighted regression

Received: 25 October 2004 / Accepted: 21 February 2005
© Springer-Verlag 2005

Abstract Present methodological research on geographically weighted regression (GWR) focuses primarily on extensions of the basic GWR model, while ignoring well-established diagnostics tests commonly used in standard global regression analysis. This paper investigates multicollinearity issues surrounding the local GWR coefficients at a single location and the overall correlation between GWR coefficients associated with two different exogenous variables. Results indicate that the local regression coefficients are potentially collinear even if the underlying exogenous variables in the data generati
findings, applied GWR resea
interpreting the spatial patter
disease-mapping example is u
problem. Controlled experime
coefficient dependency issues
models that use eigenvectors
variables.

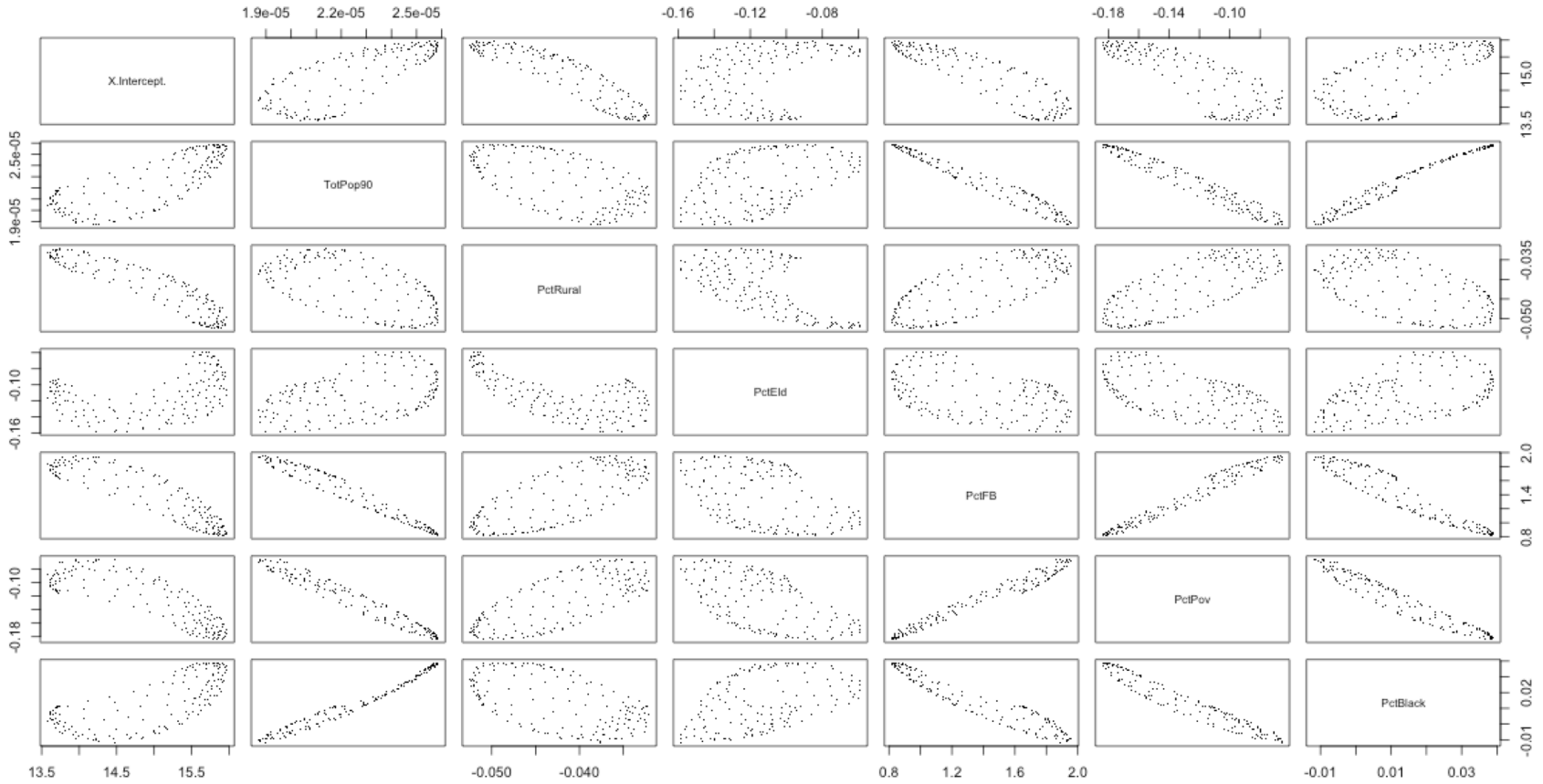
his paper demonstrates the need for diagnostic tools, especially regardin
e issues of multicollinearity that may arise in GWR. This paper also found
at the effects of multicollinearity are substantially stronger in the GWR
odel than in global regression models. In GWR, moderate to strong cor
lation of two explanatory variables makes their associated local parameter
most completely interdependent. This correlation of local regressio
efficients potentially invalidates any interpretation of individual GWR
arameter estimates and can facilitate misleading conclusions if the situatio

Checking correlation between coefficients.



The coefficients do not seem highly correlated in out model

Checking correlation between coefficients.



The this is an example of a model with highly correlated in out model

Reasons to use GWR

- GWR is a useful tool for local exploratory analysis
- GWR is truly a spatial technique that focuses on local heterogeneity.
  - It uses geographic information as well as attribute information
  - It employs a spatial weighting function with the assumption that near places are more similar than distant ones
  - The outputs are location specific hence mappable for further analysis
- The maps look “scientific” ☺

GWR

```
#######################################
##GEOGRAPHICALLY WEIGHTED REGRESSION
#######################################
bwG <- gwr.sel(PPOV ~ PFHH + PUNEM + PBLK + P65UP, data = soco, gweight=gwr.Gauss, verbose=T)

soco_gwr <- gwr(PPOV ~ PFHH + PUNEM + PBLK + P65UP, data = soco, bandwidth=bwG, gweight=gwr.Gauss)

## Residuals
res <- soco_gwr$SDF$gwr.e

##SETUP TO MAP RESIDUALS
classes_fx <- classIntervals(res, n=5, style="quantile")
pal <- brewer.pal(5, "YlOrRd")
cols <- findColours(classes_fx,pal)

##MAP A COEF
coef <- soco_gwr$SDF$PUNEM
classes_fx <- classIntervals(coef, n=5, style="quantile")
cols <- findColours(classes_fx,pal)
plot(soco,col=cols, border="transparent")
legend(x="bottom",cex=1,fill=attr(cols,"palette"),bty="n",legend=names(attr(cols, "table")),title="Local Coefficient Estimates (PUNEM)",ncol=3)

##MAP R2
classes_fx <- classIntervals(soco_gwr$SDF$localR2, n=5, style="quantile")
cols <- findColours(classes_fx,pal)
pal <- brewer.pal(5, "YlOrRd")
plot(soco,col=cols, border="grey")
legend(x="bottom",cex=1,fill=attr(cols,"palette"),bty="n",legend=names(attr(cols, "table")),title="Local R^2",ncol=3)

##CHECK FOR LOCAL CORRELATION IN THE COEFFICIENTS
pairs(as(soco_gwr$SDF, "data.frame")[,2:6], pch=".")
```