Applied Spatial Statistics in R

Yuri M. Zhukov

IQSS, Harvard University

January 19, 2010

Overview

1. Introduction
   - Why use spatial methods?
   - The spatial autoregressive data generating process
2. Spatial Data and Basic Visualization in R
   - Points
   - Polygons
   - Grids
3. Spatial Autocorrelation
4. Spatial Weights
5. Point Processes
6. Geostatistics
7. Spatial Regression
   - Models for continuous dependent variables
   - Models for categorical dependent variables
   - Spatiotemporal models

Motivations for going spatial

Independence assumption not valid

The attributes of observation i may influence the attributes of j.

Spatial heterogeneity

The magnitude and direction of a treatment effect may vary across space.

Omitted variable bias

There may be some unobserved or latent influences shared by geographical or network "neighbors".

Illustrative examples

Epidemiology

How to model the spread of a contagious disease?

Criminology

How to identify crime hot spots?

Real estate

How to predict housing prices?

Counterinsurgency

"Oil spot" modeling and clear-hold-build

Organizational learning and network diffusion

How to model the adoption of an innovation?

Non-spatial DGP

In the linear case:

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

Assumptions

- Observed values at location i independent of those at location j
- Residuals are independent ($E[\epsilon_i\epsilon_j] = E[\epsilon_i]E[\epsilon_j] = 0$)

The independence assumption greatly simplifies the model, but may be difficult to justify in some contexts...

Spatial DGP

With two neighbors i and j:

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

Assumptions

- Observed values at location i depend on those at location j, and vice versa
- Data generating process is "simultaneous" (more on this later)

Spatial DGP

With n observations, we can generalize:

$$y_i = \rho \sum_{j=1}^{n} W_{ij} y_j + X_i \beta + \epsilon_i$$
$$\epsilon_i \sim N(0, \sigma^2), \quad i = 1, \ldots, n$$

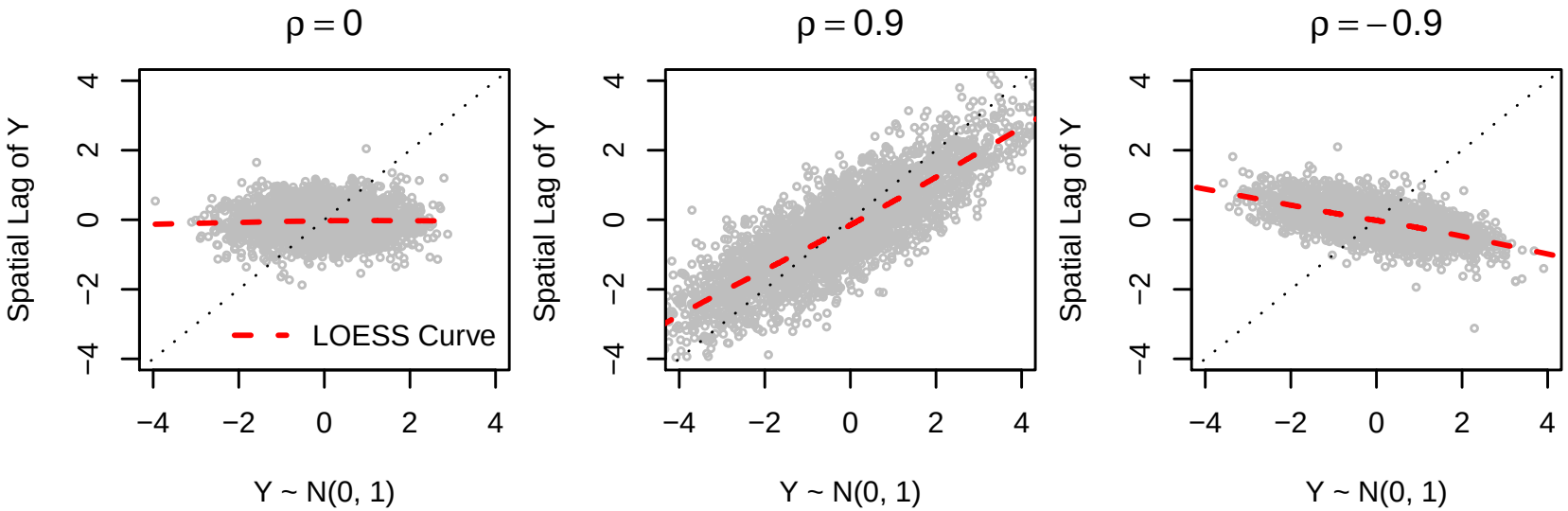
In matrix notation:

$$\mathbf{y} = \rho \mathbf{W} \mathbf{y} + \mathbf{X} \beta + \epsilon$$
$$\epsilon \sim N(0, \sigma^2 \mathbf{I}_n)$$

where $\mathbf{W}$ is the spatial weights matrix, ρ is a spatial autoregressive scalar parameter, and $\mathbf{I}_n$ is an $n \times n$ identity matrix

Spatial DGP

- When $\rho = 0$, the variable in not spatially autocorrelated. Information about a measurement in one location gives us no information about the value in neighboring locations (spatial independence).
- When $\rho > 0$, the variable in positively spatially autocorrelated. Neighboring values tend to be similar to each other (clustering).
- When $\rho < 0$, the variable in negatively spatially autocorrelated. Neighboring values tend to be different to each other (segregation).



Spatial DGP

Let's develop this further, for the moment dropping $\mathbf{X}\beta$ and introducing constant term vector of ones ι_n:

$$\begin{aligned}
\mathbf{y} &= \rho\mathbf{W}\mathbf{y} + \iota_n\alpha + \epsilon \\
(\mathbf{I}_n - \rho\mathbf{W})y &= \iota_n\alpha + \epsilon \\
y &= (\mathbf{I}_n - \rho\mathbf{W})^{-1}\iota_n\alpha + (\mathbf{I}_n - \rho\mathbf{W})^{-1}\epsilon \\
\epsilon &\sim N(0, \sigma^2\mathbf{I}_n)
\end{aligned}$$

Spatial DGP

Assuming $|\rho| < 1$, the inverse can be expressed as an infinite series

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

implying that

$$\begin{aligned} y = \iota_n\alpha + \rho\mathbf{W}\iota_n\alpha + \rho^2\mathbf{W}^2\iota_n\alpha + \dots \\ + \epsilon + \rho\mathbf{W}\epsilon + \rho^2\mathbf{W}^2\epsilon + \dots \end{aligned}$$

Since α is a scalar and $\mathbf{W}\iota_n = \iota_n$ (similarly, $\mathbf{W}(\mathbf{W}\iota_n) = \dots = \mathbf{W}^q\iota_n = \iota_n$ $\forall \quad q \geq 0$), this expression simplifies to:

$$y = (1 - \rho)^{-1}\iota_n\alpha + \epsilon + \rho\mathbf{W}\epsilon + \rho^2\mathbf{W}^2\epsilon + \dots$$

Spatial DGP

- Let's say that the rows of the weights matrix $\mathbf{W}$ represent first-order neighbors.
- Then by matrix multiplication, the rows of $\mathbf{W}^2$ would represent second-order neighbors (neighbors of one's neighbors), $\mathbf{W}^3$ third-order neighbors, and so on.
- But wait a minute... isn't i is a second-order neighbor of itself?
- This introduces simultaneous feedback into the model, where each observation y_i depends on the disturbances associated with both first- and higher-order neighbors.
- The influence of higher order neighbors declines when ρ is small (ρ can be interpreted as a discount factor reflecting a decay of influence for more distant observations)
- ...but we still have a mean and VCov structure for observations in the vector $\mathbf{y}$ that depends in a complicated way on other observations.

Spatial DGP

- Simultaneous feedback is not necessarily a bad thing...
- It can be useful if we're modeling spatial spillover effects from neighboring observations to an origin location *i* where the initial impact occurred.
- This approach effectively treats all observations as potential origins of an impact.
- But we also have to be very careful in how we treat spatial data, and how we conceive of the feedback process with regard to time.
- With cross-sectional data, observations are often taken to represent an equilibrium outcome of the spatial process we are modeling.
- But if spatial feedback is modeled as a dynamic process, the measured spatial dependence may vary with the time scale of data collection.

Further Reading

- A.D. Cliff and J.K. Ord (1973), *Spatial Autocorrelation* (London: Pion)
- B.D. Ripley(1981), *Spatial Statistics* (New York: Wiley)
- L. Anselin (1988), *Spatial Econometrics: Methods and Models* (Dordrecht, The Netherlands: Kluwer Academic Publishers)
- P.J. Diggle (2003), *Statistical Analysis of Spatial Point Patterns* (London: Arnold)
- R.S. Bivand, E.J. Pebesma and V. Gomez-Rubio (2008), *Applied Spatial Data Analysis with R* (New York: Springer)
- J. Le Sage and R.K. Pace (2009), *Introduction to Spatial Econometrics* (CRC Press)

Outline

Software options

Application	Availability	Learning Curve	Key Functionality
ArcGIS	License	Medium	Geoprocessing, visualization
GeoBUGS	Free	High	Bayesian analysis
GeoDa	Free	Low	ESDA, ML spatial regression
GRASS	Free	High	Image processing, spatial modeling
R	Free	High	Weights, spatial econometrics, geostatistics
STARS	Free	Low	Space-time analysis

Spatial Analysis in R

Task	Packages
Data management	`sp`, `rgdal`, `maptools`
Integration with other GIS	`rgdal`, `RArcInfo`, `SQLiteMap`, `RgoogleMaps`, `spgrass6`, `RPyGeo`, `R2WinBUGS`, `geonames`
Point pattern analysis	`spatstat`, `splancs`, `spatialkernel`
Geostatistics	`gstat`, `geoR`, `geoRglm`, `spBayes`
Disease mapping	`DCluster`, `spgwr`, `glmmBUGS`, `diseasemapping`
Spatial regression	`spdep`, `spatcounts`

Where to Find Spatial Data?

Coordinates and Basemaps:

Geographical Place Names `http://www.geonames.org/`

Global Administrative Areas `http://gadm.org/country`

Land Cover and Elevation `http://eros.usgs.gov/#/Find_Data`

Geo-referenced Data:

2000 U.S. Census Data
`http://disasternets.calit2.uci.edu/census2000/`

Natural Resources `http://www.prio.no/CSCW/Datasets/Geographical-and-Resource/`

International Conflict Data `http://www.acleddata.com/`

A large number of links is also available at `http://gis.harvard.edu/`

Points

Points are the most basic form of spatial data

- Points are pairs of coordinates (x, y), representing events, observation posts, individuals, cities or any other discrete object defined in space.
- Let's take a look at the dataset `crime`, which is just a table of geographic coordinates (decimal degrees) for crime locations in Baltimore, MD.

```
head(crime)
```

```
  ID   LONG        LAT
1  1 -76.65159 39.23941
2  2 -76.47434 39.35274
3  3 -76.51726 39.25874
4  4 -76.52607 39.40707
5  5 -76.51001 39.33571
6  6 -76.70375 39.26605
```

- To work with these data in R, we will need to create a `spatial` object from this table.

Points

Create matrix of coordinates

```
 sp_point <- cbind(crime$LONG, crime$LAT)
colnames(sp_point) <- c("LONG","LAT")
```

Define Projection: UTM Zone 17

```
 proj <- CRS("+proj=utm +zone=17
+datum=WGS84")
```

Create spatial object

```
 data.sp <- SpatialPointsDataFrame(
coords=sp_point, data=crime,
proj4string=proj)
```

Plot the data

```
 plot(data.sp, pch=16, cex=.5, axes=T)
```

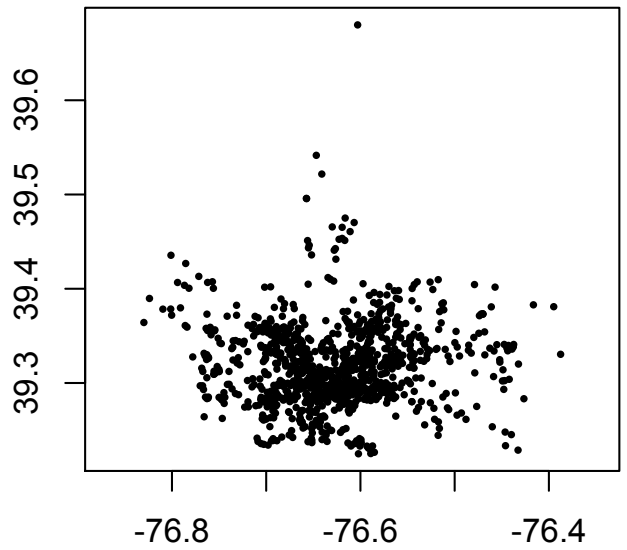


Figure: Baltimore Crime Locations

Polygons and Lines

Polygons can be thought of as sequences of connected points, where the first point is the same as the last.

- An open polygon, where the sequence of points does not result in a closed shape with a defined area, is called a line.
- In the R environment, line and polygon data are stored in objects of classes `SpatialPolygons` and `SpatialLines`:

```
getClass("Polygon")
```

```
Class Polygon [package "sp"]
 Name:      labpt      area      hole   ringDir   coords
 Class:   numeric   numeric   logical   integer   matrix
```

```
getClass("SpatialPolygons")
```

```
Class SpatialPolygons [package "sp"]
 Name:    polygons   plotOrder      bbox   proj4string
 Class:       list     integer    matrix           CRS
```

Polygons and Lines

Let's take a look at the `election` dataset.

```
summary(election)
```

```
 Object of class SpatialPolygonsDataFrame
                           min         max
Coordinates:    r1   -124.73142   -66.96985
                r2     24.95597    49.37173
Is projected:  TRUE
proj4string :  [+proj=lcc+lon_0=90w +lat_1=20n +lat_2=60n]
```

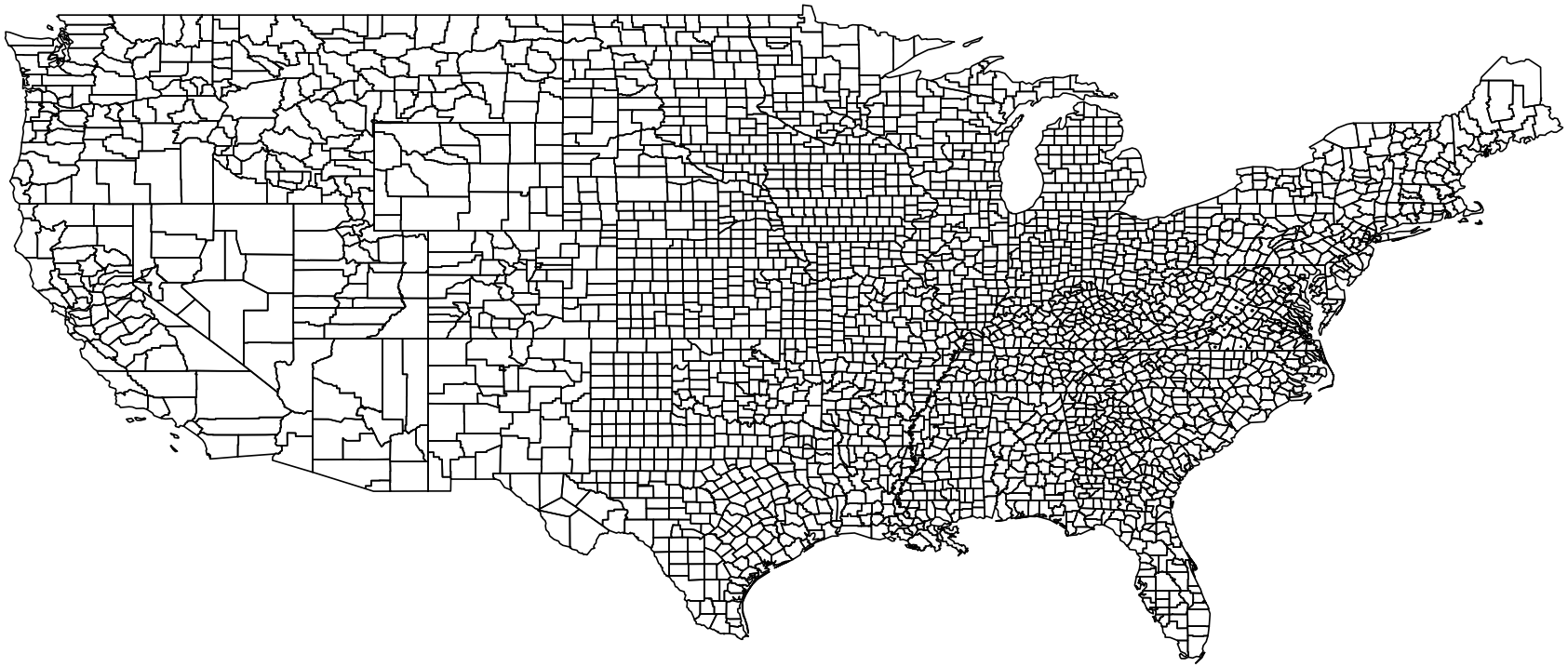
- The data are stored as a `SpatialPolygonsDataFrame`, which is a subclass of `SpatialPolygons` containing a data.frame of attributes.
- In this case, the polygons represent U.S. counties and attributes include results from the 2004 Presidential Election.

```
names(election)
```

```
 [1] "NAME" "STATE_NAME" "STATE_FIPS" "CNTY_FIPS" "FIPS" "AREA" "FIPS_num" "Bush"
[9] "Kerry" "County_F" "Nader" "Total" "Bush_pct" "Kerry_pct" "Nader_pct"
```

Polygons and Lines: Visualization

Let's visualize the study region with `plot(election)`.



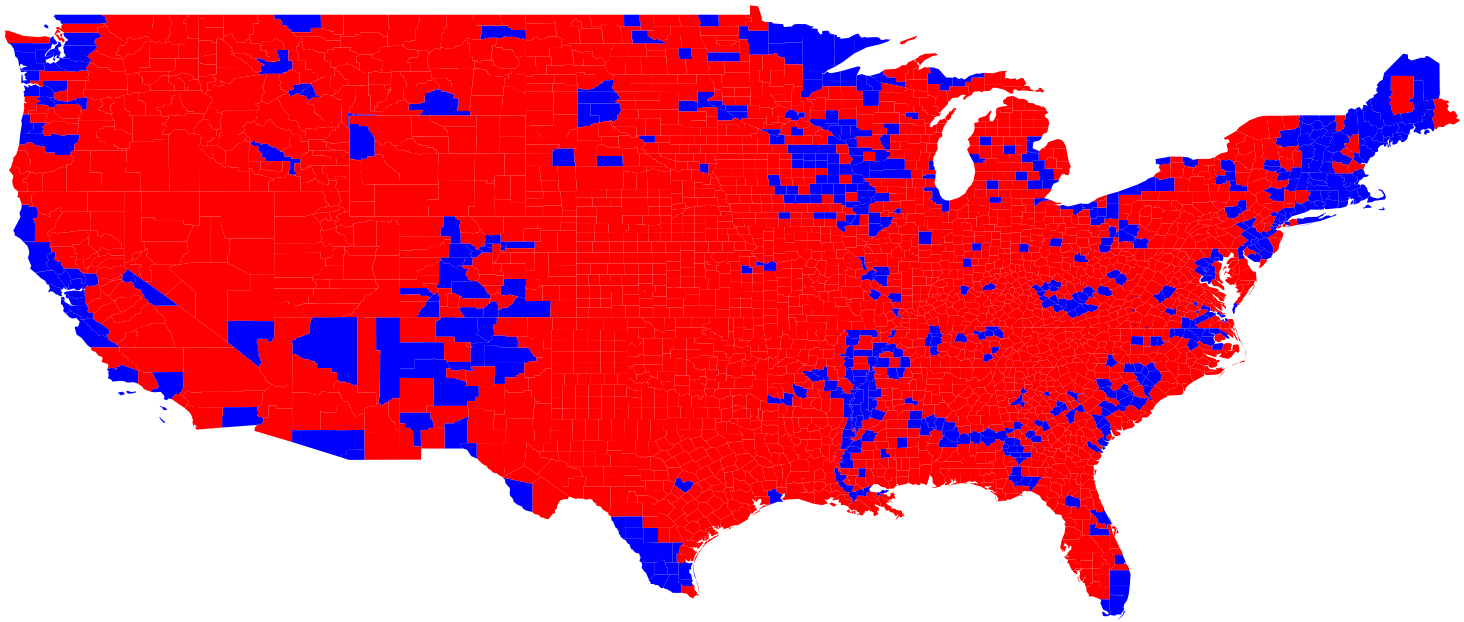
Polygons and Lines: Visualization

- For a `categorical` variable (win/lose), visualization is simple...
  1. Create a vector of colors, where each county won by Bush is coded `"red"` and every each county won by Kerry is `"blue"`.

     ```
     cols <- ifelse(election$Bush > election$Kerry,"red","blue")
     ```

  2. Use the resulting color vector with the `plot()` command.

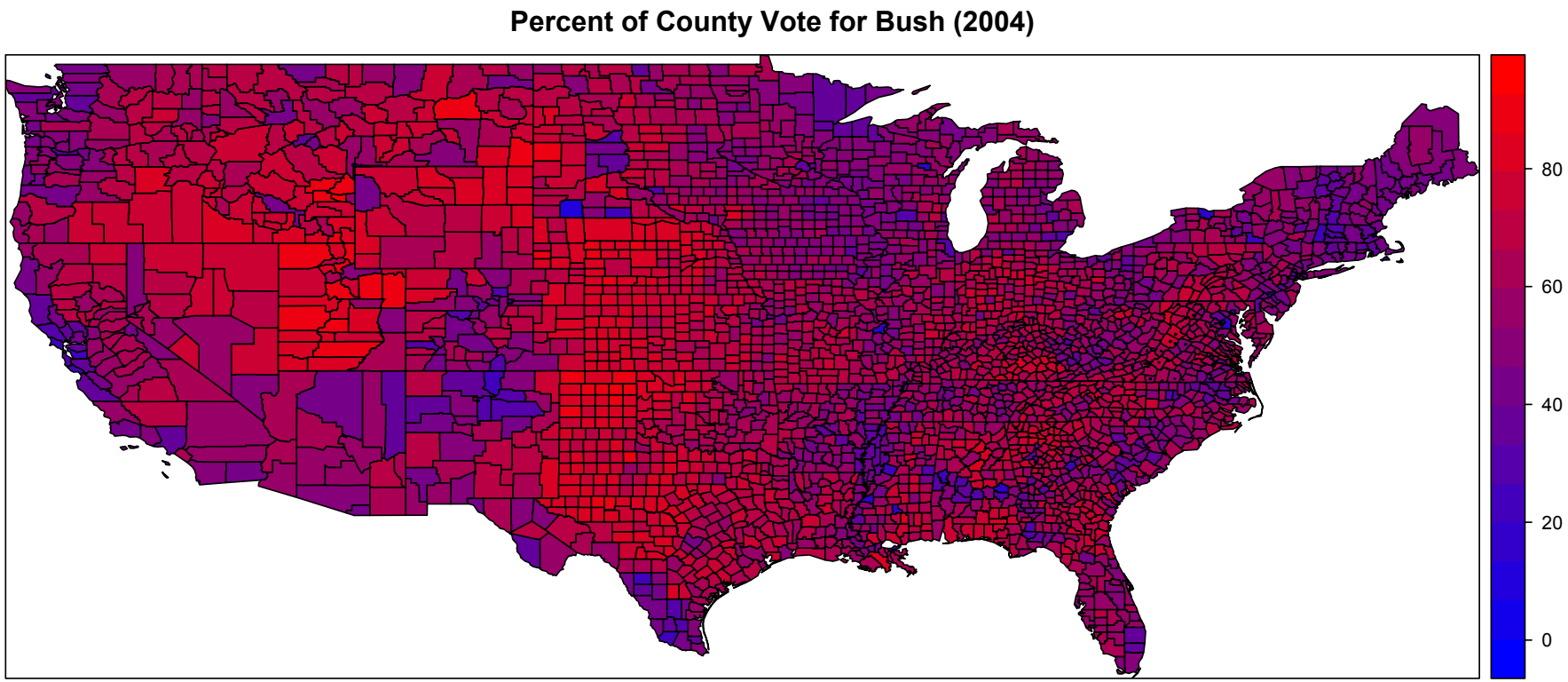
     ```
     plot(election,col=cols,border=NA)
     ```



Polygons and Lines: Visualization

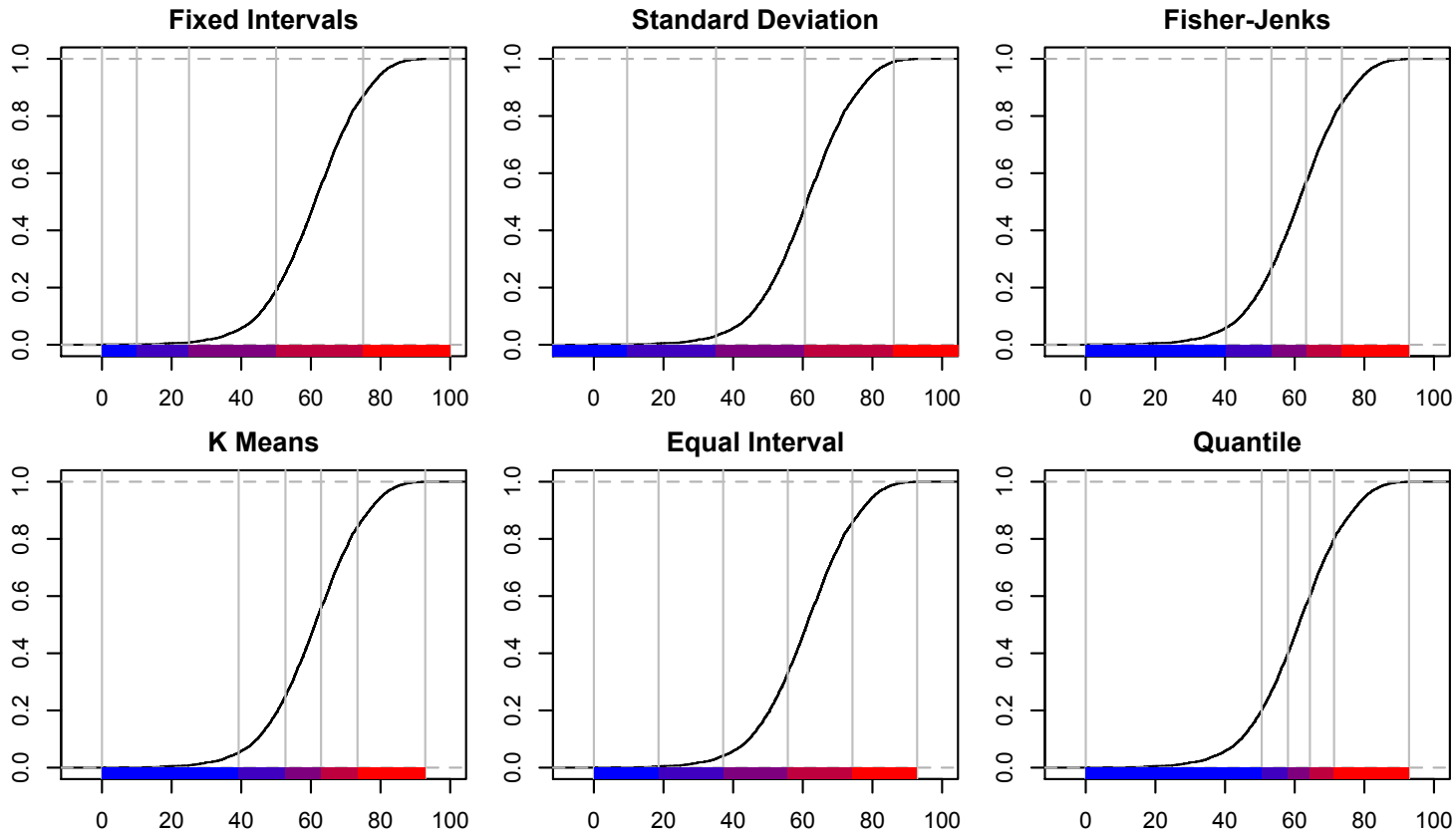
With a continuous variable, the same logic applies. A relatively simple approach is to create a custom color palette and use `spplot()`.

```
br.palette <- colorRampPalette(c("blue", "red"), space = "rgb")
spplot(data, zcol="Bush_pct", col.regions=br.palette(100))
```



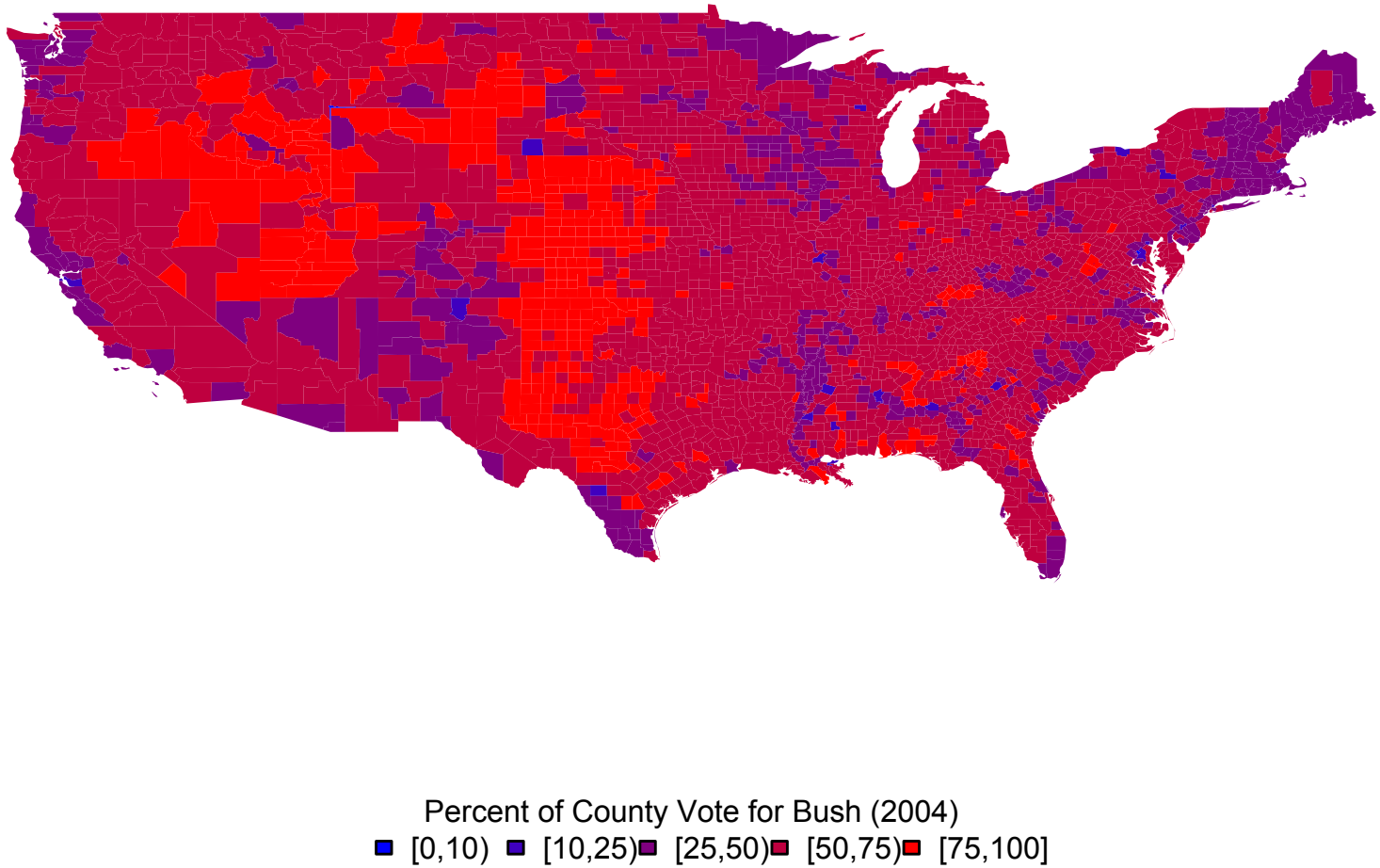
Polygons and Lines: Visualization

- We can also create a color palette for custom classification intervals with the `classInt` package.
- Here is a comparison of six such palettes for the variable `Bush_pct`, or percentage of popular vote won by George W. Bush.



Polygons and Lines: Visualization

- Here is a plot of county results using the fixed intervals:



Grids

A raster grid divides the study region into a set of identical, regularly-spaced, discrete elements (pixels), each of which records the value or presence/absence of a quantity of interest.

- Rasters originated in image processing, and are used to record properties varying continuously with space.
- Common uses include remote sensing data, elevation models and spatial prediction (weather forecasts, disease risk, etc.).

Take a look at the data structure of the `volcano` dataset, a grid of elevation measures for the Maunga Whau Volcano in New Zealand:

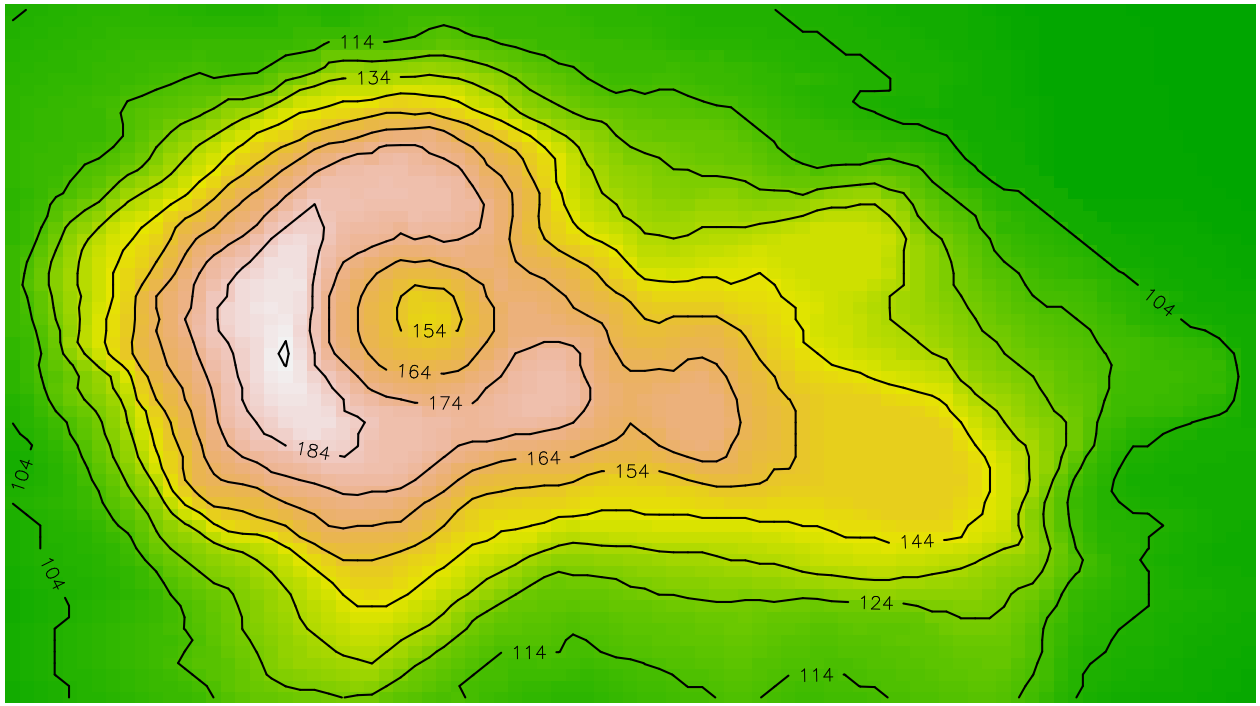
```
head(volcano)[,1:6]
```

	1	2	3	4	5	6
1	100.00	100.00	101.00	101.00	101.00	101.00
2	101.00	101.00	102.00	102.00	102.00	102.00
3	102.00	102.00	103.00	103.00	103.00	103.00
4	103.00	103.00	104.00	104.00	104.00	104.00
5	104.00	104.00	105.00	105.00	105.00	105.00
6	105.00	105.00	105.00	106.00	106.00	106.00

Grids: Visualization

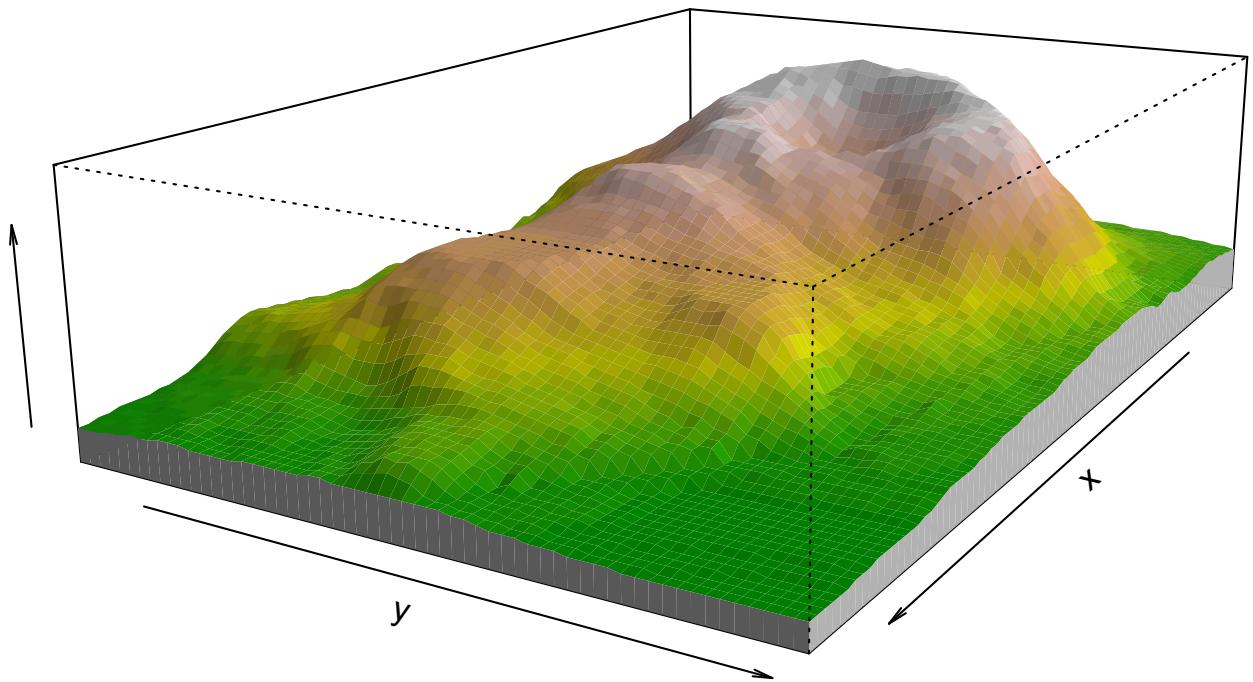
Grids can be visualized in two dimensions as contour plots, as images with a color gradient, or both.

```
image(x=10*(1:nrow(z)), y=10*(1:ncol(z)), z=volcano,
col=terrain.colors(100), axes=F)
contour(x=10*(1:nrow(z)), y=10*(1:ncol(z)), z=volcano,
levels=seq(from=min(z), to=max(z), by=10),axes=F, add=T)
```



Grids: Visualization

Grids can also be visualized in three dimensions with the `persp()` command and a grid of palette colors (similar to vector of colors from previous example).



Examples in R

Switch to R tutorial script. Section 1.

Outline

What is Spatial Autocorrelation?

- Spatial autocorrelation measures the degree to which a phenomenon of interest is correlated to itself in space (Cliff and Ord 1973, 1981).
- Tests of spatial autocorrelation examine whether the observed value of a variable at one location is independent of values of that variable at neighboring locations.
- Positive spatial autocorrelation indicates that similar values appear close to each other, or cluster, in space
- Negative spatial autocorrelation indicates that neighboring values are dissimilar or, equivalenty, that similar values are dispersed.
- Null spatial autocorrelation indicates that the spatial pattern is random.

Global autocorrelation: Moran's $\mathcal{I}$

- The Moran's $\mathcal{I}$ coefficient calculates the ratio between the product of the variable of interest and its spatial lag, with the product of the variable of interest, adjusted for the spatial weights used.

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

- where y_i is the value of a variable for the ith observation, $\bar{y}$ is the sample mean and w_{ij} is the spatial weight of the connection between i and j.
- Values range from –1 (perfect dispersion) to +1 (perfect correlation). A zero value indicates a random spatial pattern.
- Under the null hypothesis of no autocorrelation, $\mathbb{E}[\mathcal{I}] = \frac{-1}{n-1}$

Global autocorrelation: Moran's $\mathcal{I}$

- Calculating the variance of Moran's $\mathcal{I}$ is a little more involved:

$$
\begin{aligned}
Var(\mathcal{I}) =& \frac{n\mathfrak{S}_1 - \mathfrak{S}_2\mathfrak{S}_3}{(n-1)(n-2)(n-3)(\sum_i \sum_j w_{ij})^2} \\
\mathfrak{S}_1 =& (n^2 - 3n + 3)\left(\frac{1}{2}\sum_i\sum_j (w_{ij} + w_{ji})^2\right) \\
& - n\left(\sum_i (\sum_j w_{ij} + \sum_j w_{ji})^2\right) + 3(\sum_i \sum_j w_{ij})^2 \\
\mathfrak{S}_2 =& \frac{n^{-1}\sum_i (y_i - \bar{x})^4}{(n^{-1}\sum_i (y_i - \bar{x})^2)^2} \\
\mathfrak{S}_3 =& \frac{1}{2}\sum_i\sum_j (w_{ij} + w_{ji})^2 - 2n\left(\frac{1}{2}\sum_i\sum_j (w_{ij} + w_{ji})^2\right) \\
& + 6\left(\sum_i\sum_j w_{ij}\right)^2
\end{aligned}
$$

Global autocorrelation: Geary's $\mathcal{C}$

- The Geary's $\mathcal{C}$ uses the sum of squared differences between pairs of data values as its measure of covariation.

$$\mathcal{C} = \frac{(n-1)\sum_i \sum_j w_{ij}(y_i - y_j)^2}{2(\sum_i \sum_j w_{ij}) \sum_i (y_i - \bar{y})^2}$$

- where y_i is the value of a variable for the ith observation, $\bar{y}$ is the sample mean and w_{ij} is the spatial weight of the connection between i and j.
- Values range from 0 (perfect correlation) to 2 (perfect dispersion). A value of 1 indicates a random spatial pattern.

Global autocorrelation: Join Counts

- When the variable of interest is *categorical*, a join count analysis can be used to assess the degree of clustering or dispersion.
- A binary variable is mapped in two colors (Black & White), such that a join, or edge, is classified as either *WW* (0-0), *BB* (1-1), or *BW* (1-0).
- Join count statistics can show
  - positive spatial autocorrelation (clustering) if the number of *BW* joins is significantly *lower* than what we would expect by chance,
  - negative spatial autocorrelation (dispersion) if the number of *BW* joins is significantly *higher* than what we would expect by chance,
  - null spatial autocorrelation (random pattern) if the number of *BW* joins is approximately *the same* as what we would expect by chance.

Global autocorrelation: Join Counts

- By the naive definition of probability, if we have n_B Black units and $n_W = n - n_B$ White units, the respective probabilities of observing the two types of units are:

$$P_B = \frac{n_B}{n} \qquad P_W = \frac{n - n_B}{n} = 1 - P_B$$

- The probabilities of BB and WW in two adjacent cells are

$$P_{BB} = P_B P_B = P_B^2 \qquad P_{WW} = (1 - P_B)(1 - P_B) = (1 - P_B)^2$$

- The probability of BW in two adjacent cells is

$$P_{BW} = P_B(1 - P_B) + (1 - P_B)P_B = 2P_B(1 - P_B)$$

Global autocorrelation: Join Counts

- The expected counts of each type of join are:

$$\mathbb{E}[BB] = \frac{1}{2}\sum_i \sum_j w_{ij} P_B^2 \qquad \mathbb{E}[WW] = \frac{1}{2}\sum_i \sum_j w_{ij}(1-P_B)^2$$

$$\mathbb{E}[BW] = \frac{1}{2}\sum_i \sum_j w_{ij} 2P_B(1-P_B)$$

- Where $\frac{1}{2}\sum_i \sum_j w_{ij}$ is the total number of joins (of any type) on a map, assuming a binary connectivity matrix.
- The observed counts are:

$$BB = \frac{1}{2}\sum_i \sum_j w_{ij} y_i y_j \qquad WW = \frac{1}{2}\sum_i \sum_j w_{ij}(1-y_i)(1-y_j)$$

$$BW = \frac{1}{2}\sum_i \sum_j w_{ij}(y_i - y_j)^2$$

- where $y_i = 1$ if unit i is Black and $y_i = 0$ if White.

Global autocorrelation: Join Counts

- The variance of BW is calculated as

$$\begin{aligned}\sigma^2_{BW} =& \mathbb{E}[BW^2] - \mathbb{E}[BW]^2 \\ =& \frac{1}{4}\left(\frac{2\mathfrak{s}_2 n_B(n-n_B)}{n(n-1)} + \frac{(\mathfrak{s}_3-\mathfrak{s}_1)n_B(n-n_B)}{n(n-1)}\right. \\ &\left. + \frac{4(\mathfrak{s}_1^2+\mathfrak{s}_2-\mathfrak{s}_3)n_B(n_B-1)(n-n_B)(n-n_B-1)}{n(n-1)(n-2)(n-3)}\right) - \mathbb{E}[BW]^2\end{aligned}$$

$$\mathfrak{s}_1 = \sum_i \sum_j w_{ij}$$

$$\mathfrak{s}_2 = \frac{1}{2}\sum_i \sum_j (w_{ij} - w_{ji})^2$$

$$\mathfrak{s}_3 = \sum_i (\sum_j w_{ij} + \sum_j w_{ji})^2$$

Global autocorrelation: Join Counts

- A test statistic for the BW join count is

$$\mathcal{Z}(BW) = \frac{BW - \mathbb{E}[BW]}{\sqrt{\sigma^2_{BW}}}$$

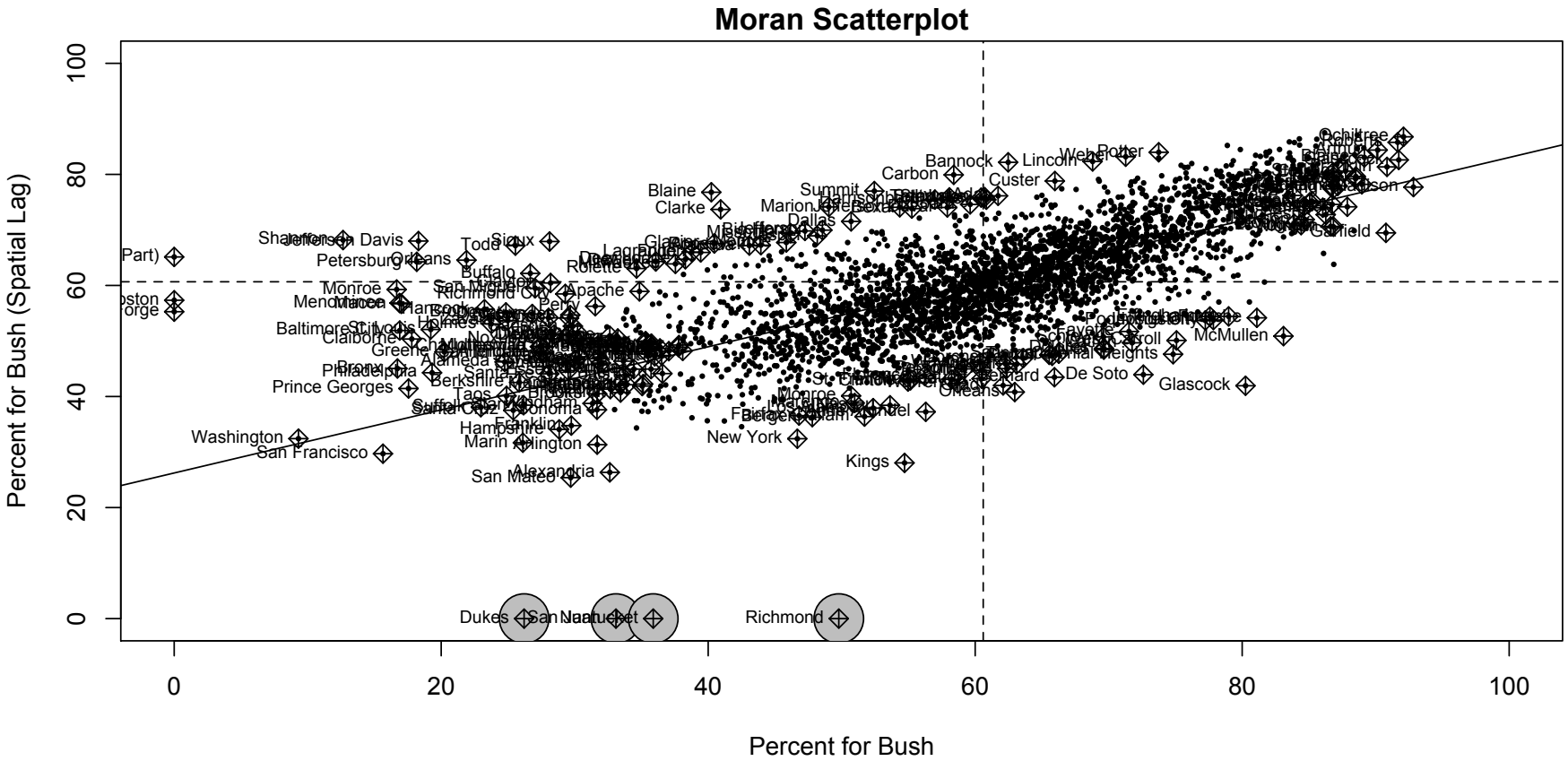
- The join count statistic is assumed to be asymptotically normally distributed under the null hypothesis of no spatial autocorrelation.
- The test of significance is then provided by evaluating the BW statistic as a standard deviate (Cliff and Ord, 1981).

Local autocorrelation

- Global tests for spatial autocorrelation are calculated from local relationships between observed values at spatial units and their neighbors.
- It is possible to break these measures down into their components, thus constructing local tests for spatial autocorrelation.
- These tests can be used to detect
  - Clusters, or units with similar neighbors
  - Hotspots, or units with dissimilar neighbors

Local autocorrelation

Below is a scatterplot of county vote for Bush and its spatial lag (average vote received in neighboring counties). The Moran's $\mathcal{I}$ coefficient is drawn as the slope of the linear relationship between the two. The plot is partitioned into four quadrants: low-low, low-high, high-low and high-high.



Local autocorrelation: Local Moran's $\mathcal{I}$

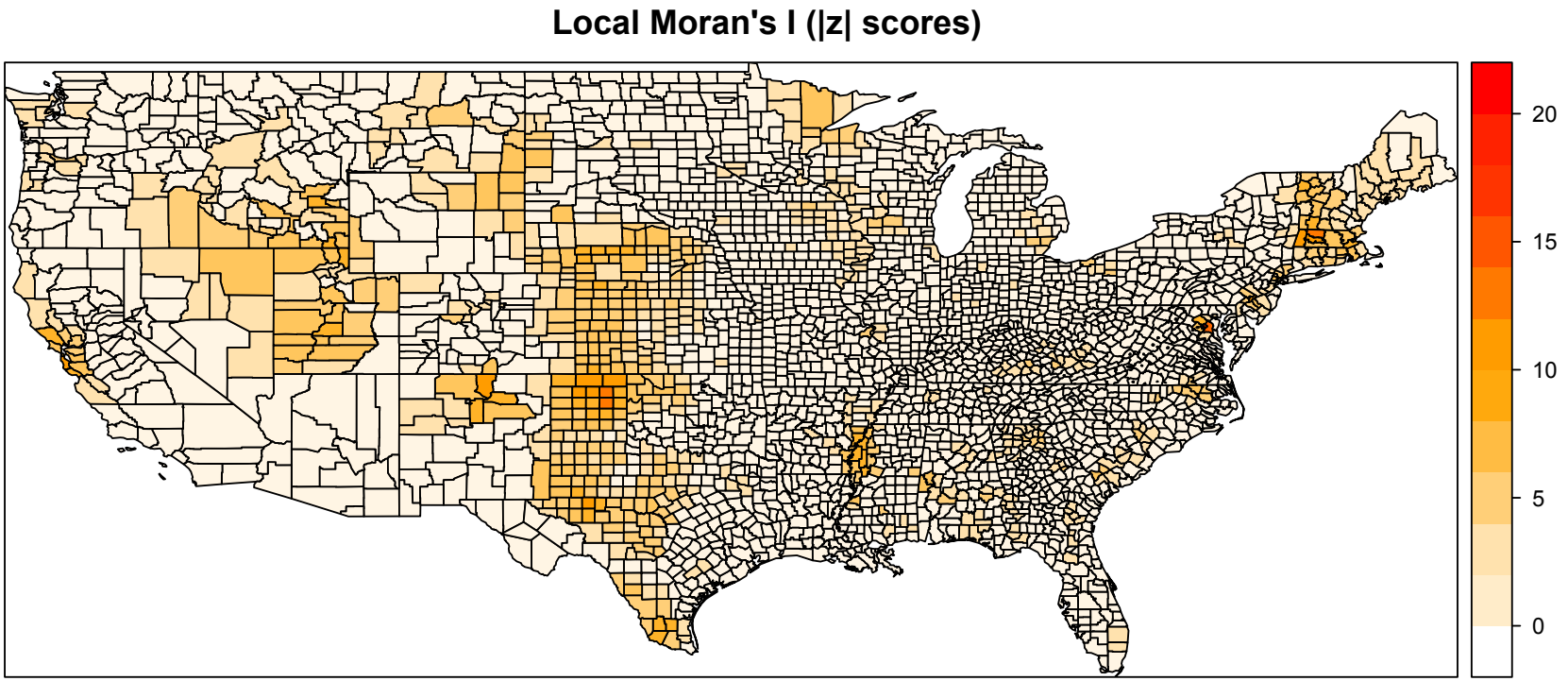
- A local Moran's $\mathcal{I}$ coefficient for unit i can be constructed as one of the n components which comprise the global test:

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

- As with global statistics, we assume that the global mean $\bar{y}$ is an adequate representation of the variable of interest.
- As before, local statistics can be tested for divergence from expected values, under assumptions of normality.

Local autocorrelation: Local Moran's $\mathcal{I}$

Below is a plot of Local Moran $|z|$-scores for the 2004 Presidential Elections. Higher absolute values of z scores (red) indicate the presence of "hot spots", where the percentage of the vote received by President Bush was significantly different from that in neighboring counties.



Words of Caution

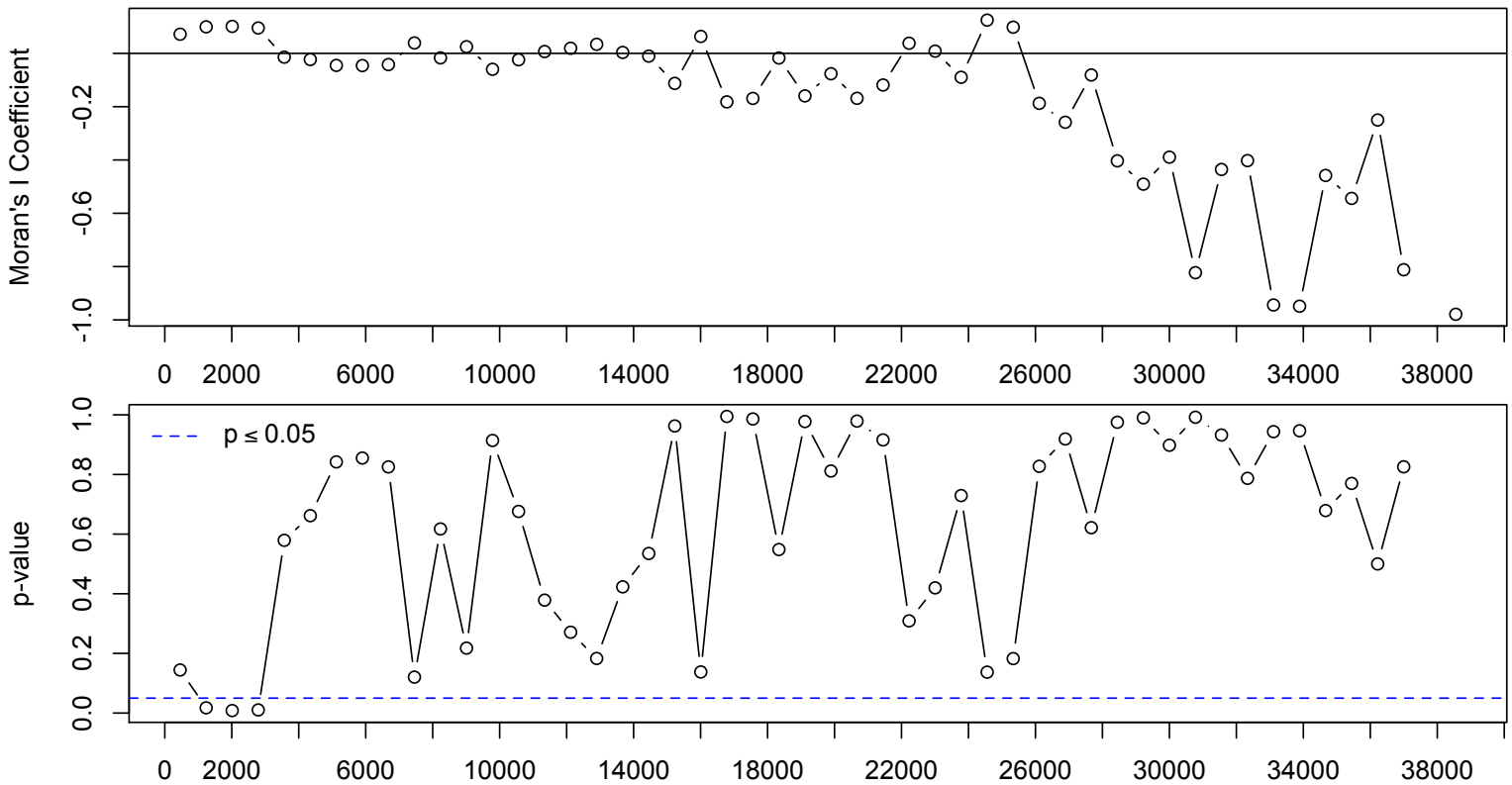
1. Autocorrelation tests are highly sensitive to spatial patterning in the variable of interest from any source. But by assuming that the mean model removes such systematic spatial patterning, spatial autocorrelation tests do not always produce useful insights into the DGP.

Words of Caution

1. Autocorrelation tests are highly sensitive to spatial patterning in the variable of interest from any source. But by assuming that the mean model removes such systematic spatial patterning, spatial autocorrelation tests do not always produce useful insights into the DGP.
2. These tests are also highly sensitive to one's choice of spatial weights. Where the weights do not reflect the "true" structure of spatial interaction, estimated autocorrelation (or lack thereof) may actually stem from misspecification.

Words of Caution

Below is a correlogram of Moran's $\mathcal{I}$ coefficients for Polity IV country democracy scores in 2008. The x-axis represents distances between country capitals, in kilometers. Here, democracy is significantly ($p \leq .05$) spatially autocorrelated only at distances of 3,000 km and below. So, autocorrelation estimates will depend highly on choice of lag distance.



Words of Caution

1. Autocorrelation tests are highly sensitive to spatial patterning in the variable of interest from any source. But by assuming that the mean model removes such systematic spatial patterning, spatial autocorrelation tests do not always produce useful insights into the DGP.
2. These tests are also highly sensitive to one's choice of spatial weights. Where the weights do not reflect the "true" structure of spatial interaction, estimated autocorrelation (or lack thereof) may actually stem from misspecification.
3. As originally designed, spatial autocorrelation tests assumed there are no neighborless units in the study area. When this assumption is violated, the size of n may be adjusted (reduced) to reflect the fact that some units are effectively being ignored. Not doing so will generally bias the absolute value for the autocorrelation statistic upward and the variance downward.

Examples in R

Switch to R tutorial script. Section 2.

Outline

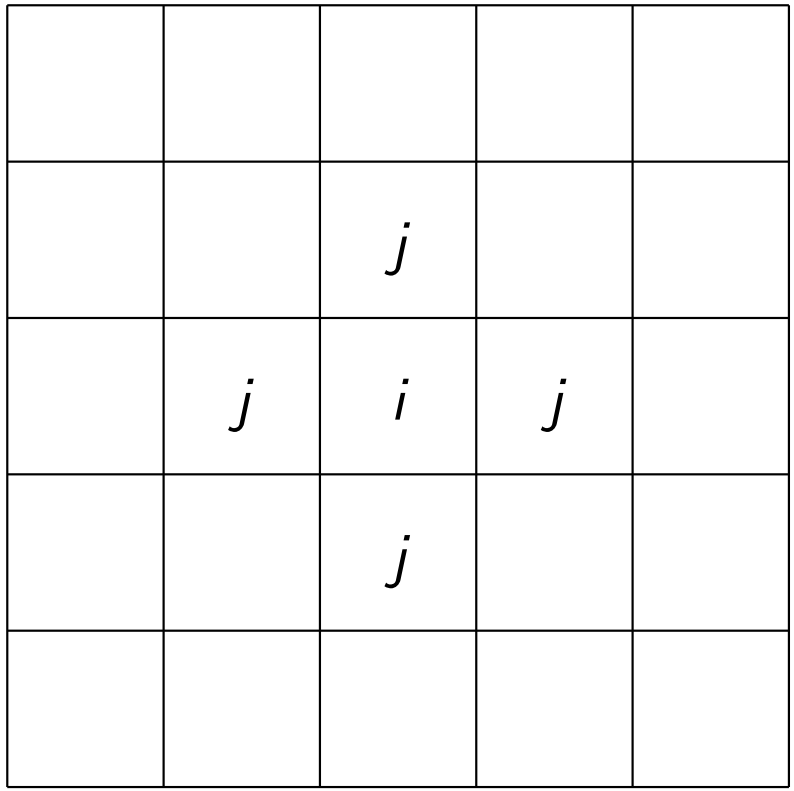
Choosing your neighbors?

- Most spatial weights matrices **W** are based on some version of a connectivity matrix **C**.
- **C** is an $n \times n$ binary matrix, where $i = \{1, 2, \ldots, n\}$ and $j = \{1, 2, \ldots, n\}$ are the units in the system (for example, countries in the international system).
- Entry $c_{ij} = 1$ if two units $i \neq j$ are considered connected, and $c_{ij} = 0$ if they are not.
- The tricky part is how the word "connected" is defined.

Areal Contiguity I: Regular Grids

Rook's case

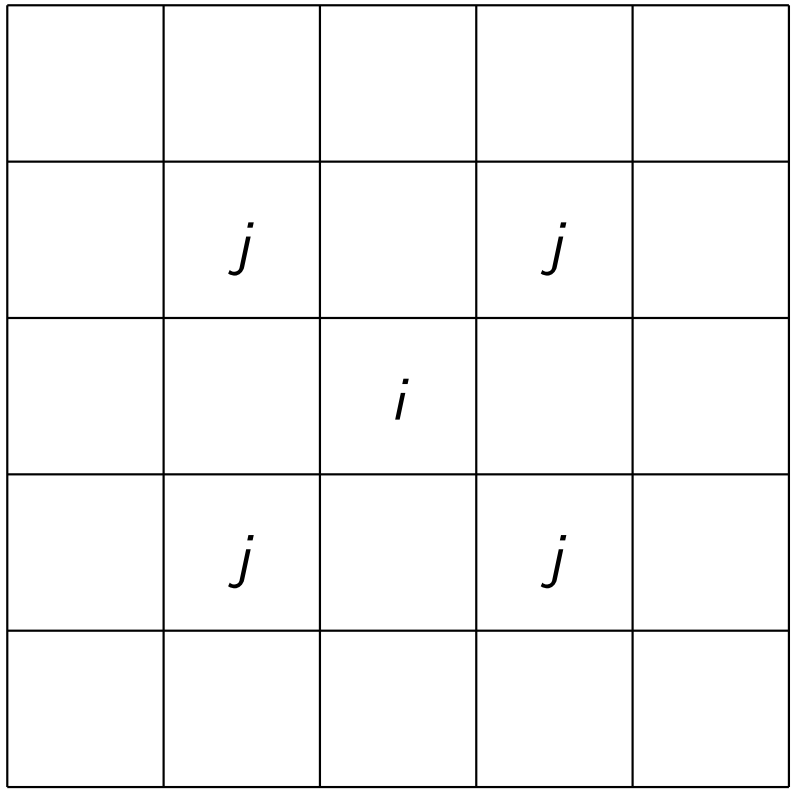
Cells sharing a common edge are considered contiguous



Areal Contiguity I: Regular Grids

Bishop's case

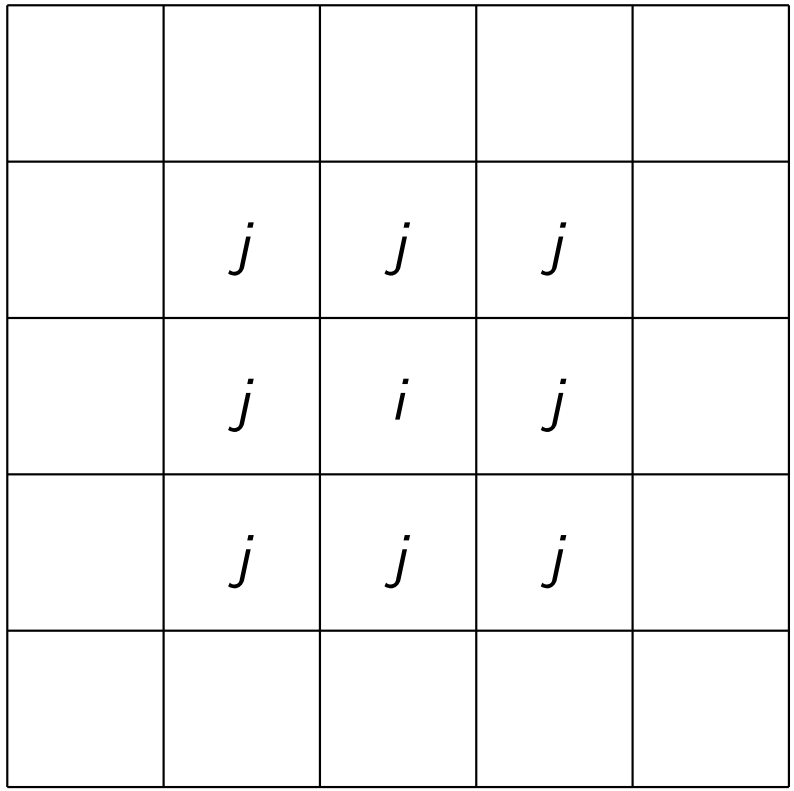
Cells sharing a common vertex are considered contiguous



Areal Contiguity I: Regular Grids

Queen's case

Cells sharing a common edge or common vertex are considered contiguous



Areal Contiguity I: Regular Grids

Second-order neighbors: (rook's case)

Cells sharing a common edge with first-order neighbors are considered contiguous

		k		
	k	j	k	
k	j	k	j	k
	k	j	k	
		k		

Areal Contiguity I: Regular Grids

- These conceptions of contiguity are useful when dealing with regular square grids or rectangular lattices, where the spatial structure can be easily summarized in elegant mathematical terms.
- But when spatial units consist of irregularly-shaped polygons, as is the case in most applied work (countries, census tracts, various administrative units), this simple characterization of contiguity breaks down...

Areal Contiguity II: Polygons ($\mathbf{W}_{CONT}$)

Two polygons x_i and x_j are neighbors if they share a common boundary.

Advantage

- Makes substantive sense

Disadvantage

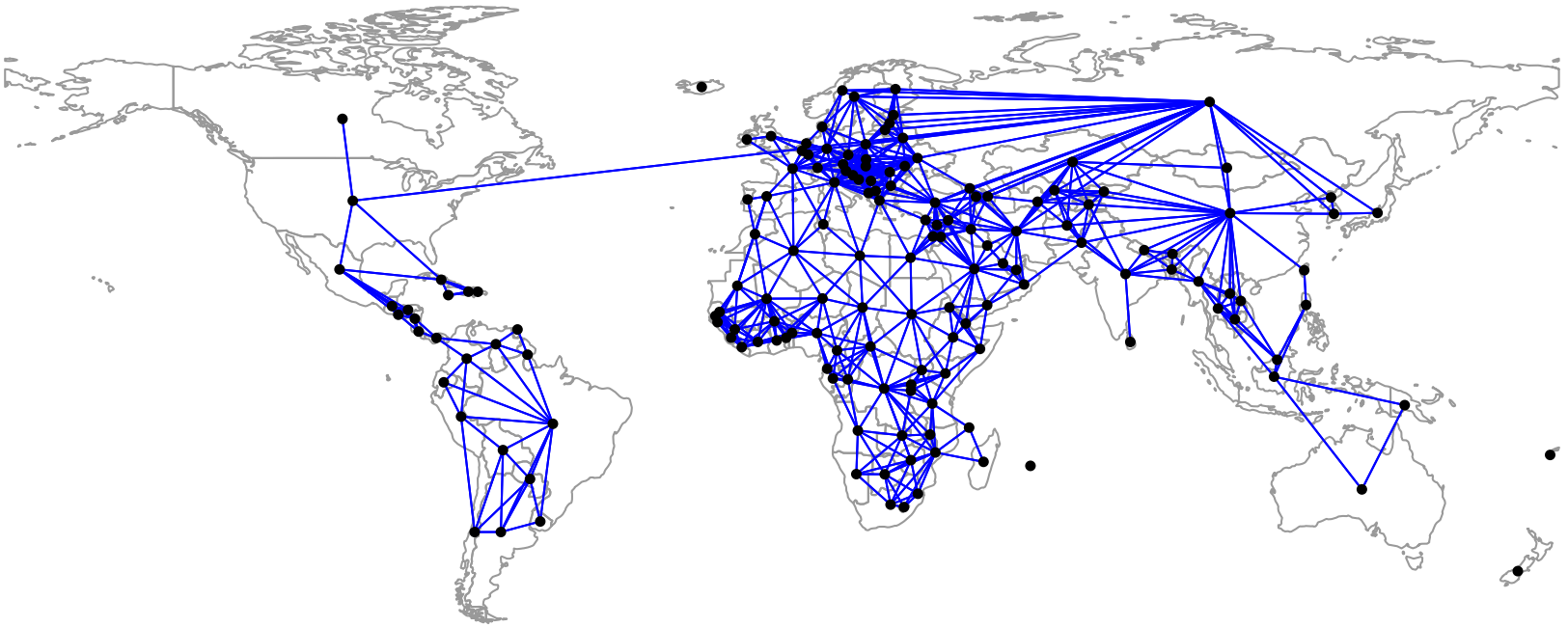
- Neighborless units

Lists with no-neighbor areas are problematic for the estimation of spatial weights.

- Should the weight representation of the empty set be a numeric zero or a missing value?
- This choice, and the resulting size of n, is highly consequential for tests of spatial autocorrelation.

Areal Contiguity II: Visualization of Connections

Figure: Contiguity neighbors with 500 km snap distance

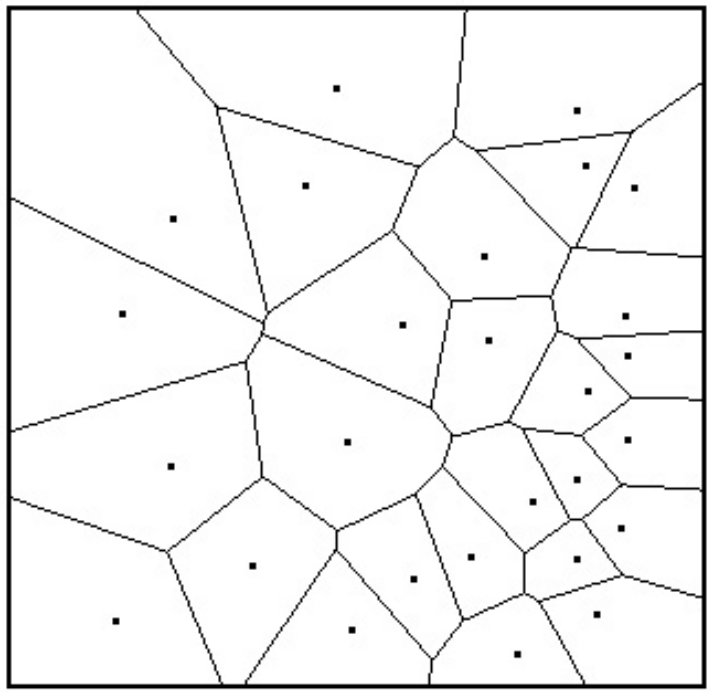


Areal Contiguity II: Polygons

- While intuitive, polygon contiguity is not always appropriate or feasible.
- When the spatial units consist of points, such as cities or event locations, aggregation into polygons is often undesirable due to the Modifiable Areal Unit Problem (MAUP) (Openshaw and Taylor 1979, 1981).
  - Aggregation of point data is only meaningful if the underlying phenomenon is homogeneous across space.
  - Otherwise, any aggregation scheme which does not account for heterogeneity and structural instability will be misleading.
  - Furthermore, the level of aggregation affects the magnitude of various measures of association, such as autocorrelation coefficients and estimated regression parameters.
  - The MAUP is closely conceptually similar to the ecological fallacy problem (King 1997).

Areal Contiguity II: Polygons

- Another approach is to draw polygons around each point by spatial tessellation, as in Voronoi or Dirichlet diagrams.
- But here, notions of boundary locations, length and area are largely artificial constructs, determined by the particular tessellation algorithm used.



Interpoint Distance Neighbors I: Minimum Distance Neighbors ($\mathbf{W}_{MDN}$)

Neighbors of unit x_i defined by interpoint distance:

- Lower bound: 0
- Upper bound: $\max_{i=1}^{n} \left(\min_{j \neq i}^{n-1} d(x_i, x_j) \right)$

Advantage

- No neighborless units

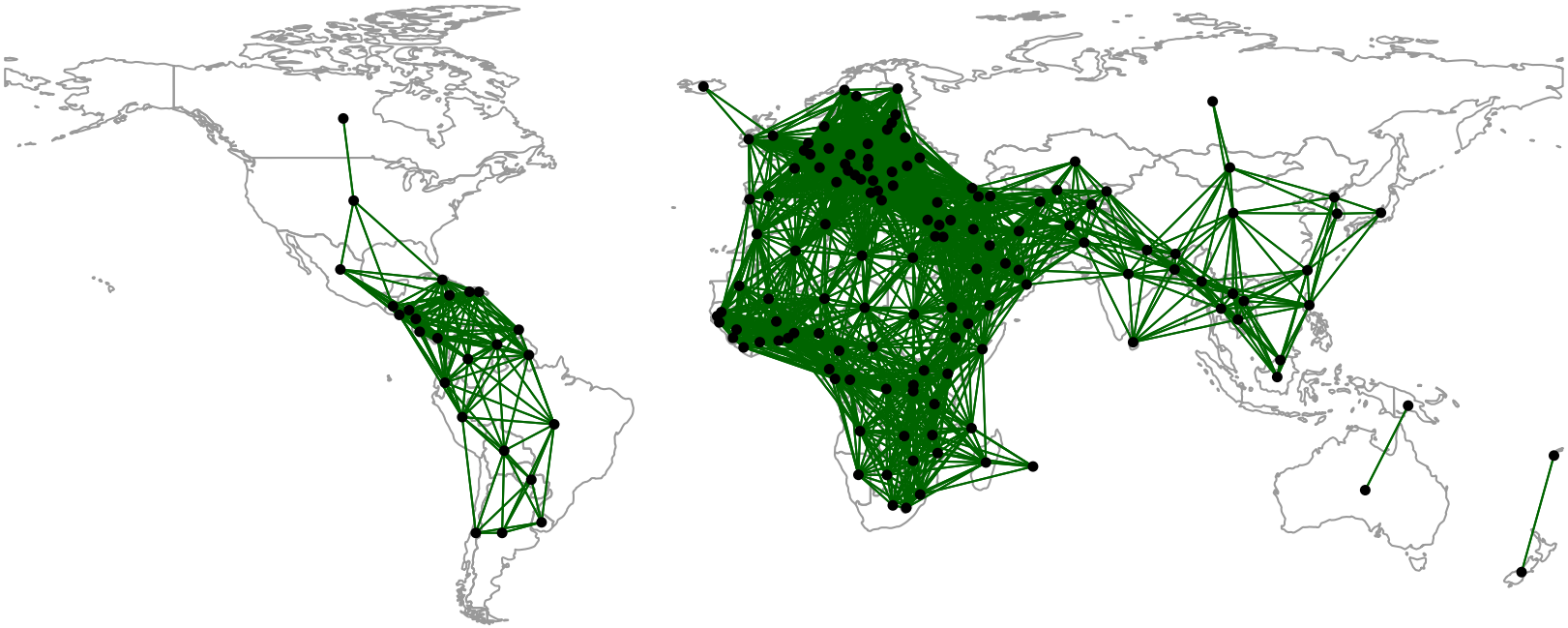
Disadvantage

- Inefficient for irregularly-spaced data
- Potentially high number of politically irrelevant connections

Choice of points (centroids vs. capital cities) is potentially quite significant and requires theoretical justification

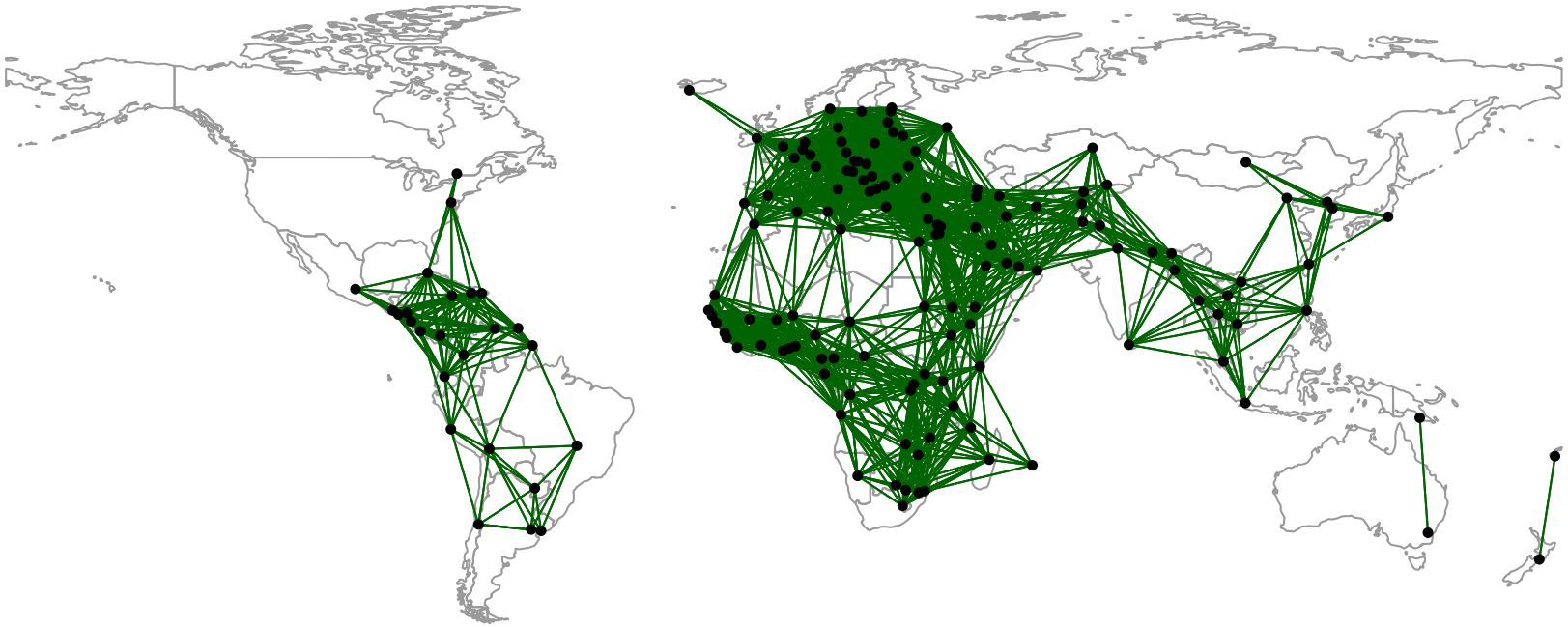
Interpoint Distance Neighbors I: Visualization of Connections

Figure: Minimum distance neighbors (polygon centroids)



Interpoint Distance Neighbors I: Visualization of Connections

Figure: Minimum distance neighbors (capital cities)



Interpoint Distance Neighbors II: k Nearest Neighbors ($\mathbf{W}_{KNN}$)

Neighbors of unit x_i defined by user-defined parameter k. x_j is a neighbor of x_i if $x_j \in N_k x_i$, where $N_k x_i$ are the k nearest neighbors of x_i.

Advantage

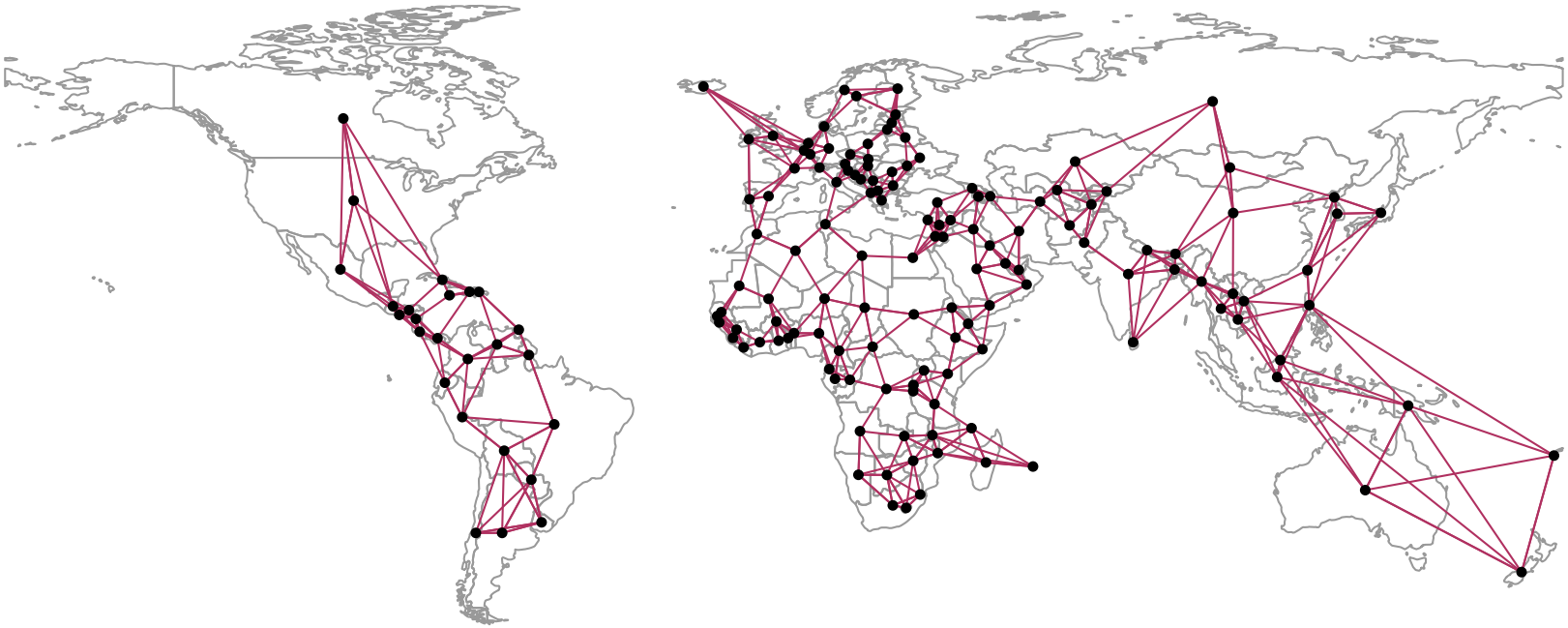
- No neighborless units
- Less noisy than $\mathbf{W}_{MDN}$

Disadvantage

- Parameter selection may not reflect 'true' level of connectedness or isolation

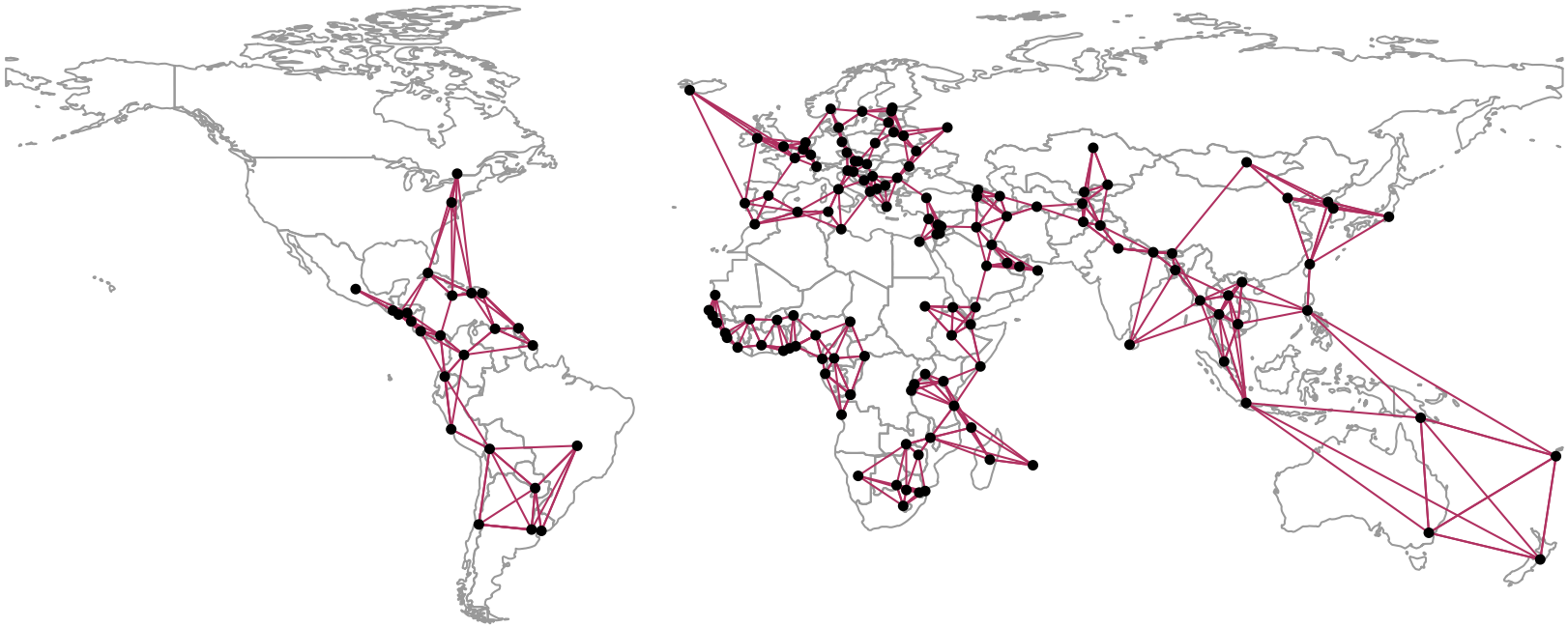
Interpoint Distance Neighbors II: Visualization of Connections

Figure: $k = 4$ Nearest Neighbors (polygon centroids)



Interpoint Distance Neighbors II: Visualization of Connections

Figure: $k = 4$ Nearest Neighbors (capital cities)



Graph-Based Neighbors: Sphere of Influence Neighbors ($\mathbf{W}_{SOI}$)

For each point $x \in S = \{x_1, \ldots, x_n\}$,

- Let $r_i = \min_{k \neq i} d(x_i, x_k)$.
- Let C_i be a circle of radius r_i, centered at x_i.

Points x_i and x_j are neighbors whenever C_i and C_j intersect in exactly two points.

Advantage

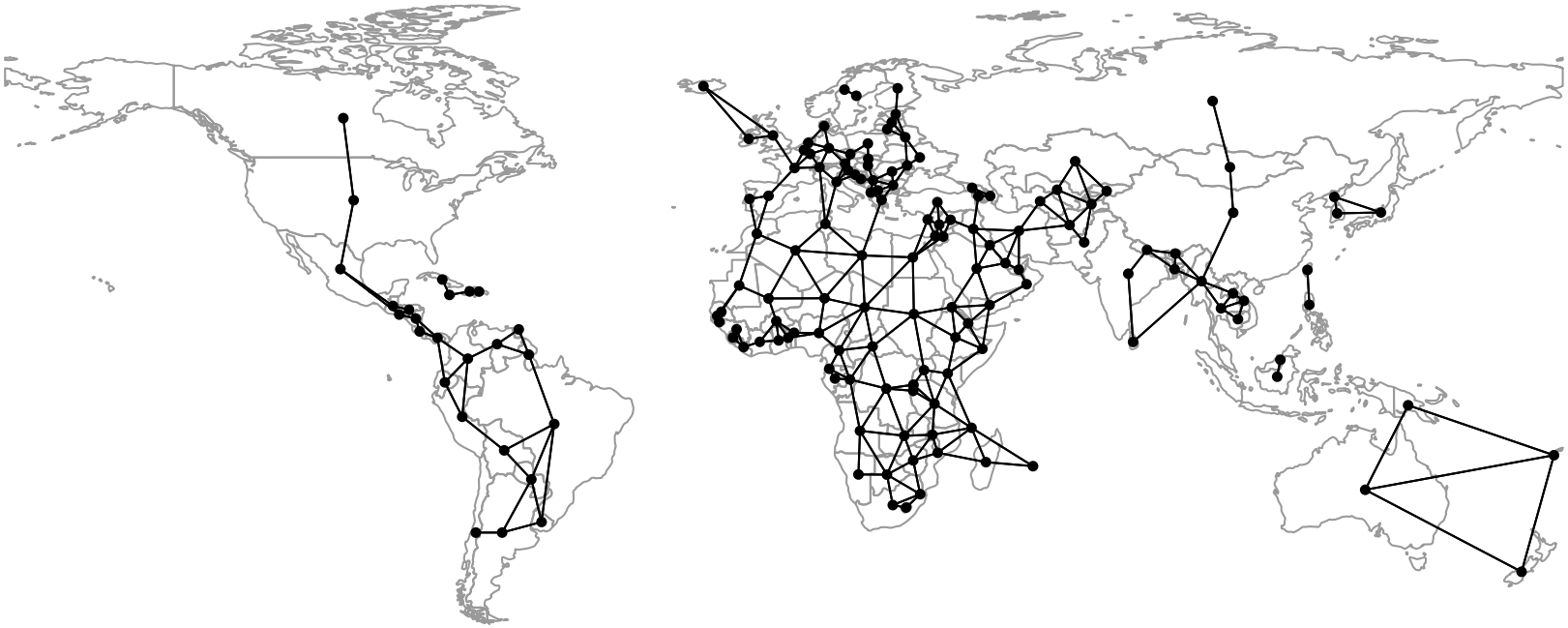
- No neighborless units
- Less noisy than $\mathbf{W}_{MDN}$
- Less arbitrary than $\mathbf{W}_{KNN}$

Disadvantage

- Uses Euclidean, not Great Circle Distances

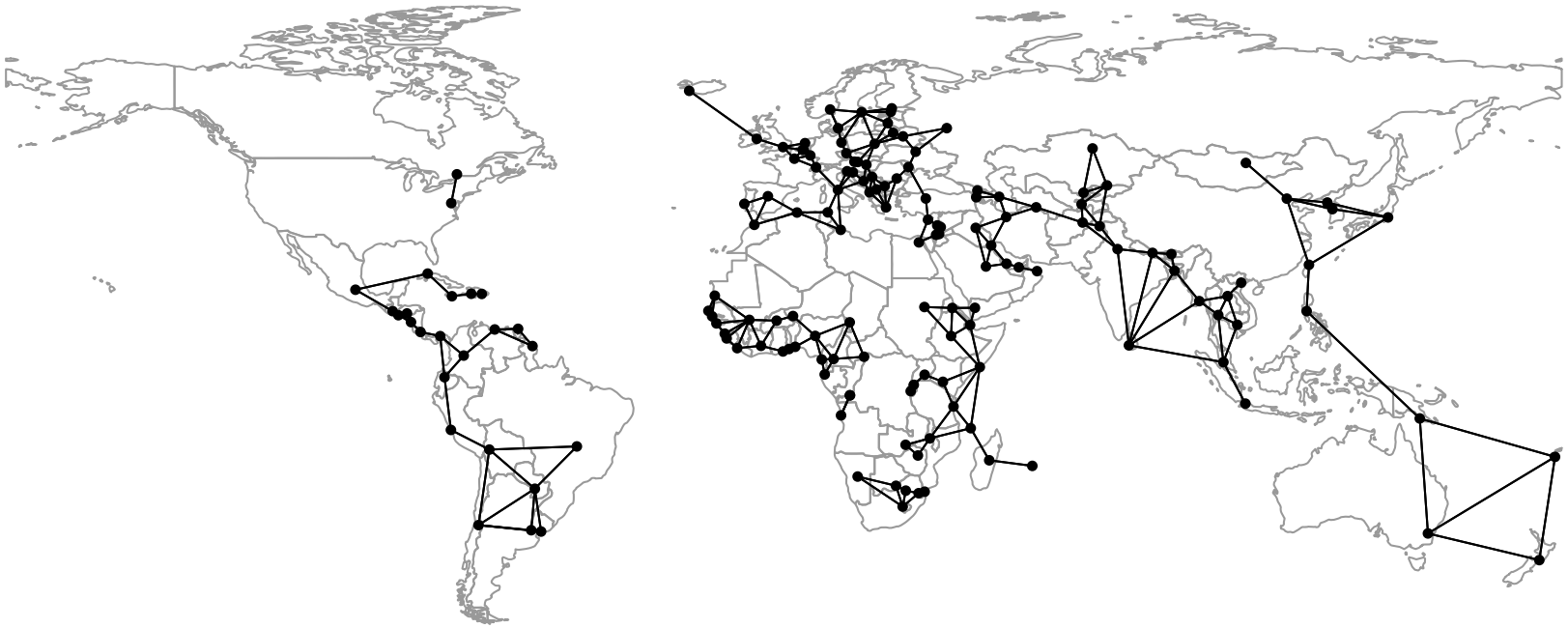
Graph-Based Neighbors: Visualization of Connections

Figure: Sphere of Influence Neighbors (polygon centroids)



Graph-Based Neighbors: Visualization of Connections

Figure: Sphere of Influence Neighbors (capital cities)



Application: Democratic Diffusion

Changes of political regime modeled as a first-order Markov chain process with the transition matrix

$$\mathbf{K} = \begin{bmatrix} Pr(y_{i,t} = 0 | y_{i,t-1} = 0) & Pr(y_{i,t} = 1 | y_{i,t-1} = 0) \\ Pr(y_{i,t} = 0 | y_{i,t-1} = 1) & Pr(y_{i,t} = 1 | y_{i,t-1} = 1) \end{bmatrix}$$

where $y_{i,t} = 1$ if an (A)utocratic regime exists in country i at time t, and $y_{i,t} = 0$ if the regime is (D)emocratic.
. . . in other words:

$$\mathbf{K} = \begin{bmatrix} Pr(D \rightarrow D) & Pr(D \rightarrow A) \\ Pr(A \rightarrow D) & Pr(A \rightarrow A) \end{bmatrix}$$

Estimation

Conditional transition probabilities are estimated by a probit link:

$$Pr(y_{i,t} = 1 | y_{i,t-1}, \mathbf{x}_{i,t}) = \Phi[\mathbf{x}_{i,t}^T \beta + y_{i,t-1} \mathbf{x}_{i,t}^T \alpha]$$

Previous uses:

- Takeshi Amemiya, *Advanced Econometrics* (Cambridge, MA: Harvard University Press, 1985)
- Adam Przeworski and Fernando Limongi, "Modernization: Theories and Facts," *World Politics* 49 (1997): 155-83.
- Kristian S. Gleditsch and Michael D. Ward, "Diffusion and the International Context of Democratization," *International Organization* 60 (2006): 911-33.

Equilibrium Effects of Democratic Transition

If a regime transition takes place in country i, what is the change in predicted probability of a regime transition in country j (country i's neighbor)?

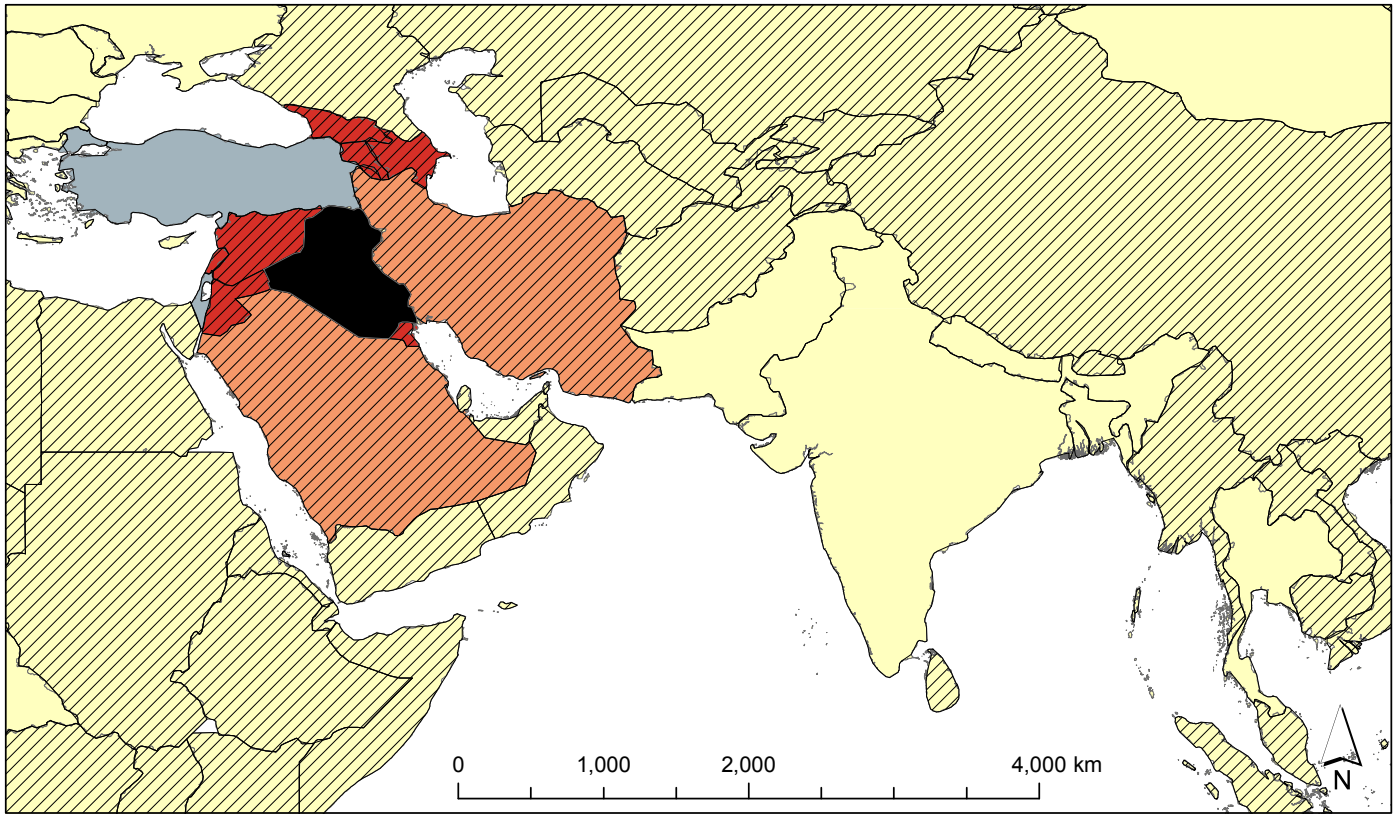
$$\text{QI} = Pr(y_{j,t}|y_{i,t} = y_{i,t-1}) - Pr(y_{j,t}|y_{i,t} \neq y_{i,t-1})$$

where $y_{i,t} = 0$ if country i is a democracy at time t and $y_{i,t} = 1$ if it is an autocracy. All other covariates are held constant.

Illustrative cases

- Iraq transitions from autocracy to democracy.
- Russia transitions from democracy to autocracy.

Iraq's democratization and regional regime stability



Contiguity + 500 km

Iraq transitions from autocracy to democracy (1998 data)

Monte Carlo simulation (1,000 runs)

Iraq

Regime Type

Democracy

Autocracy

Change in Transition Probability

-0.05- -0.025

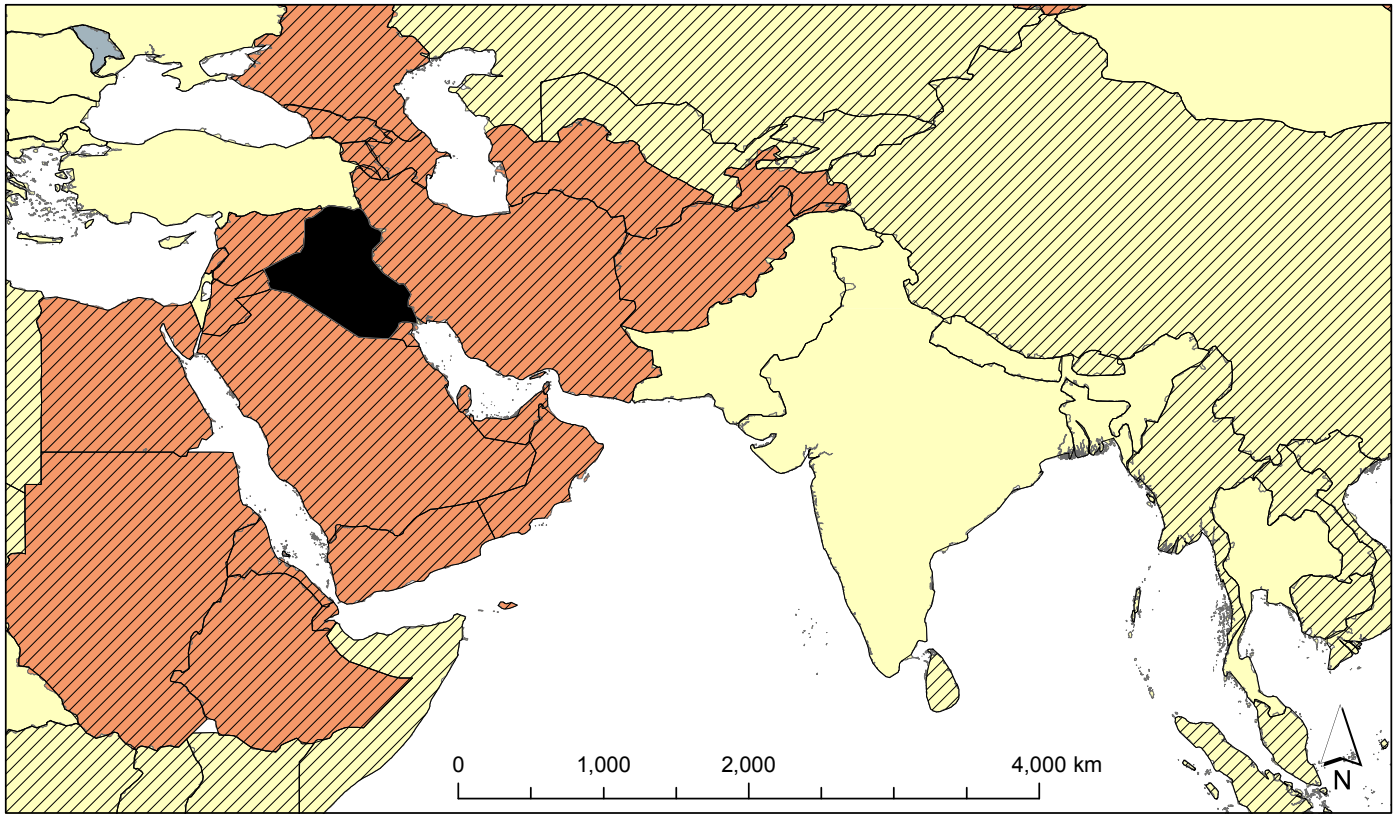
-0.025 - -0.001

0

0.001 - 0.025

0.025 - 0.05

Iraq's democratization and regional regime stability



Minimum Distance

Iraq transitions from autocracy to democracy (1998 data)

Monte Carlo simulation (1,000 runs)

Iraq

Regime Type

Democracy

Autocracy

Change in Transition Probability

-0.05- -0.025

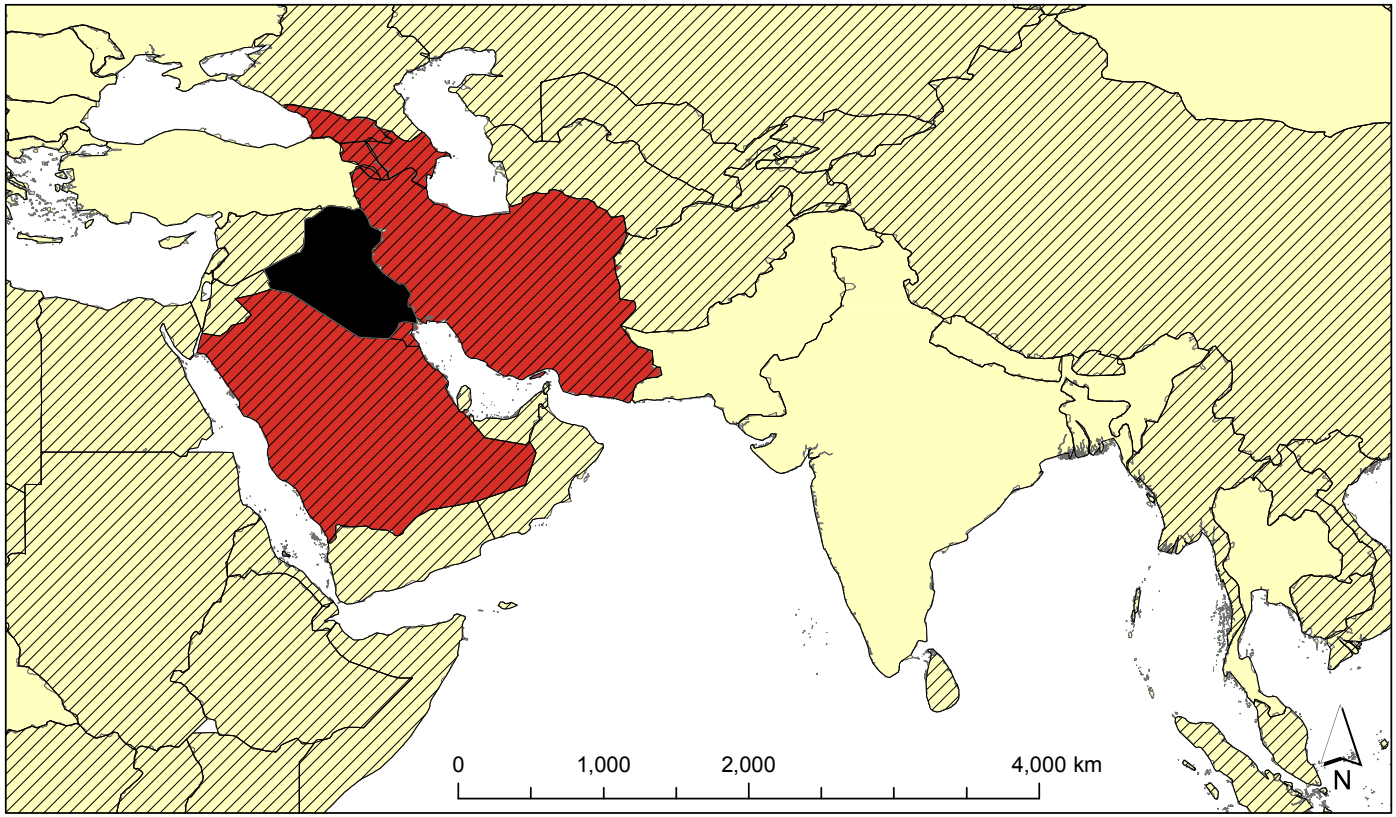
-0.025 - -0.001

0

0.001 - 0.025

0.025 - 0.05

Iraq's democratization and regional regime stability



k = 4 Nearest Neighbors

Iraq transitions from autocracy to democracy (1998 data)

Monte Carlo simulation (1,000 runs)

Iraq

Regime Type

Democracy

Autocracy

Change in Transition Probability

-0.05- -0.025

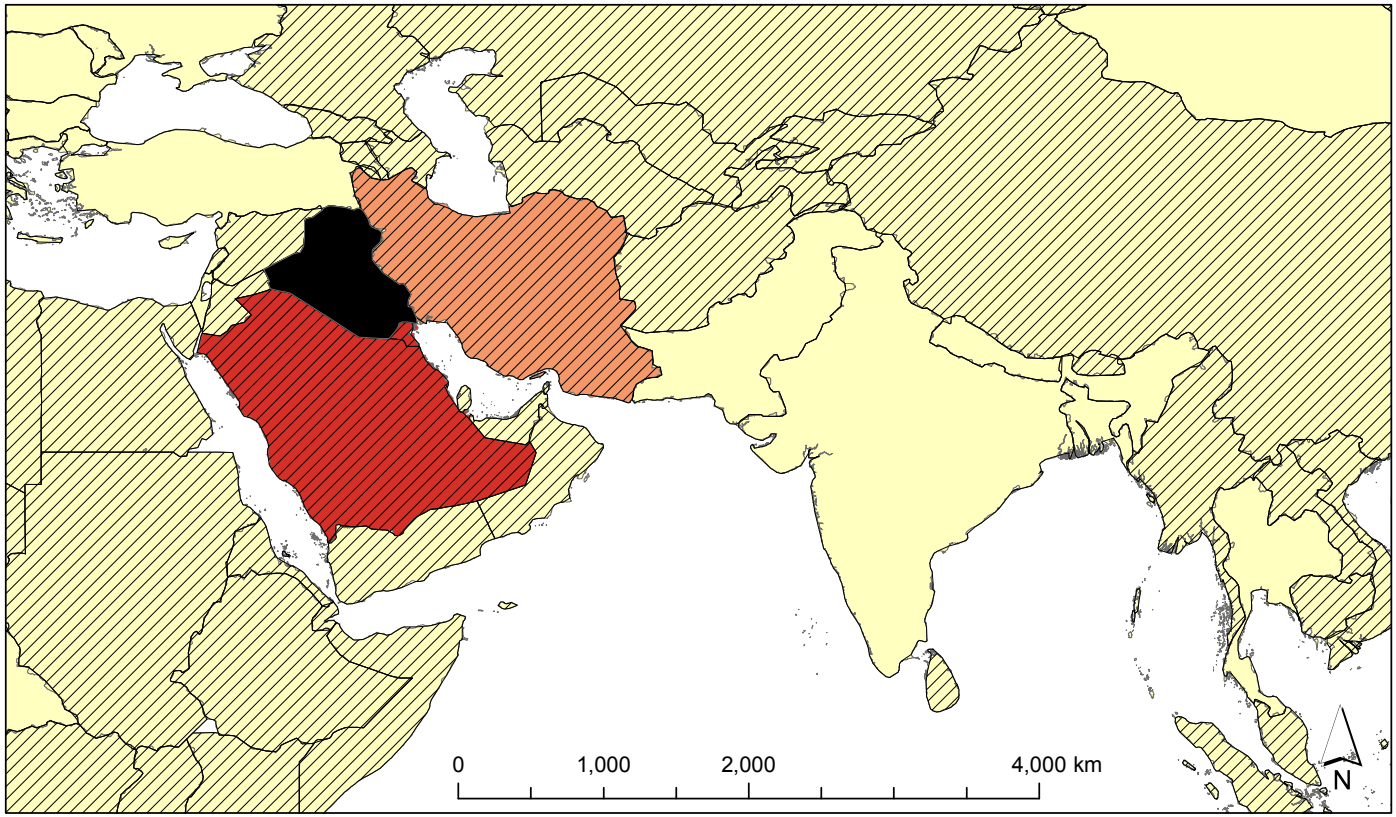
-0.025 - -0.001

0

0.001 - 0.025

0.025 - 0.05

Iraq's democratization and regional regime stability



Sphere of Influence

Iraq transitions from autocracy to democracy (1998 data)

Monte Carlo simulation (1,000 runs)

Iraq

Regime Type

Democracy

Autocracy

Change in Transition Probability

-0.05- -0.025

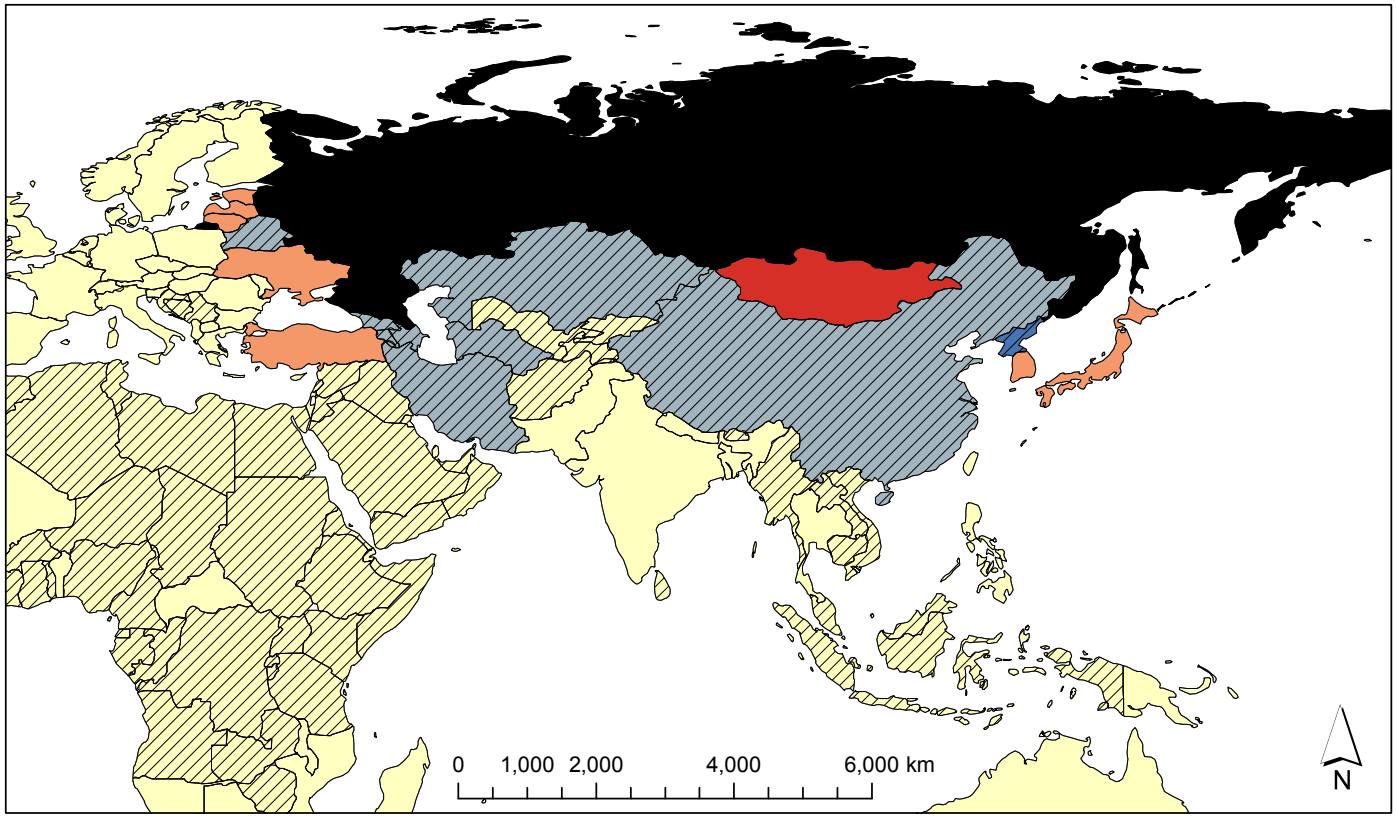
-0.025 - -0.001

0

0.001 - 0.025

0.025 - 0.05

Russia's autocratization and regional regime stability



Contiguity + 500 km

Russia transitions from democracy to autocracy (1998 data)

Monte Carlo simulation (1,000 runs)

Russia

Regime Type

Democracy

Autocracy

Change in Transition Probability

-0.05- -0.025

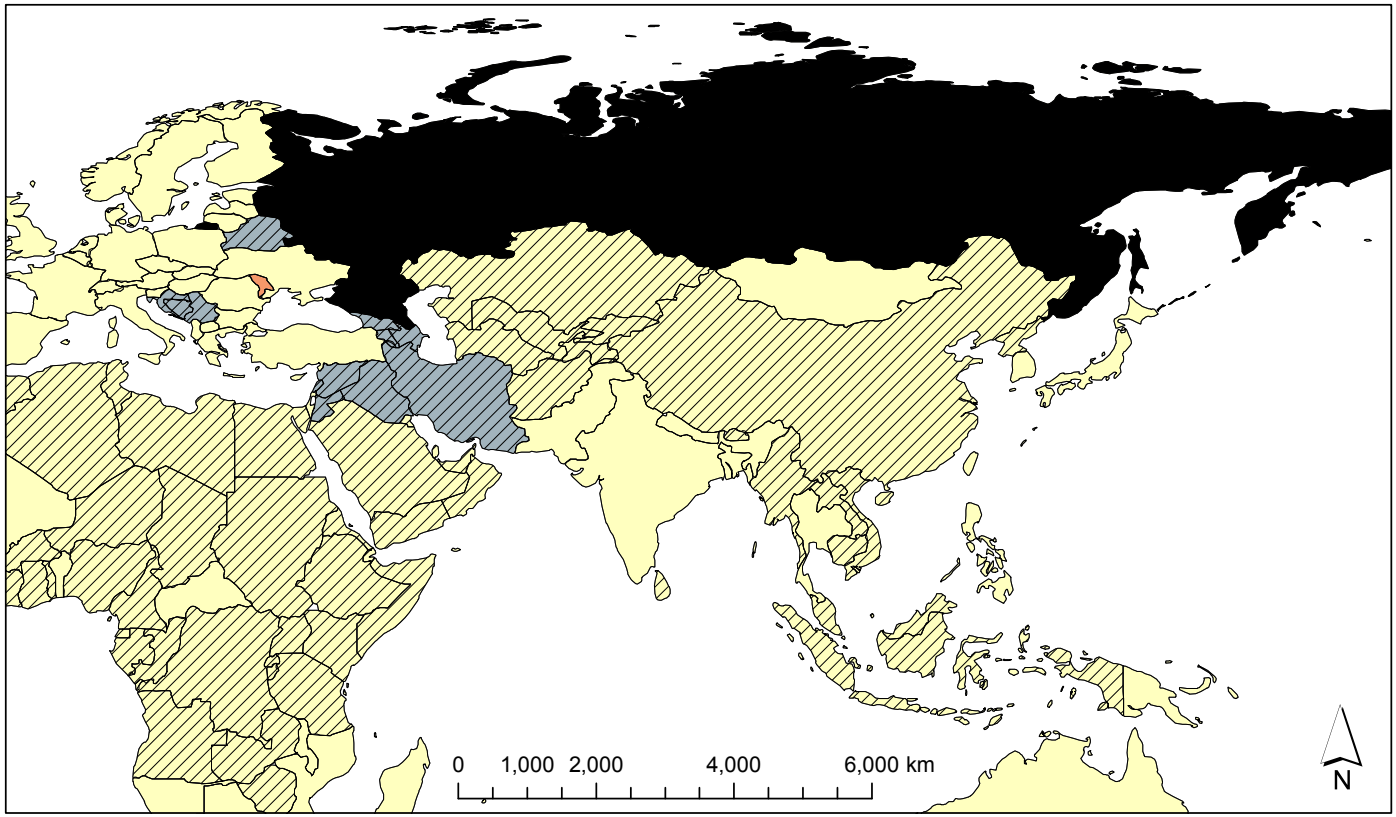
-0.025 - -0.001

0

0.001 - 0.025

0.025 - 0.05

Russia's autocratization and regional regime stability



Minimum Distance

Russia transitions from democracy to autocracy (1998 data)

Monte Carlo simulation (1,000 runs)

Russia

Regime Type

Democracy

Autocracy

Change in Transition Probability

-0.05- -0.025

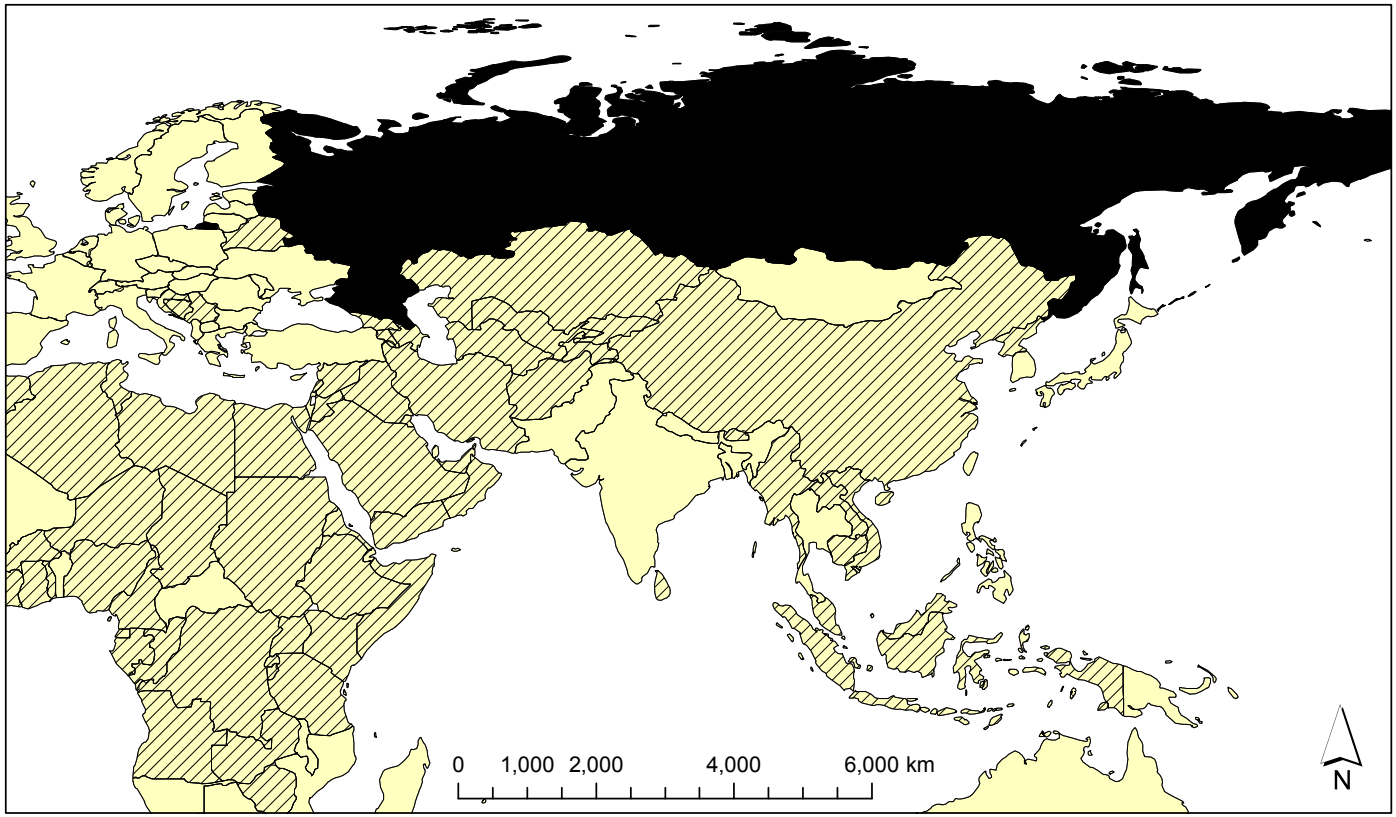
-0.025 - -0.001

0

0.001 - 0.025

0.025 - 0.05

Russia's autocratization and regional regime stability



k = 4 Nearest Neighbors

Russia transitions from democracy to autocracy (1998 data)

Monte Carlo simulation (1,000 runs)

Russia

Regime Type

Democracy

Autocracy

Change in Transition Probability

-0.05- -0.025

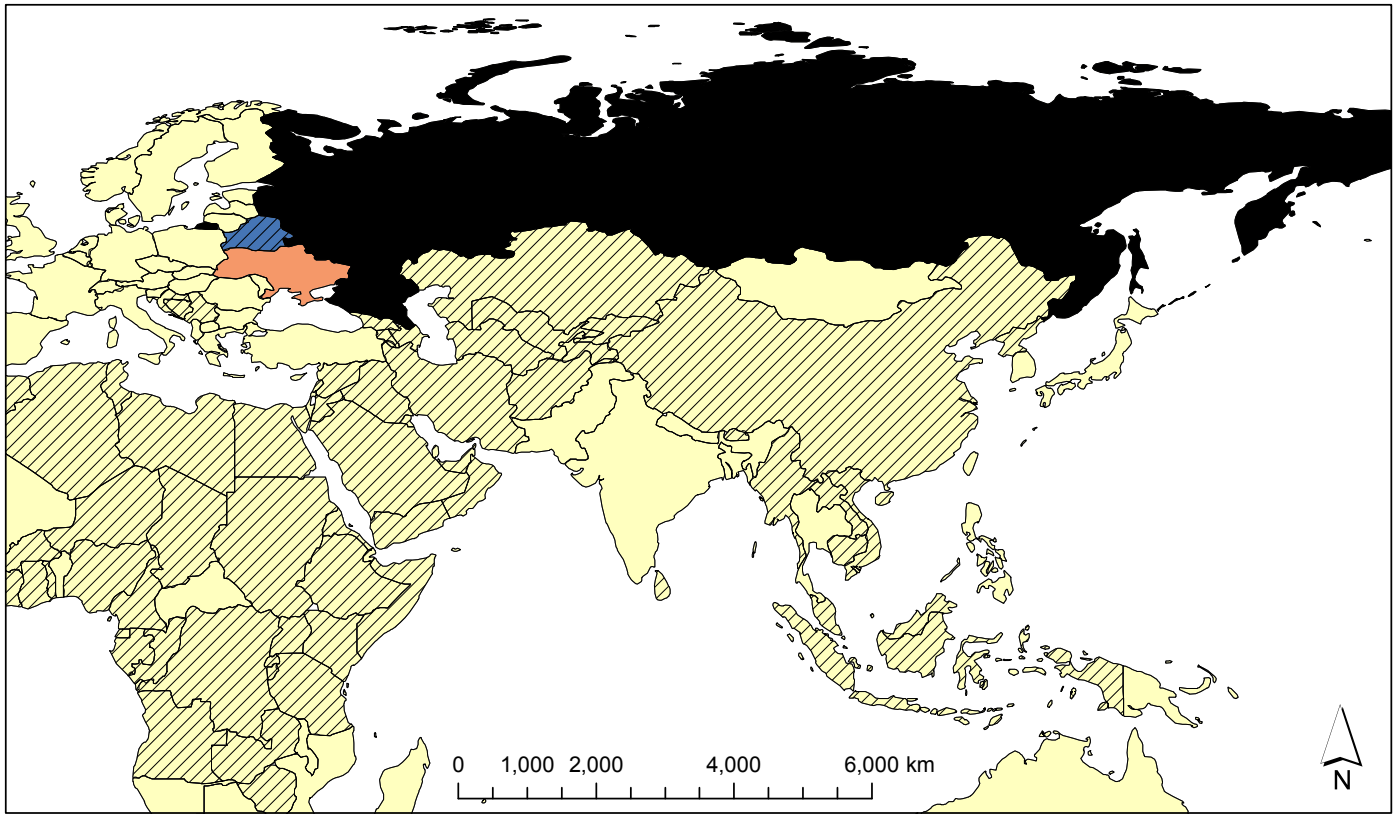
-0.025 - -0.001

0

0.001 - 0.025

0.025 - 0.05

Russia's autocratization and regional regime stability



Sphere of Influence

Russia transitions from democracy to autocracy (1998 data)

Monte Carlo simulation (1,000 runs)

Russia

Regime Type

Democracy

Autocracy

Change in Transition Probability

-0.05- -0.025

-0.025 - -0.001

0

0.001 - 0.025

0.025 - 0.05

Network neighbors

The structure of spatial dependence can be non-geographic. Any theoretically-relevant dyadic relationship can form the basis of connectivity.

- **Individual level**: friendship, frequency of communication, citations, kinship.
- **Organizational level**: market competition, joint enterprises, personnel exchanges.
- **International level**: alliance relationship, trade flows, joint organizational membership, diplomatic contacts, cultural exchanges, migration flows.

From Connections to Weights

- Once a definition of connectivity is made, one must translate binary indicators into weights, which will form the elements w_{ij} of matrix $\mathbf{W}$.
- A plethora of options exist: inverse distance (IDW), negative exponentials of distance, length of shared boundary, relative area, accessibility...
- The resulting weights will often be asymmetric, unless the study region is a regular lattice.

From Connections to Weights

- The rows of **W** are often row-standardized, so that $\sum_{j=1}^{n} w_{ij} = 1$
- There is no mathematical or statistical requirement for this, but row standardization facilitates interpretation of lagged variables as a weighted average of neighboring values of some variable.
- This is not always desirable, however.
- Row-standardization also implies competition among neighbors: the fewer the neighbors, the stronger their individual influence on i.
- Further, when weights are based on some measure of distance decay (ie: IDW), scaling the rows to sum to one results in a loss of that interpretation.
- **Bottom line**: the weights should bear a direct relation to one's theoretical conceptualization of the structure of dependence.

Sparse vs. Dense Matrices

Sparsity carries a number of substantive and computational advantages:

- Dense matrices are noisy and contain a potentially large number of irrelevant connections.
- Dense matrices will bias downward indirect effects of a change in observation j (the individual weights of non-zero entries in row-standardized weight matrices will be smaller).
- Dense matrices can be computationally intensive to the point that even simple matrix operations are infeasible.

Sparse vs. Dense Matrices

Consider the following example with 2000 U.S. Census data:

Tracts

$n = 65,443$

31.90 GB of storage required for dense matrix, .01 GB for sparse matrix.

Block Groups

$n = 208,790$

324.80 GB of storage required for dense matrix, .03 GB for sparse matrix.

Blocks

$n = 8,205,582$

501,659.33 GB of storage required for dense matrix, 1.10 GB for sparse.

Here, dense and sparse matrices have n^2 and $6/n$ nonzero elements, respectively. For spatially random data on a plane, each unit will have an average of 6 contiguity neighbors (LeSage and Pace 2009).

Ordering of Weights Matrix

Ordering of rows and columns matters greatly for computation times.

- Consider an $n \times n$ permutation matrix $\mathbf{P}$, which has exactly one entry 1 in each row and each column and 0's elsewhere. Each permutation matrix can produce a reordered weights matrix $\mathbf{W}_P$, by the operation $\mathbf{W}_P = \mathbf{P}\mathbf{W}\mathbf{P}'$.
- Note that $\mathbf{P}^{-1} = \mathbf{P}'$, $|\mathbf{P}| = 1$ and $|\mathbf{P}(\mathbf{I}_n - \rho\mathbf{W})\mathbf{P}'| = |\mathbf{P}||\mathbf{I}_n - \rho\mathbf{W}||\mathbf{P}'| = |\mathbf{I}_n - \rho\mathbf{W}| = |\mathbf{I}_n - \rho\mathbf{P}\mathbf{W}\mathbf{P}'|$
- Thanks to these properties, log-determinant calculation and other matrix operations will not be affected by the reordering of $\mathbf{W}$.
- But computation times for these operations are affected.

Ordering of Weights Matrix

Efficiency is increased if ordering is **geographic** (north-south or east-west)

- This ordering concentrates nonzero elements around the diagonal, which reduces the bandwidth of a matrix ($max|i-j|$ for nonzero elements).
- For a sample of 62,226 U.S. Census Tracts, calculation of a single log-determinant requires over 12 GB of memory for a randomly ordered weights matrix, making calculation infeasible on most machines.
- The same operation takes less than a minute for a geographically-ordered matrix (LeSage and Pace 2009).

Examples in R

Switch to R tutorial script. Section 3.

Outline

Point Pattern Analysis: Aerial Bombardment

- During World War II, Germany launched 1,358 V-2 Rockets at London.
- The V-2's speed and trajectory made it invulnerable to anti-aircraft guns and fighters.
- But its guidance systems were thought to be too primitive to hit specific targets.
- After the strikes began in 1944, bomb damage maps were interpreted by some analysts as showing that impact sites were clustered.
- This evidence appeared to contradict existing intelligence on the V-2 program.
- If the rocket strikes were spatially clustered, the guidance systems must have been more advanced than previously thought.

Point Pattern Analysis: Aerial Bombardment

Figure: Distribution of V-2 Rocket Strikes on Central London, 1944

Point Pattern Analysis: Aerial Bombardment

- R.D. Clarke (1946) decided to apply a statistical test to assess whether any support could be found for the clustering hypothesis.
- He selected an area of 144 km^2 in south London, which he divided into 576 squares of $1/4$ km^2.
- For each square, Clark recorded the total number of observed bomb hits. There were 537 total in the study area.
- He then recorded the number of squares with $k = 1, 2, 3, \ldots$ hits.
- The expected number of squares with k hits was derived from the Poisson distribution $\sum_{k=1}^{n} \frac{e^{-\lambda}\lambda^k}{k!}$, with $\lambda = \frac{537}{576}$ and $n = 576$.

Point Pattern Analysis: Aerial Bombardment

No. of bombs per square	Expected	Observed
1	226.74	229
2	211.39	211
3	98.54	93
4	7.14	7
5+	1.57	1
	$\chi^2 = 1.17, p = 0.88$	

- It is clear from the cross-tabulation that the distribution of V-2 hits conforms quite closely to the Poisson distribution.
- The occurrence of clustering would have been reflected in an excess number of squares with either a high number of bombs or none at all, and fewer squares in the intermediate classes.
- The closeness of fit suggested that V-2 impact sites were random, rather than clustered.

Point Pattern Processes

Point patterns have first- and second- order properties:

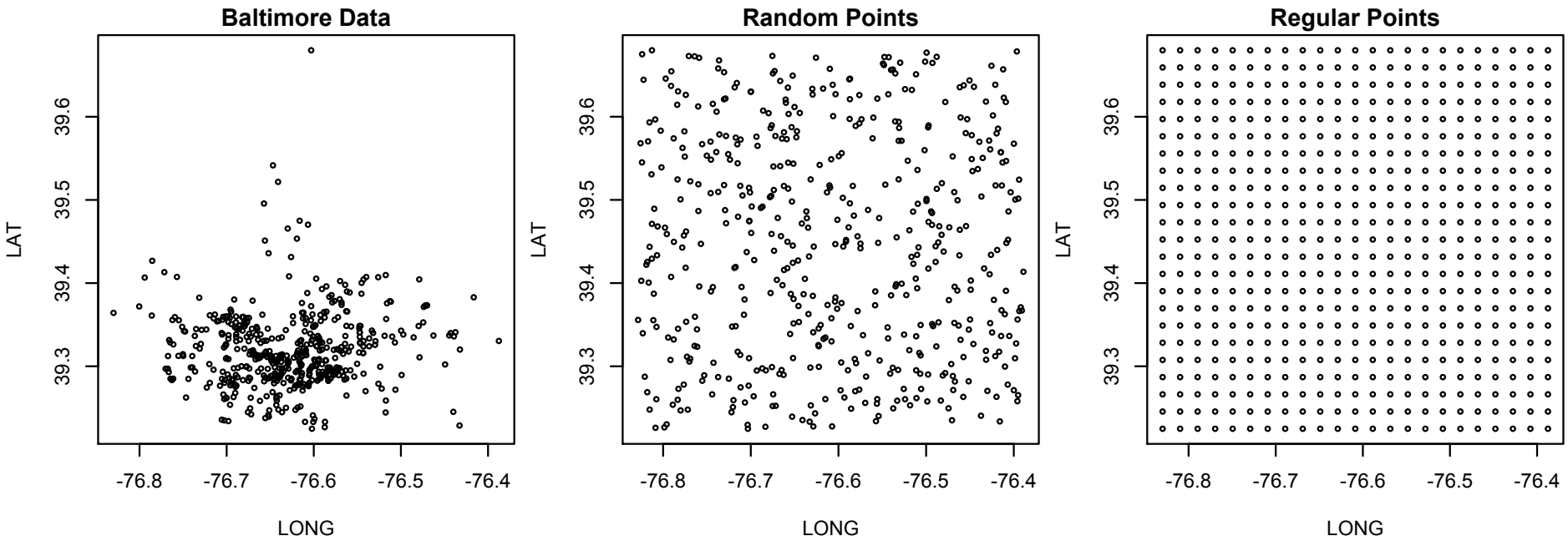
1. First-order properties measure the distribution of events in a study region: intensity and spatial density.
2. Second-order properties measure the tendency of events to appear clustered, independently, or regularly-spaced.

Point Pattern Processes: Complete Spatial Randomness

- The most basic test which can be performed is that of Complete Spatial Randomness (CSR). Under CSR, events are distributed independently and uniformly over a study area.
- A point process which is CSR point process is formally defined as a homogeneous Poisson process (HPP).
  - Under HPP, the location of one point in space does not affect the probabilities of other points' appearing nearby. The intensity of the point process in area A is a constant $\lambda(y) = \lambda > 0, \quad \forall y \in A$.
- A generalization of HPP which allows for non-constant intensity $\lambda(y)$ is called an inhomogeneous Poisson process (IPP).

Point Pattern Processes: Complete Spatial Randomness

- Let's explore conformity to CSR among three point patterns: (1) real data on crime locations in Baltimore, (2) points drawn from uniform distribution over the same study area, (3) regularly-spaced point pattern.



Point Pattern Processes: $\mathcal{G}$ Function

- The $\mathcal{G}$ Function measures the distribution of distances from an arbitrary event to its nearest neighbors.

$$\hat{\mathcal{G}}(r) = \frac{\sum_{i=1}^{n} I_i}{n}$$

$$I_i = \begin{cases} 1 & \text{if } d_i \in \{d_i : d_i \leq r, \forall i\} \\ 0 & \text{otherwise} \end{cases}$$

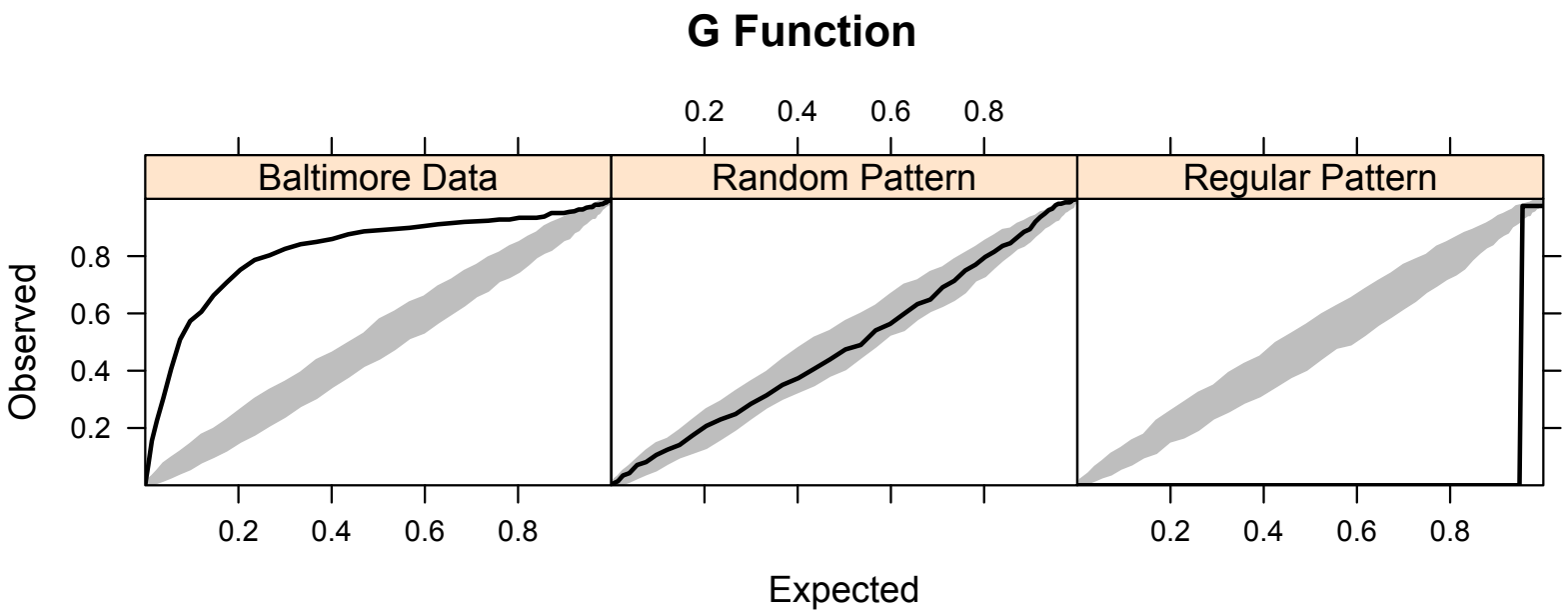
- where $d_i = \min_j\{d_{ij}, \forall j \neq i \in S\}, i = 1, \ldots, n$.
- So, the $\mathcal{G}$ function represents the number of elements in the set of distances up to some threshold r, normalized by the total number of points n in point pattern S.
- Under CSR, the value of the $\mathcal{G}$ function becomes:

$$\mathcal{G}(r) = 1 - e^{\lambda \pi r^2}$$

- where λ is the mean number of events per unit (intensity).

Point Pattern Processes: $\mathcal{G}$ Function

- The comparability of a point process with CSR can be assessed by plotting the empirical function $\hat{\mathcal{G}}(r)$ against the theoretical expectation $\mathcal{G}(r)$.
- For a clustered pattern, observed locations should be closer to each other than expected under CSR. A regular pattern should have higher nearest-neighbor distances than expected under CSR.
- This is shown below for the Baltimore crime locations dataset.



Point Pattern Processes: $\mathcal{F}$ Function

- The $\mathcal{F}$ Function measures the distribution of **all** distances from an arbitrary point k in the plane to the nearest observed event j.

$$\hat{\mathcal{F}}(r) = \frac{\sum_{k=1}^{m} I_k}{m}$$

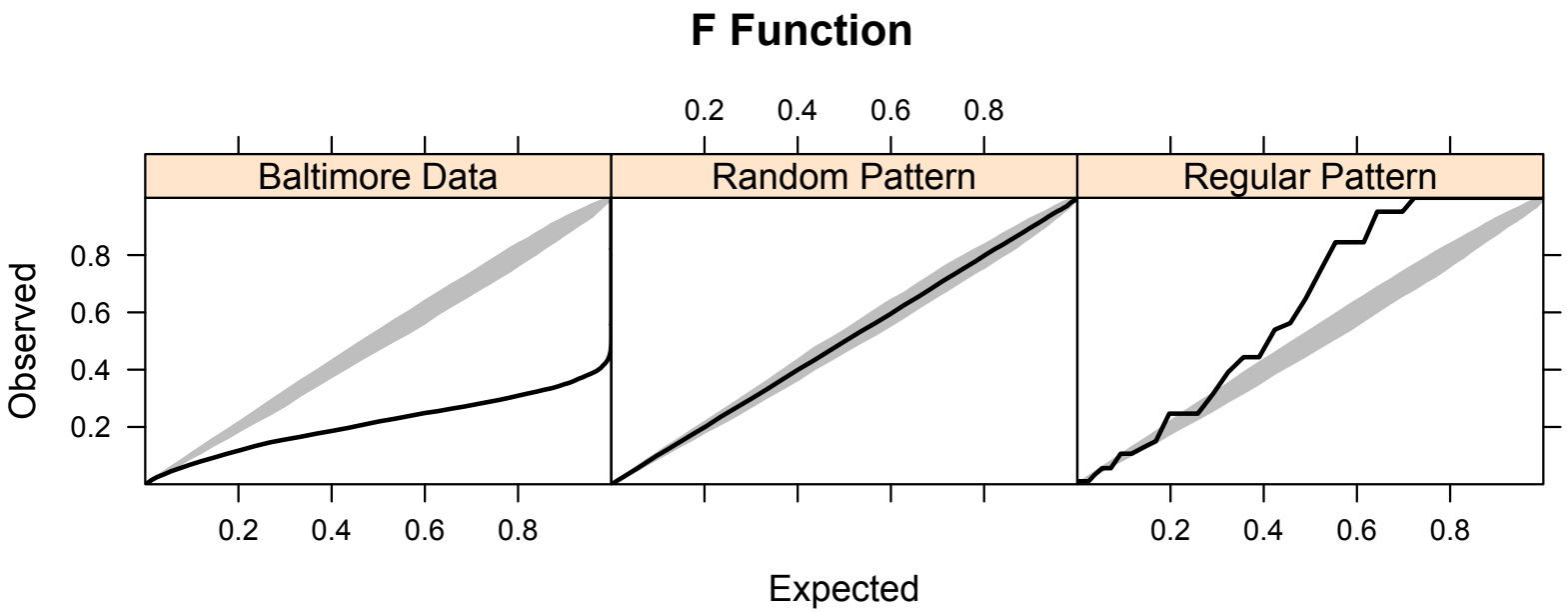
$$I_k = \begin{cases} 1 & \text{if } d_k \in \{d_k : d_k \leq r, \forall k\} \\ 0 & \text{otherwise} \end{cases}$$

- where $d_k = \min_j\{d_{kj}, \forall j \in S\}, k = 1, \ldots, m, \quad j = 1, \ldots, n.$
- Under CSR, the expected value is also

$$\mathcal{F}(r) = 1 - e^{\lambda \pi r^2}$$

Point Pattern Processes: $\mathcal{F}$ Function

- As before, we can plot the empirical function $\hat{\mathcal{F}}(r)$ against its theoretical expectation $\mathcal{F}(r)$.
- For a clustered pattern, observed locations j should be farther away from random points k than expected under CSR. In a regular pattern, random locations should be closer to observed points.
- This is again shown below for the Baltimore crime locations dataset.



Point Pattern Processes: Intensity

- For an HPP point process, intensity is a constant $\lambda(x) = \lambda = \frac{n}{|A|}$, where n is the number of points observed in region A, and $|A|$ is the area of region A.
- For an IPP point process, intensity is non-constant and can be estimated non-parametrically with kernel smoothing (Diggle 1985, Berman and Diggle 1989, Bivand et. al. 2008).

Point Pattern Processes: Kernel Density

- The kernel density estimator is:

$$\hat{\lambda}(x) = \frac{1}{h^2} \sum_{i=1}^{n} \frac{\kappa\left(\frac{||x - x_i||}{h}\right)}{q(||x||)}$$

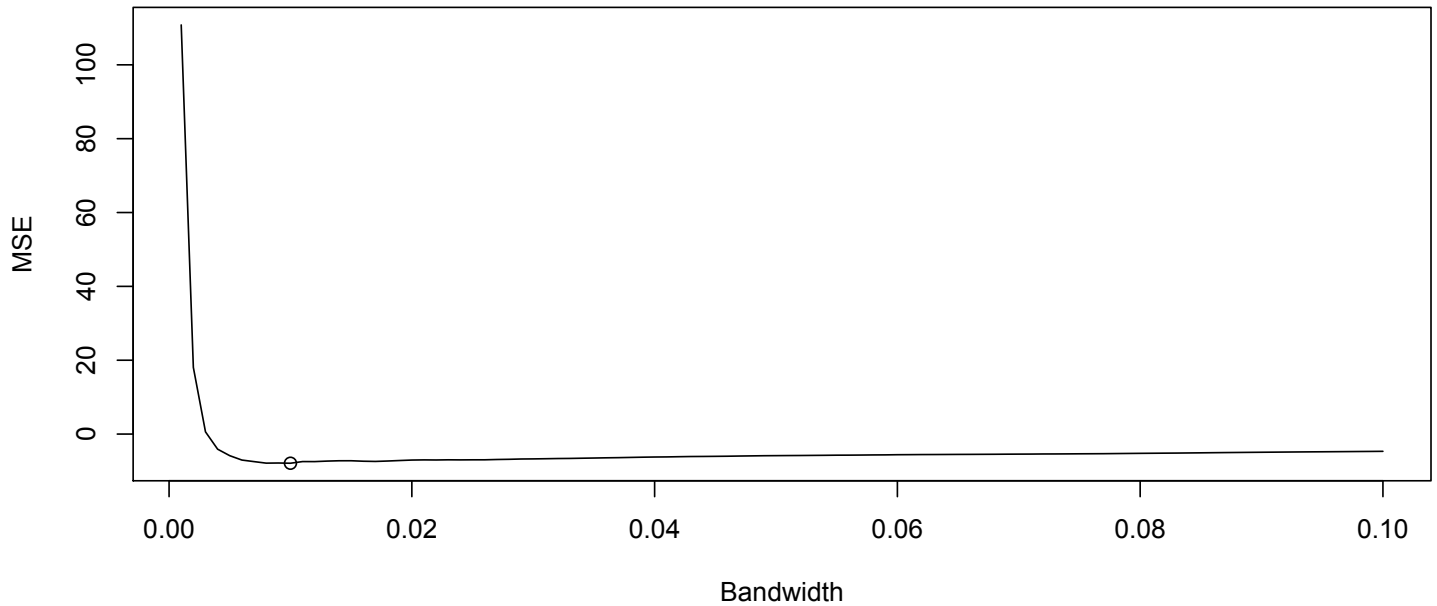
- where $x_i \in \{x_1, \ldots, x_n$ is an observed point, h is the bandwidth, $q(||x||)$ is a border correction to compensate for observations missing due to edge effects, and $\kappa(u)$ is a bivariate and symmetrical kernel function.
- R currently implements a two-dimensional quartic kernel function:

$$\kappa(u) = \begin{cases} \frac{3}{\pi}(1 - ||u||^2)^2 & \text{if } u \in (-1, 1) \\ 0 & \text{otherwise} \end{cases}$$

- where $||u||^2 = u_1^2 + u_2^2$ is the squared norm of point $u = (u_1, u_2)$

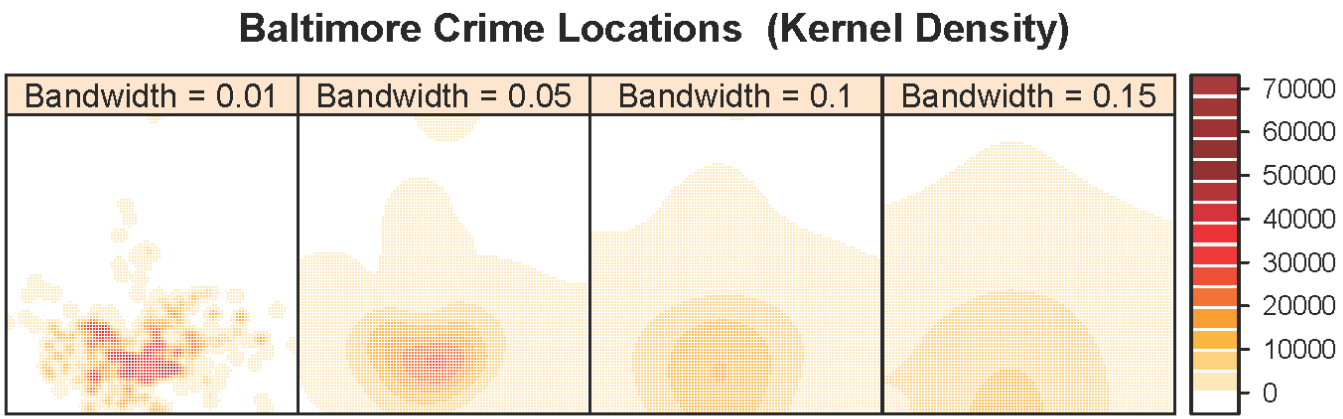
Point Pattern Processes: Kernel Density

- There is no general rule for selecting the bandwidth h, which governs the level of smoothing.
- Small bandwidth $\rightarrow$ spiky map; large bandwidth $\rightarrow$ smooth map.
- Berman and Diggle (1989) propose a criterion based on minimization of mean square error (MSE) of the kernel smoothing estimator.
- The plot below implements this approach for the Baltimore crime dataset. The "optimal" bandwidth here is 0.01.



Point Pattern Processes: Kernel Density

- The plot below shows kernel density estimates for the Baltimore crime locations at different values of the bandwidth h.
- Lighter values indicate greater intensity of the point process.
- Clearly, different bandwidths tell very different stories about the spatial intensity of crime in Baltimore...



Examples in R

Switch to R tutorial script. Section 4.

Outline

Geostatistics

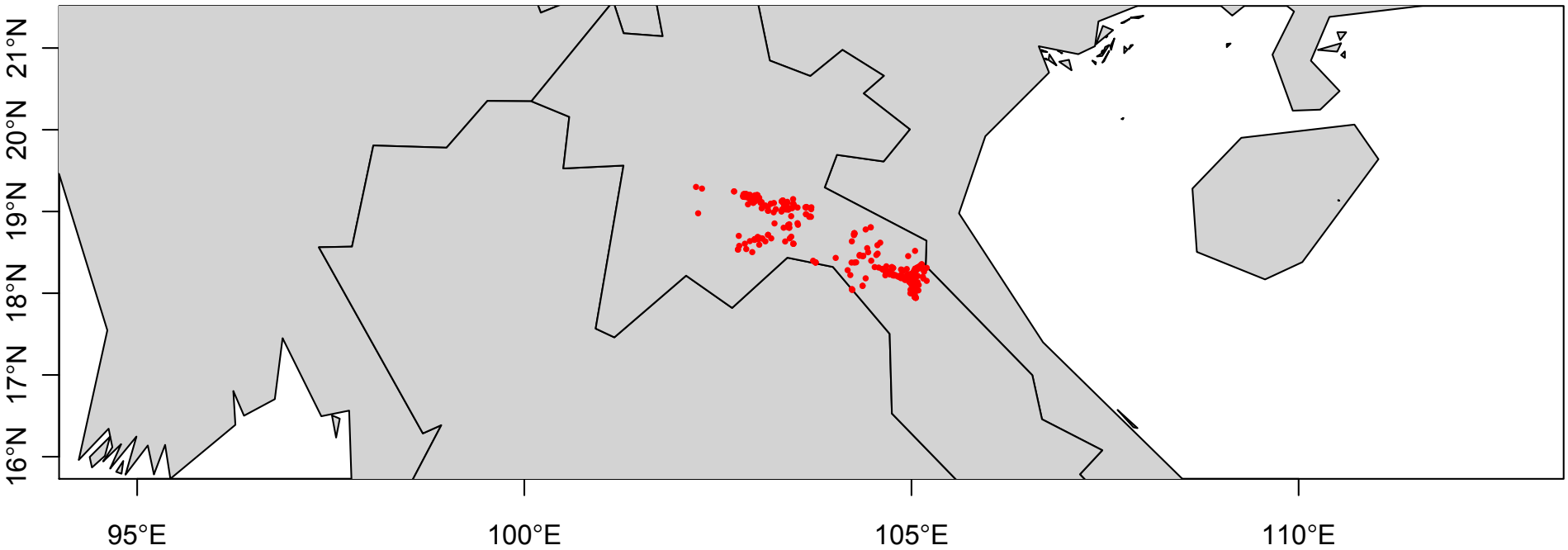
What if the pattern of point locations is not of primary interest? You may wish to...

- determine where new data should be collected,
- identify which observations are spatial outliers,
- perform spatial prediction,
- interpolate missing data from nearby observed locations,
- estimate local averages of spatially autocorrelated variables.

These problems are the domain of a subfield called geostatistics.

Geostatistics

- Let's take a look at some data on U.S. Air Strikes in Laos during the Vietnam War.
- The variable we are interested is `LOAD_LBS`, the payload of each bomb dropped.



Geostatistics: IDW Interpolation

- Spatial interpolation is the prediction of values of attributes at unsampled locations x_0 from existing measurements at x_i.
- This procedure converts a sample of point observations into an alternative representation, such as a contour map or grid.
- One approach to interpolation is to use a locally-weighted average of nearby values.
- Inverse-distance weighted (IDW) interpolation computes one such weighted average:

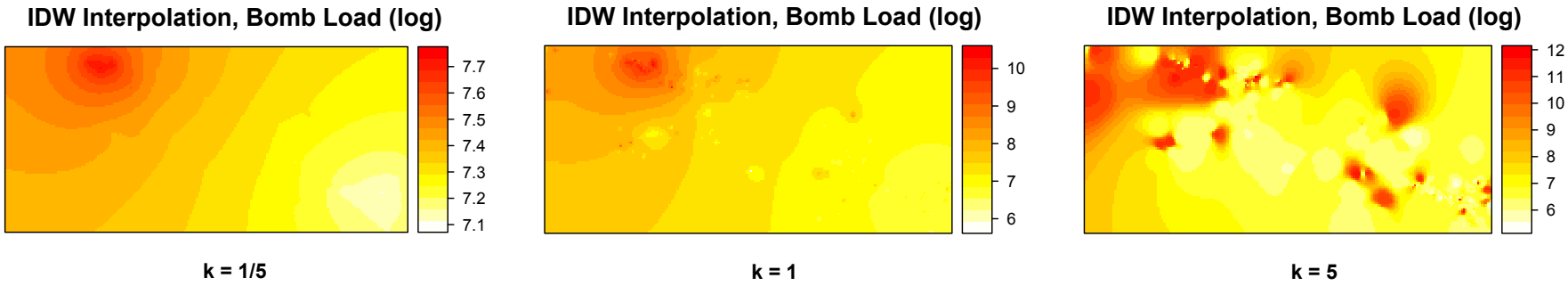
$$\hat{Z}(x_0) = \sum_{i=1}^{n} w_{i0} Z(x_i)$$

where weights w_{ij} are determined according to the distance between points x_i and x_j, and scaled by parameter k.

$$w_{ij} = \frac{1}{d_{ij}^k}$$

Geostatistics: IDW Interpolation

- Predicted values and variance for Laos bombing data are shown below.
- Values of $k > 1$ reduce the relative impact of distant points and produce a peaky map.
- Values of $k < 1$ increase the impact of distant points and produce a smooth map.



Geostatistics: Variogram

- In geostatistics, spatial autocorrelation has traditionally been modelled by a variogram, which describes the degree to which nearby locations have similar values.
- A variogram cloud is a scatterplot of data pairs, in which the semivariance is plotted against interpoint distance.
- The semivariance $\gamma(d)$ is formally defined as the squared difference in height between locations:

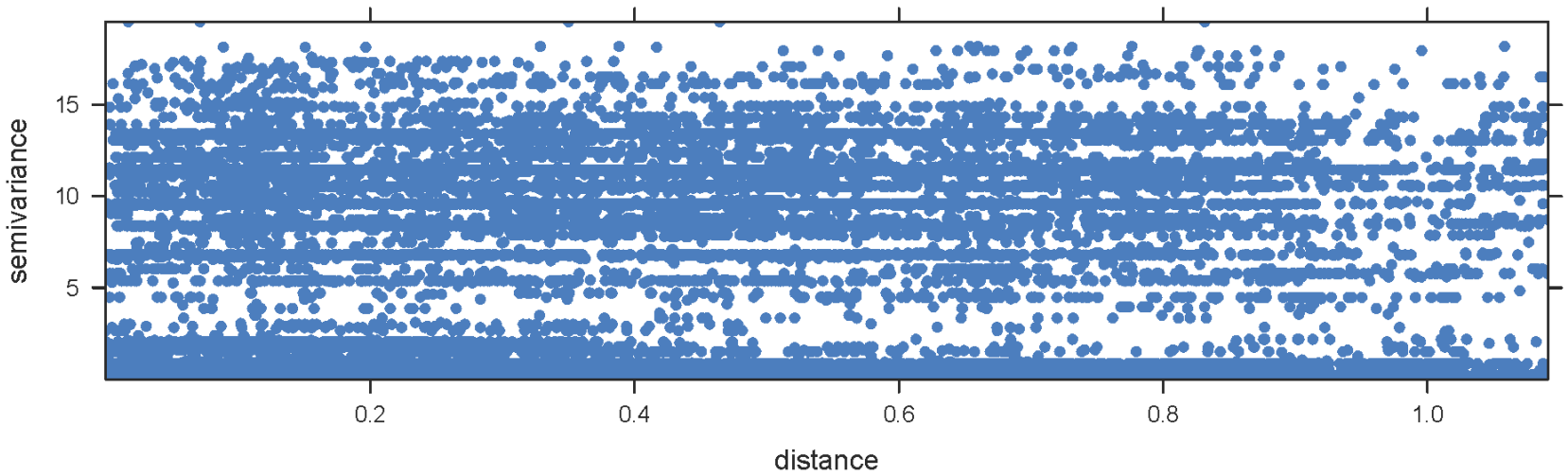
$$\hat{\gamma}(d) = \frac{1}{2n(d)} \sum_{d_{ij}=d} (Z(x_i) - Z(x_j))^2$$

- where $n(d)$ is the number of point pairs separated by distance d, and $Z(x_i)$ is the value of a variable at location x_i.

Geostatistics: Variogram

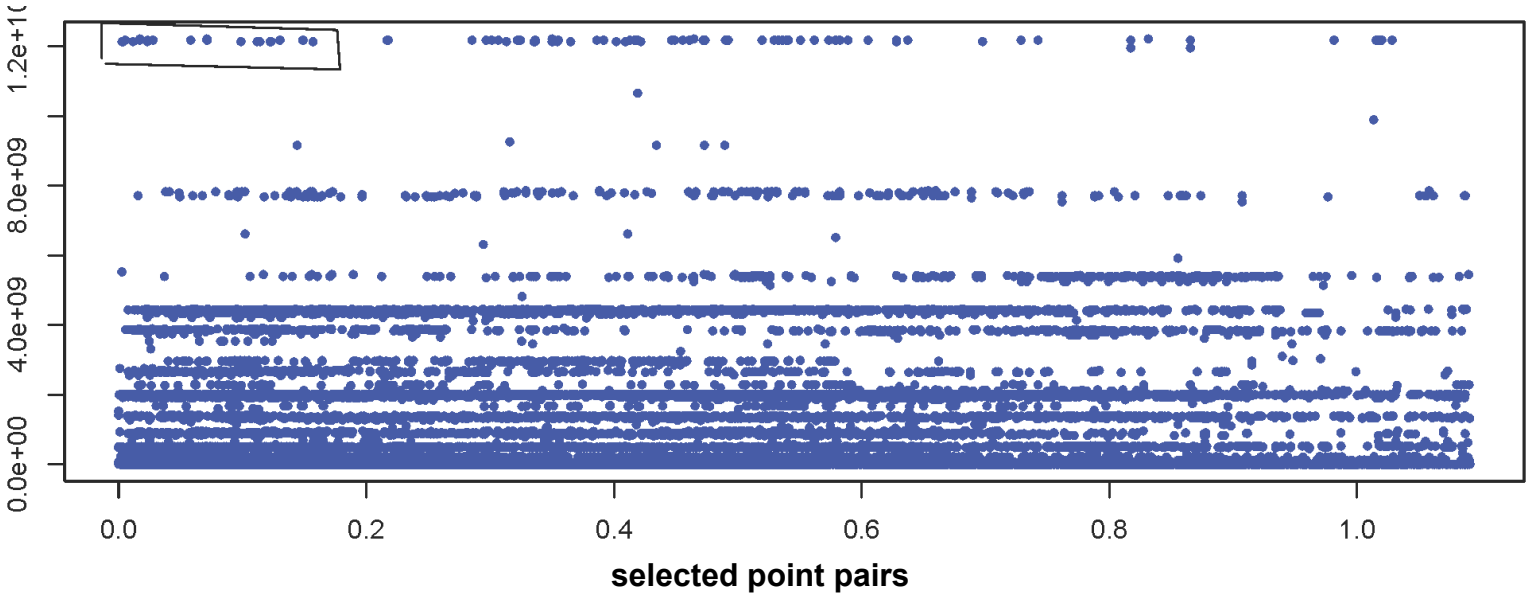
Below is the variogram cloud for Laos bomb load (natural log).

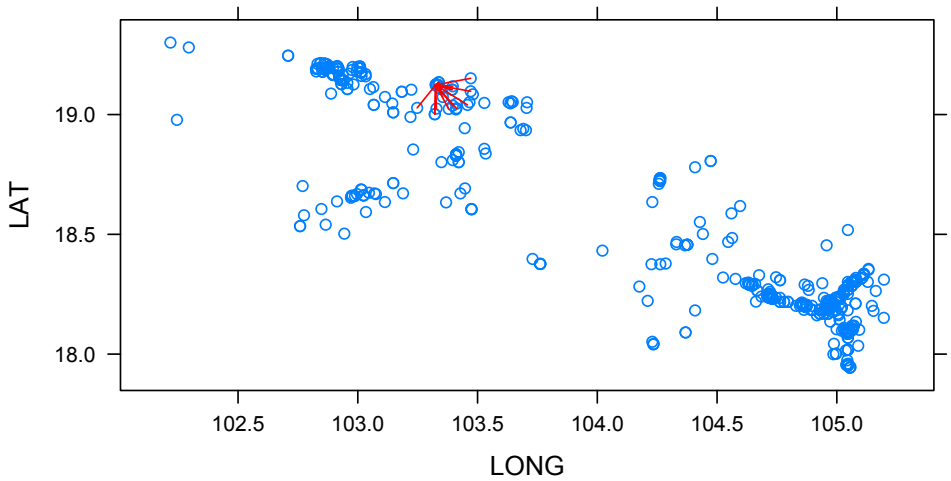
- **Upper left corner**: point pairs are close together, but have very different values.
- **Lower left corner**: close together, similar values.
- **Upper right corner**: far apart, different values.
- **Lower right corner**: far apart, similar values.



Geostatistics: Variogram

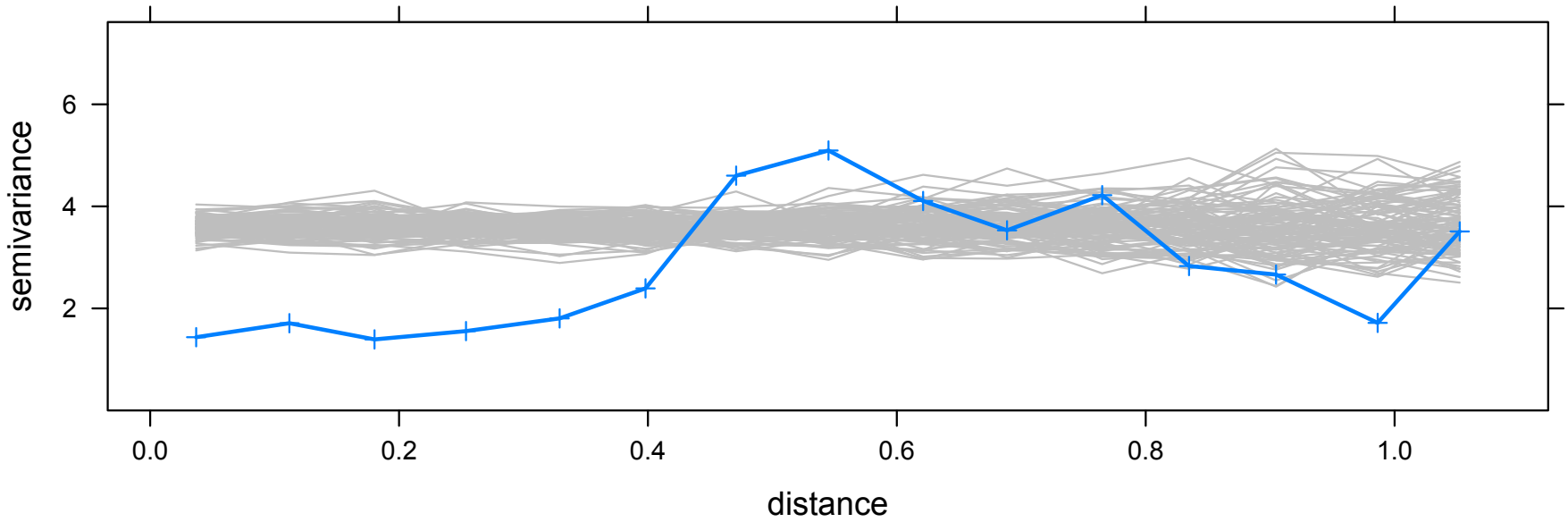
A variogram can be used to identify outliers...





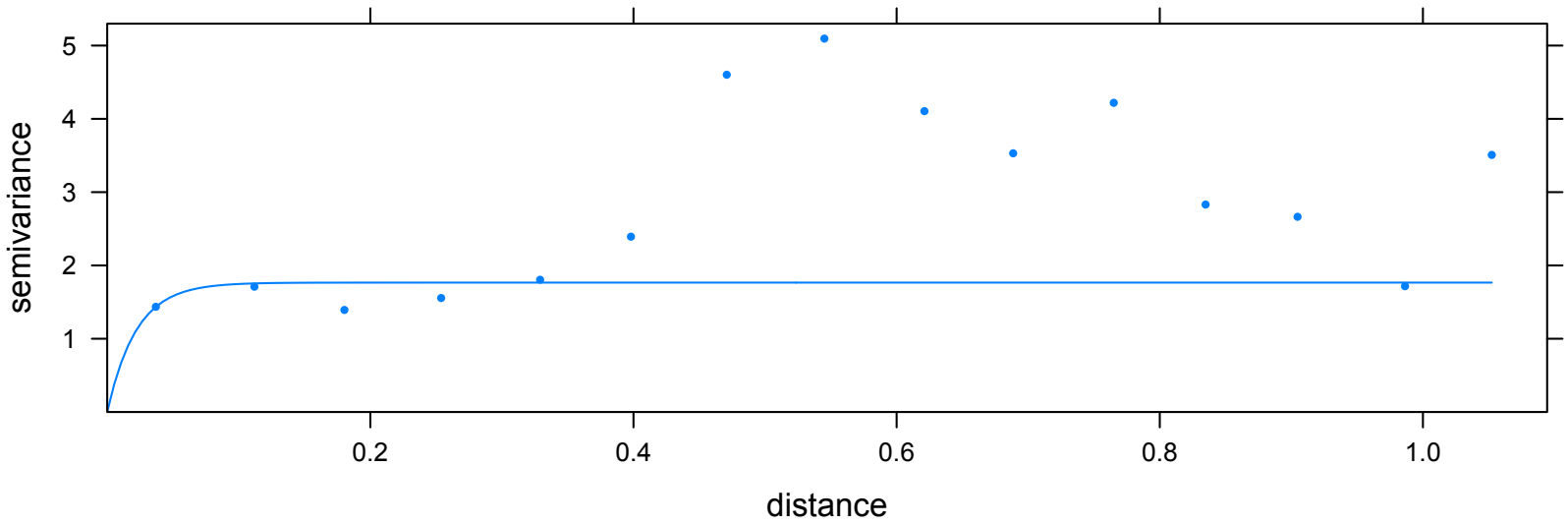
Geostatistics: Variogram

To test the null hypothesis that an increase in semivariance with distance is due to chance, we can use simulation to generate 100 spatially random datasets and check whether the sample variogram falls within the range of the random variograms. As shown below, the CSR hypothesis seems unlikely for the Laos bombing data.



Geostatistics: Variogram

- The variogram can be used for spatial prediction.
- This can be done by fitting a parametric model to the variogram.
- In the Laos example below, an exponential model was used.
- The shape of the curve indicates that at small separation distances, the variance in z is small. After a certain level of separation (.5 degrees), the variance in z values becomes somewhat random and the model flattens out to a value corresponding to the average variance.



Geostatistics: Ordinary Kriging

- Kriging is used to interpolate a value $Z(x_0)$ of a random field $Z(x)$ at unobserved location x_0, using data from observed location x_i.
- Allows variance to be non-constant, dependent on distance between points as modeled by the variogram $\gamma(d)$.
- The kriging estimator is given by

$$\hat{Z}(x_0) = \sum_{i=1}^{n} w_i(x_0) Z(x_i)$$

where $w_i(x_0),\ i = 1, \ldots, n$ is a spatial weight.

- Kriging is very similar to IDW interpolation, expect that the weights used in kriging are based on the model variogram, rather than an arbitrary function of distance.

Geostatistics: Ordinary Kriging

- To interpolate at a point x_0 based on points $x_1, \ldots, x_n$, the weights $w_1, \ldots w_n$ must be found. This can be done by solving the system of linear equations:

$$\begin{bmatrix} \gamma(d_{11}) & \gamma(d_{12}) & \cdots & \gamma(d_{1n}) & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma(d_{n1}) & \gamma(d_{n2}) & \cdots & \gamma(d_{nn}) & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{bmatrix} = \begin{bmatrix} w_1 \\ \vdots \\ w_n \\ \lambda \end{bmatrix} \begin{bmatrix} \gamma(d_{10}) \\ \vdots \\ \gamma(d_{n0}) \\ 1 \end{bmatrix}$$

  where $\gamma(d_{ij})$ is the semivariance for the distance between points x_i and x_j, and λ is the trend parameter.
- Ordinary kriging assumes an unknown constant trend: $\lambda(x) = \lambda$.

Geostatistics: Ordinary Kriging

Once weights are estimated, interpolation by ordinary kriging is given by:

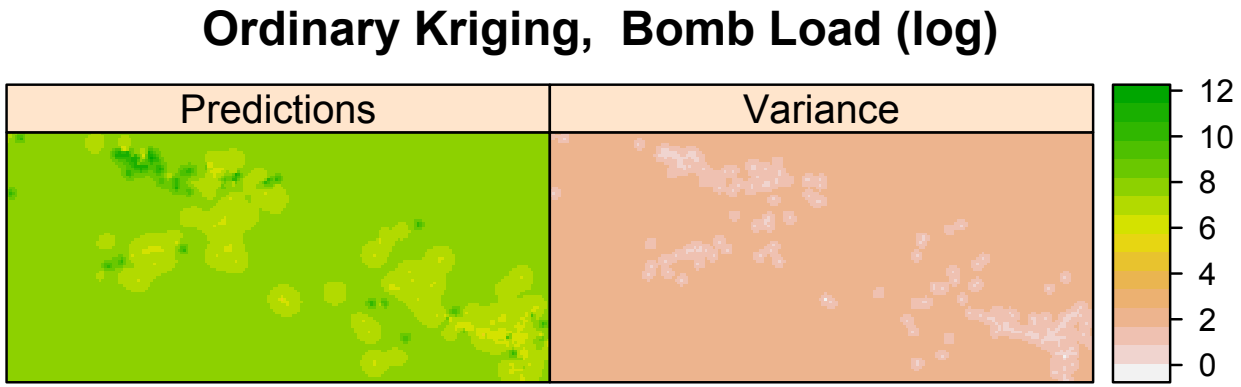
$$\hat{Z}(x_0) = \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix}' \begin{pmatrix} Z(x_1) \\ \vdots \\ Z(x_n) \end{pmatrix}$$

The ordinary kriging error is given by:

$$var\left(\hat{Z}(x_0) - Z(x_0)\right) = \begin{pmatrix} w_1 \\ \vdots \\ w_n \\ \lambda \end{pmatrix}' \begin{pmatrix} \gamma(d_{10}) \\ \vdots \\ \gamma(d_{n0}) \\ 1 \end{pmatrix}$$

Geostatistics: Ordinary Kriging

- Predicted values and variance for ordinary kriging is shown below for the Laos bombing data.



Examples in R

Switch to R tutorial script. Section 5.

Outline

Inefficiency of OLS estimators

- In a time-series context, the OLS estimator remains consistent even when a lagged dependent variable is present, as long as the error term does not show serial correlation.
- While the estimator may be biased in small samples, it can still be used for asymptotic inference.
- In a spatial context, this rule does not hold, irrespective of the properties of the error term.
- Consider the first-order SAR model (covariates omitted):

$$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \epsilon$$

- The OLS estimate for ρ would be:

$$\hat{\rho} = \Big((\mathbf{W}\mathbf{y})'(\mathbf{W}\mathbf{y})\Big)^{-1}(\mathbf{W}\mathbf{y})'\mathbf{y} = \rho + \Big((\mathbf{W}\mathbf{y})'(\mathbf{W}\mathbf{y})\Big)^{-1}(\mathbf{W}\mathbf{y})'\epsilon$$

- Similar to time series, the second term does not equal zero and the estimator will be biased.

Inefficiency of OLS estimators

- Asymptotically, the OLS estimator will be consistent if two conditions are met:

$$\text{plim } N^{-1}(\mathbf{Wy})'(\mathbf{Wy}) = \mathbf{Q} \quad \text{a finite and nonsingular matrix}$$
$$\text{plim } N^{-1}(\mathbf{Wy})'\epsilon = 0$$

- While the first condition can be satisfied with proper constraints on ρ and the structure of $\mathbf{W}$, the second does not hold in the spatial case:

$$\text{plim } N^{-1}(\mathbf{Wy})'\epsilon = \text{plim } N^{-1}\epsilon'(\mathbf{W})(\mathbf{I}_n - \rho\mathbf{W})^{-1}\epsilon \neq 0$$

- The presence of $\mathbf{W}$ in the expression results in a quadratic form in the error term.
- Unless $\rho = 0$, the plim will not converge to zero.

Properties of Maximum Likelihood Estimators

By contrast with OLS, maximum likelihood estimators (MLE) have attractive asymptotic properties, which apply in the presence of spatially lagged terms. ML estimates will exhibit consistency, efficiency and asymptotic normality if the following conditions are met:

- A log-likelihood for parameters of interest must exist (i.e.: non-degenerate $\ln L$)
- The log-likelihood must be continuously differentiable
- Boundedness of various partial derivatives
- The existence, positive definiteness and/or non-singularity of covariance matrices
- Finiteness of various quadratic forms

The various conditions are typically met when the structure of spatial interaction, expressed jointly by the autoregressive coefficient and the weights matrix, is nonexplosive (Anselin 1988).

Two-stage techniques

Instrumental variable estimation has similar asymptotic properties to MLE, but can be easier to implement numerically.

- Recall that the failure of OLS in models with spatially lagged DV's is due the correlation between the spatial variable and the error term (plim $N^{-1}(\mathbf{Wy})'\epsilon \neq 0$)
- This endogeneity issue can be addressed with two-stage methods based on the existence of a set of instruments $\mathbf{Q}$, which are strongly correlated with the original variables $\mathbf{Z} = [\mathbf{Wy} \quad \mathbf{X}]$, but asymptotically uncorrelated with the error term.

Two-stage techniques

- Where $\mathbf{Q}$ is of the same column dimension as $\mathbf{Z}$, the instrumental variable estimate θ_{IV} is

$$\theta_{IV} = [\mathbf{Q}'\mathbf{Z}]^{-1}\mathbf{Q}'\mathbf{y}$$

- In the general case where the dimension of $\mathbf{Q}$ is larger than $\mathbf{Z}$, the problem can be formulated as a minimization of the quadratic distance from zero:

$$\min\Phi(\theta) = (\mathbf{y} - \mathbf{Z}\theta)'\mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'(\mathbf{y} - \mathbf{Z}\theta)$$

- The solution to this optimization problem is the IV estimator θ_{IV}

$$\theta_{IV} = [\mathbf{Z}'\mathbf{P}_Q\mathbf{Z}]^{-1}\mathbf{Z}'\mathbf{P}_Q\mathbf{y}$$

with $\mathbf{P}_Q = \mathbf{Q}[\mathbf{Q}'\mathbf{Q}]^{-1}\mathbf{Q}'$ an idempotent projection matrix

Two-stage techniques

- $\mathbf{P}_Q\mathbf{Z}$ can be seen to correspond to a matrix of predicted values from regressions of each variable in $\mathbf{Z}$ on the instruments in $\mathbf{Q}$

$$\mathbf{P}_Q\mathbf{Z} = \mathbf{Q}\{[\mathbf{Q}'\mathbf{Q}]^{-1}\mathbf{Q}'\mathbf{Z}\}$$

- where the bracketed term is the OLS estimate for a regression of $\mathbf{Z}$ on $\mathbf{Q}$.
- Let $\mathbf{Z}_p$ be the predicted values of $\mathbf{Z}$. Then the IV estimator can also be expressed as

$$\theta_{IV} = [\mathbf{Z}_p'\mathbf{Z}]^{-1}\mathbf{Z}_p'\mathbf{y}$$

- which is the 2SLS estimator.

Two-stage techniques

Instrumental variable approaches are highly sensitive to the choice of instruments. Several options exist:

- Spatially lagged predicted values from a regression of $\mathbf{y}$ on non-spatial regressors ($\mathbf{W}\mathbf{y}^*$) (Anselin 1980).
- Spatial lags of exogenous variables ($\mathbf{WX}$) (Anselin 1980, Kelejian and Robinson 1993).
- In a spatiotemporal context, a time-wise lagged dependent variable or its spatial lag ($\mathbf{W}\mathbf{y}_{t-1}$) (Haining 1978).

Spatial autoregressive model (SAR): Likelihood function

- The **full log-likelihood** has the form:

$$\ln L = -\frac{n}{2}\ln(\pi\sigma^2) + \ln|\mathbf{I}_n - \rho\mathbf{W}| - \frac{e'e}{2\sigma^2}$$
$$e = (\mathbf{I}_n - \rho\mathbf{W})y - \mathbf{X}\beta$$

- It follows that the maximization of the likelihood is equivalent to a minimization of squared errors, corrected by the determinants from the Jacobian (Anselin 1988).
- This correction – and particularly the spatial term in $|\mathbf{I}_n - \rho\mathbf{W}|$ – will keep the least squares estimate from being equivalent to MLE.

Spatial autoregressive model (SAR): Likelihood function

- The most demanding part of the functions called to optimize the spatial autoregressive coefficient is the calculation of the Jacobian, the log-determinant of the $n \times n$ matrix $|\mathbf{I}_n - \rho\mathbf{W}|$
- One option is to express the determinant as a function of the eigenvalues ω of $\mathbf{W}$ (Ord 1975):

$$\ln|\mathbf{I}_n - \rho\mathbf{W}| = \ln\prod_{i=1}^{n}(1 - \rho\omega_i) = \sum_{i=1}^{n}\ln(1 - \rho\omega_i)$$

- An alternative approach is brute-force calculation of the determinant and inverse matrix at each iteration.

OLS vs. SAR

Consider the following linear regression of percent of county vote won by President Bush ($\mathbf{y}$) on per capita income in the county ($\mathbf{X}$): $\mathbf{y} = \mathbf{X}\beta + \epsilon$.

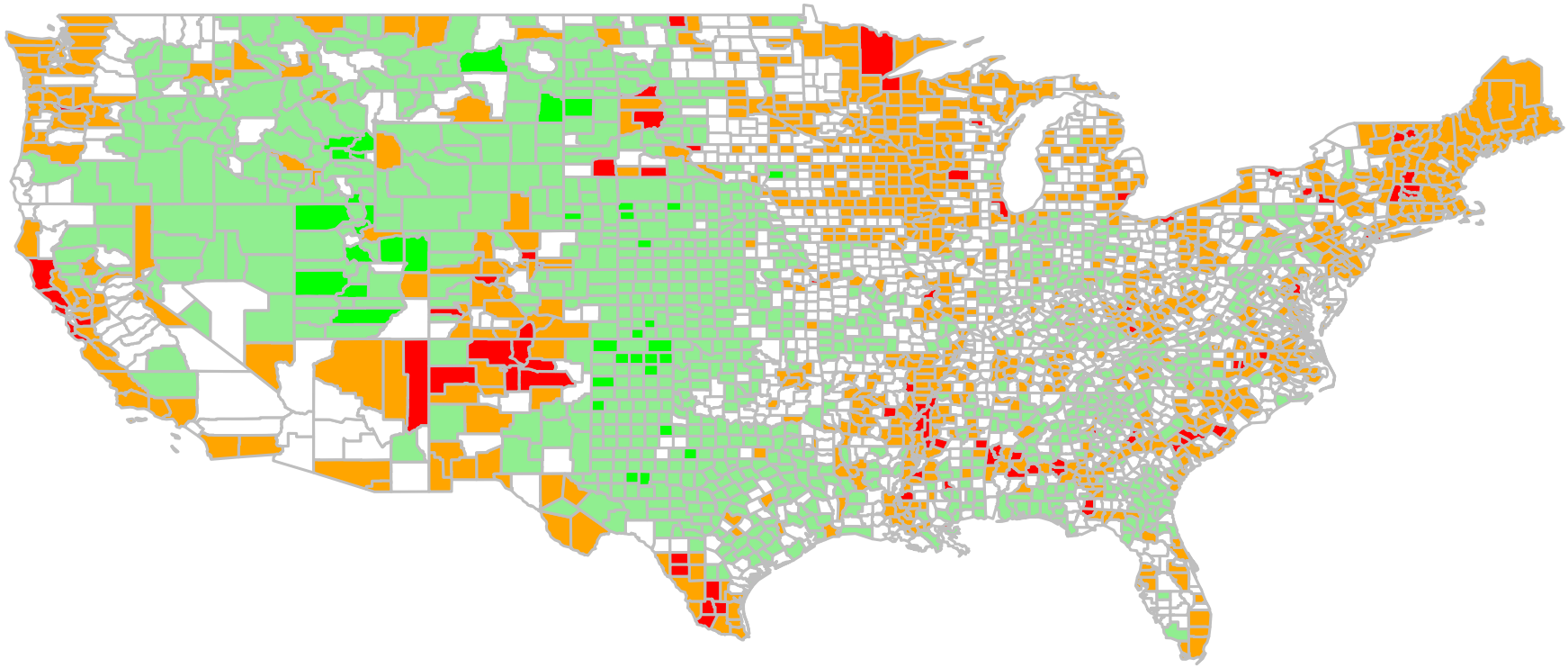
	OLS
(Intercept)	63.4340
	(0.8893)***
Per capita income	-0.0002
	(0.0000)***
AIC	24,666
N	3,111
Moran's $\mathcal{I}$ Residuals	0.550
Moran's $\mathcal{I}$ Std. Deviate	51.138***

$^*p \leq .05$, $^{**}p \leq .01$, $^{***}p \leq .001$

The Moran's $\mathcal{I}$ statistic shows a significant amount of spatial autocorrelation in the residuals.

OLS Residuals

Below is a map of residuals from a linear regression of percent of country vote received by Bush on per capita income.



OLS vs. SAR

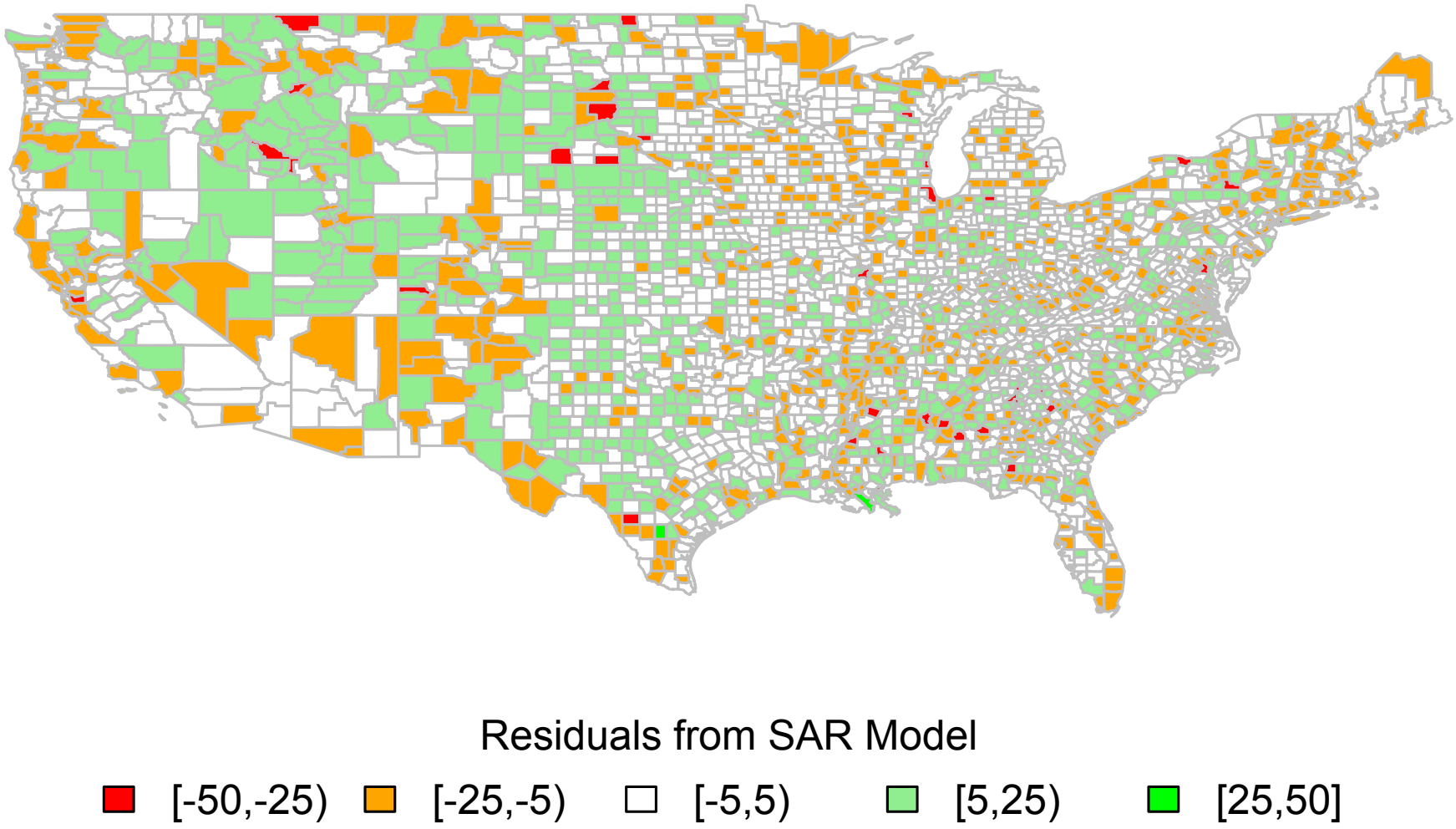
And the same model estimated by SAR: $\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \mathbf{X}\beta + \epsilon$.

	OLS	SAR
(Intercept)	63.4340	14.073
	(0.8893)***	(1.0572)***
Per capita income	-0.0002	5.46e-05
	(0.0000)***	(3.38e-05)
ρ		0.7510
		(0.0143)***
AIC	24,666	22,860
N	3,111	3,111
Moran's $\mathcal{I}$ Residuals	0.550	-0.0410
Moran's $\mathcal{I}$ Std. Deviate	51.138***	-3.7788
	*$p \leq .05$, **$p \leq .01$, ***$p \leq .001$	

The ρ coefficient is positive and highly significant, indicating strong spatial autocorrelation in the dependent variable. The Moran's $\mathcal{I}$ statistic indicates that the residuals are no longer spatially clustered.

SAR Residuals

Below is a map of residuals from the SAR model.



SAR Equilibrium Effects

- Because of the dependence structure of the SAR model, coefficient estimates do not have the same interpretation as in OLS.
- The β parameter reflects the short-run direct impact of x_i on y_i. However, we also need to account for the indirect impact of x_i on y_i, from the influence y_i exerts on its neighbors y_j, which in turn feeds back into y_i.
- The equilibrium effect of a change in x_i on y_i can be calculated as:

$$\mathbb{E}[\Delta y] = (\mathbf{I}_n - \rho \mathbf{W})^{-1} \Delta \mathbf{X}$$

where $\Delta \mathbf{X}$ is a matrix of changes to the covariates, and Δy is the associated change in the dependent variable.

- Since each unit will have a different set of connectivities to its neighbors, the impact of a hypothetical change in x_i will depend on which unit is being changed.

SAR Equilibrium Effects

- Below are the equilibrium effects (increase in percent of county vote for Bush) associated with a doubling of per capita income in Bronx County.

County	OLS	SAR
Currituck	0	4.81
Plymouth	0	4.47
Bronx	-2.86	4.31
Hunterdon	0	3.89
Lebanon	0	3.80
Mercer	0	3.70
Dare	0	3.52
Dauphin	0	3.42
Edgecombe	0	3.37
Barnstable	0	3.36

Spatially lagged error

- Use of the spatial error model may be motivated by **omitted variable bias**.
- Suppose that y is explained entirely by two explanatory variables x and z, where $x, z \sim N(0, \mathbf{I}_n)$ and are independent.

$$y = x\beta + z\theta$$

- If z is not observed, the vector $z\theta$ is nested into the error term ϵ.

$$y = x\beta + \epsilon$$

- Examples of latent variable z: culture, social capital, neighborhood prestige.

Spatially lagged error

- But we may expect the latent variable z to follow a spatial autoregressive process.

$$z = \lambda \mathbf{W} z + r$$
$$z = (\mathbf{I}_n - \lambda \mathbf{W})^{-1} r$$

- where $r \sim N(0, \sigma^2 \mathbf{I}_n)$ is a vector of disturbances, $\mathbf{W}$ is the spatial weights matrix, and λ is a scalar parameter.
- Substituting this back into the previous equation, we have the DGP for the spatial error model (SEM) :

$$y = X\beta + z\theta$$
$$y = X\beta + (\mathbf{I}_n - \lambda \mathbf{W})^{-1} u$$

- where $u = \theta r$

Spatially lagged error

- In addition to omitted variable bias, another motivation for the spatial error model might be **spatial heterogeneity**.
- Suppose we have a panel data set, with multiple observations for each unit.
- If we want our model to incorporate individual effects, we can include an $n \times 1$ vector $\mathbf{a}$ of individual intercepts for each unit:

$$\mathbf{y} = \mathbf{a} + \mathbf{X}\beta$$

- But in a cross-sectional setting, with one observation per unit, this approach is not feasible, since we'll have more parameters than observations.

Spatially lagged error

- Instead, we can treat **a** as a vector of spatial random effects.
- We assume that the vector of intercepts **a** follows a spatial autoregressive process:

$$\mathbf{a} = \lambda \mathbf{W}\mathbf{a} + \epsilon$$
$$\mathbf{a} = (\mathbf{I}_n - \lambda \mathbf{W})^{-1}\epsilon$$

- where $\epsilon \sim N(0, \sigma^2 \mathbf{I}_n)$ is a vector of disturbances
- Substituting this into the previous model yields the DGP of the SEM:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{a}$$
$$\mathbf{y} = \mathbf{X}\beta + (\mathbf{I}_n - \lambda \mathbf{W})^{-1}\epsilon$$

Spatially lagged error: Likelihood function

- The **full log-likelihood** has the form:

$$\ln L = -\frac{n}{2}\ln(\pi\sigma^2) + \ln|\mathbf{I}_n - \lambda\mathbf{W}| - \frac{e'e}{2\sigma^2}$$
$$e = (\mathbf{I}_n - \lambda\mathbf{W})(y - \mathbf{X}\beta)$$

Spatially lagged error: Interpretation of coefficients

- The SEM is essentially a generalized normal linear model with spatially autocorrelated disturbances.
- Assuming independence between **X** and the error term, least squares estimates for β are not efficient, but still unbiased.
- Because the SEM does not involve spatial lags of the dependent variable, estimated β parameters can be interpreted as partial derivatives:

$$\beta_k = \frac{\delta y_i}{\delta x_{jk}} \quad \forall \quad i, k$$

- where i indexes the observations and k indexes the explanatory variables.

SEM Estimates

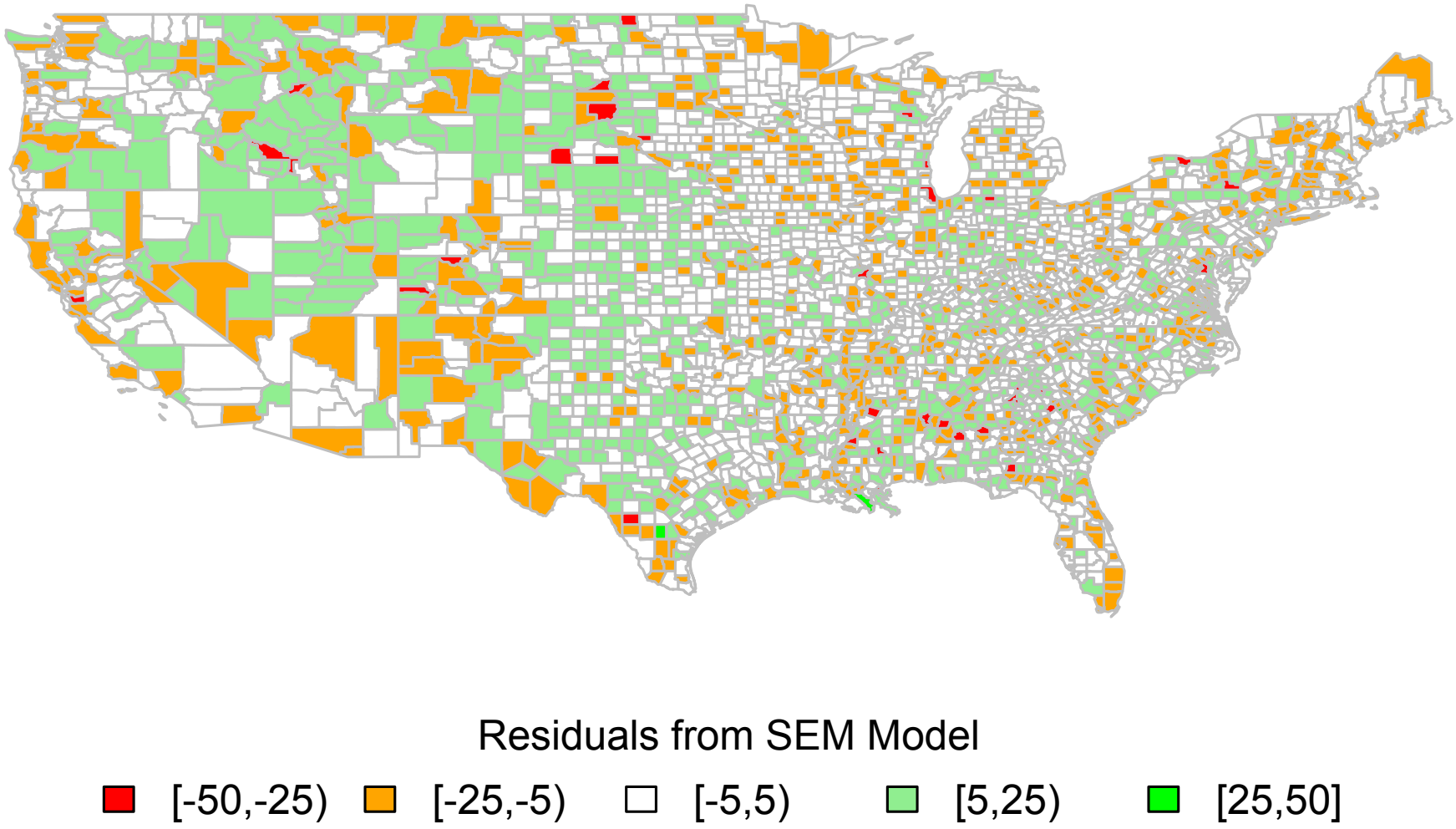
Let's run the election model from before: $\mathbf{y} = \mathbf{X}\beta + \lambda\mathbf{W}\mathbf{u} + \epsilon$.

	OLS	SAR	SEM
(Intercept)	63.434	14.073	58.3470
	(0.8893)***	(1.0572)***	(.9910)***
Per capita income	-0.0002	5.46e-05	8.02e-05
	(0.0000)***	(3.38e-05)	(4.17e-05)'
ρ		0.7510	
		(0.0143)***	
λ			0.7612
			(0.01422)***
AIC	24,666	22,860	22,864
N	3,111	3,111	3,111
Moran's $\mathcal{I}$ Residuals	0.550	-0.0410	-0.0511
Moran's $\mathcal{I}$ Std. Deviate	51.138***	-3.7788	-4.7192
	'$p \leq .1$, *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$		

The λ coefficient is positive and highly significant, indicating strong spatial dependence in the errors.

SEM Residuals

Below is a map of residuals from the SEM model.



Spatial Durbin Model

- Like the SEM, the Spatial Durbin Model can be motivated by concern over **omitted variables**.
- Recall the DGP for the SEM:

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

- Now suppose that $\mathbf{X}$ and $\mathbf{u}$ are correlated.
- One way to account for this correlation would be to conceive of $\mathbf{u}$ as a linear combination of $\mathbf{X}$ and an error term $\mathbf{v}$ that is independent of $\mathbf{X}$.

$$\mathbf{u} = \mathbf{X}\gamma + \mathbf{v}$$
$$\mathbf{v} \sim N(0, \sigma^2\mathbf{I}_n)$$

- where the scalar parameter γ and σ^2 govern the strength of the relationship between $\mathbf{X}$ and $\mathbf{z} = (\mathbf{I}_n - \lambda\mathbf{W})^{-1}$

Spatial Durbin Model

- Substituting this expression for $\mathbf{u}$, we have the following DGP:

$$\begin{aligned}
\mathbf{y} &= \mathbf{X}\beta + (\mathbf{I}_n - \lambda\mathbf{W})^{-1}(\gamma\mathbf{X} + \mathbf{v}) \\
\mathbf{y} &= \mathbf{X}\beta + (\mathbf{I}_n - \lambda\mathbf{W})^{-1}\gamma\mathbf{X} + (\mathbf{I}_n - \lambda\mathbf{W})^{-1}\mathbf{v} \\
(\mathbf{I}_n - \lambda\mathbf{W})\mathbf{y} &= (\mathbf{I}_n - \lambda\mathbf{W})\mathbf{X}\beta + \gamma\mathbf{X} + \mathbf{v}
\end{aligned}$$

$$\mathbf{y} = \lambda\mathbf{W}\mathbf{y} + \mathbf{X}(\beta + \gamma) + \mathbf{W}\mathbf{X}(-\lambda\beta) + \mathbf{v}$$

- This is the Spatial Durbin Model (SDM), which includes a spatial lag of the dependent variable $\mathbf{y}$, as well as the explanatory variables $\mathbf{X}$.

Spatial Durbin Model

- The Spatial Durbin Model can also be motivated by concern over **spatial heterogeneity**.
- Recall the vector of intercepts **a**:

$$\mathbf{a} = (\mathbf{I}_n - \lambda \mathbf{W})^{-1} \epsilon$$

- Now suppose that **X** and ϵ are correlated.
- As before, let's restate ϵ as a linear combination of **X** and random noise **v**.

$$\mathbf{a} = \mathbf{X}\gamma + \mathbf{v}$$

- Substituting this back into the SEM yields the same expression of SDM as before:

$$\mathbf{y} = \lambda \mathbf{W}\mathbf{y} + \mathbf{X}(\beta + \gamma) + \mathbf{W}\mathbf{X}(-\lambda\beta) + \mathbf{v}$$

Spatial Durbin Model: Likelihood function

- Let's restate the SDM as follows:

$$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \alpha\iota_n + \mathbf{X}\beta + \mathbf{W}\mathbf{X}\theta + \epsilon$$

- The **log-likelihood** has a similar form to the SEM:

$$\ln L = -\frac{n}{2}\ln(\pi\sigma^2) + \ln|\mathbf{I}_n - \rho\mathbf{W}| - \frac{e'e}{2\sigma^2}$$
$$e = \mathbf{y} - \rho\mathbf{W}\mathbf{y} - \mathbf{Z}\delta$$

- where $\mathbf{Z} = [\iota_n \quad \mathbf{X} \quad \mathbf{W}\mathbf{X}]$, $\delta = [\alpha \quad \beta \quad \theta]$, and ρ is bounded by $(min(\omega)^{-1}, max(\omega)^{-1})$, where ω is an $n \times 1$ vector of eigenvalues of $\mathbf{W}$.

SDM Estimates

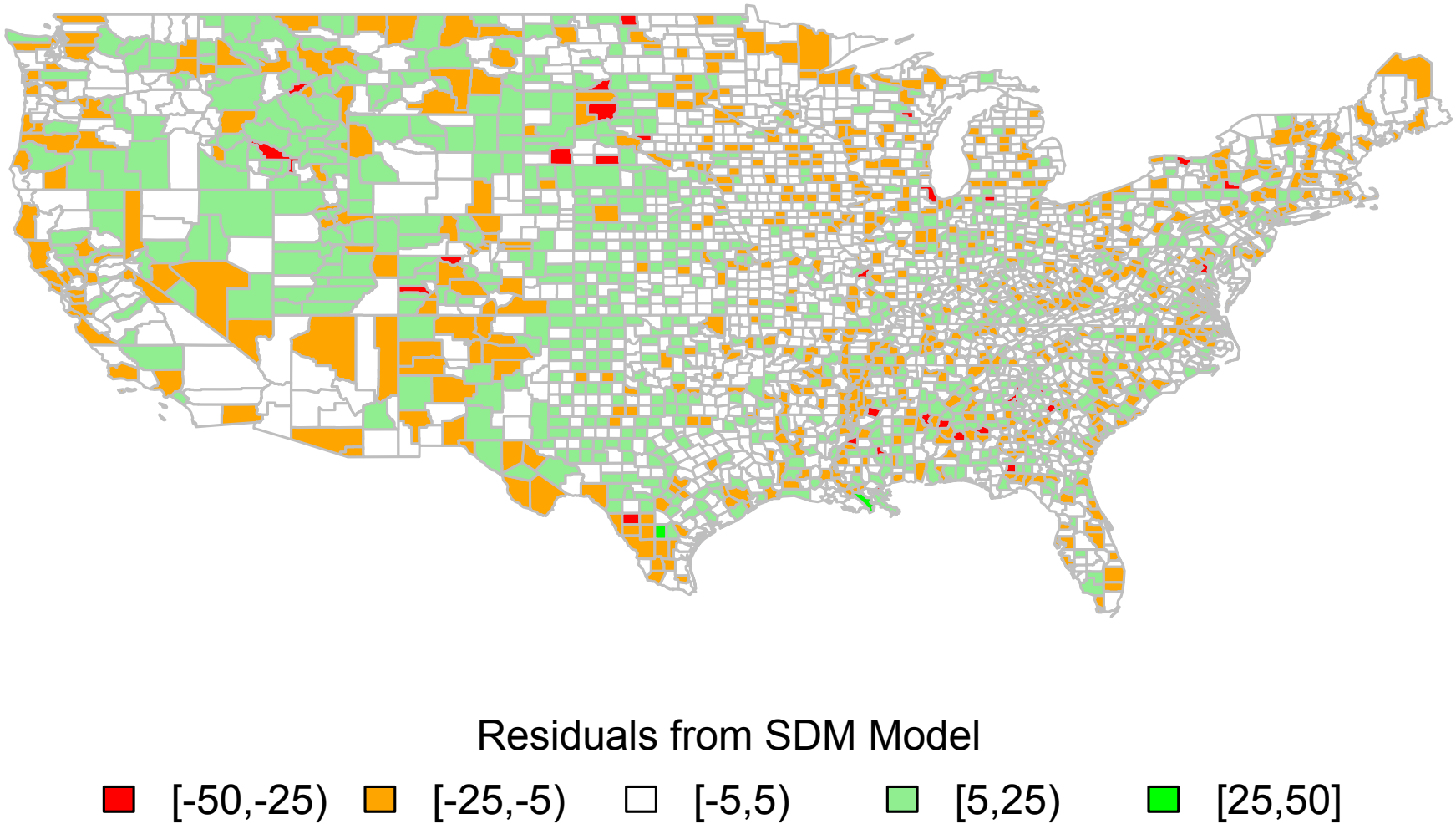
Let's try running the SDM: $\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \alpha\iota_n + \mathbf{X}\beta + \mathbf{W}\mathbf{X}\theta + \epsilon$

	OLS	SAR	SEM	SDM
(Intercept)	63.434	14.073	58.3470	16.848
	(0.8893)***	(1.0572)***	(.9910)***	(1.2588)***
Per capita income	-0.0002	5.46e-05	8.02e-05	0.0002
	(0.0000)***	(3.38e-05)	(4.17e-05)'	(0.0000)***
Lagged Bush vote (ρ)		0.7510		0.7501
		(0.0143)***		(0.0144)***
Lagged error (λ)			0.7612	
			(0.01422)***	
Lagged income (θ)				-0.0003
				(0.0001)***
AIC	24,666	22,860	22,864	22,843
N	3,111	3,111	3,111	3,111
Moran's $\mathcal{I}$ Residuals	0.550	-0.0410	-0.0511	-0.0454
Moran's $\mathcal{I}$ Std. Deviate	51.138***	-3.7788	-4.7192	-4.1894
	'$p \leq .1$, *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$			

The SDM results in a slightly better fit...

SDM Residuals

Below is a map of residuals from the SDM model.



Extensions: Spatial Autocorrelation Model (SAC)

- The SAC model contains spatial dependence in both the dependent variable and the errors, with (potentially) two different weights matrices.

$$\mathbf{y} = \rho\mathbf{W_1}\mathbf{y} + \mathbf{X}\beta + \lambda\mathbf{W_2}\mathbf{u} + \epsilon$$
$$\mathbf{y} = (\mathbf{I}_n - \rho\mathbf{W_1})^{-1}\mathbf{X}\beta + (\mathbf{I}_n - \rho\mathbf{W_1})^{-1}(\mathbf{I}_n - \lambda\mathbf{W_2})^{-1}\epsilon$$
$$\epsilon \sim N(0, \sigma^2\mathbf{I}_n)$$

- The **log-likelihood** has the form:

$$\ln L = -\frac{n}{2}\ln(\pi\sigma^2) + \ln|\mathbf{I}_n - \rho\mathbf{W_1}| + \ln|\mathbf{I}_n - \lambda\mathbf{W_2}| - \frac{e'e}{2\sigma^2}$$
$$e = (\mathbf{I}_n - \lambda\mathbf{W_2})\big((\mathbf{I}_n - \rho\mathbf{W_1})\mathbf{y} - \mathbf{X}\beta\big)$$

Extensions: Spatial Autoregressive Moving Average Model (SARMA)

- Like the SAC, the SARMA model also contains spatial dependence in the dependent variable and the errors.

$$\mathbf{y} = \iota_n \alpha + \rho \mathbf{W_1 y} + \mathbf{X}\beta + (\mathbf{I}_n - \theta \mathbf{W_2})\epsilon$$
$$\mathbf{y} = (\mathbf{I}_n - \rho \mathbf{W_1})^{-1}(\mathbf{X}\beta + \iota_n \alpha) + (\mathbf{I}_n - \rho \mathbf{W_1})^{-1}(\mathbf{I}_n - \theta \mathbf{W_2})\epsilon$$
$$\epsilon \sim N(0, \sigma^2 \mathbf{I}_n)$$

- The main distinction between the SAC and SARMA is the series representation of the inverse $(\mathbf{I}_n - \theta \mathbf{W_2})$.
- As a result, the SAC places more emphasis on higher order neighbors.

Extensions: Spatial Durbin Error Model (SDEM)

- The SDEM model contains spatial dependence in both the explanatory variables and the errors.

$$\mathbf{y} = \iota_n \alpha + \mathbf{X}\beta + \mathbf{WX}\gamma + (\mathbf{I}_n - \rho\mathbf{W})^{-1}\epsilon$$
$$\epsilon \sim N(0, \sigma^2 \mathbf{I}_n)$$

- Direct impacts correspond to the β parameters; indirect impacts correspond to the γ parameters
- The model can be generalized to incorporate two weights matrices without affecting interpretation of parameters:

$$\mathbf{y} = \iota_n \alpha + \mathbf{X}\beta + \mathbf{W_1X}\gamma + (\mathbf{I}_n - \rho\mathbf{W_2})^{-1}\epsilon$$

Examples in R

Switch to R tutorial script. Section 6.a.

Geographically Weighted Regression (GWR)

- A key assumption that we have made in the models examined thus far is that the structure of the model remains constant over the study area (no local variations in the parameter estimates).
- If we are interested in accounting for potential **spatial heterogeneity** in parameter estimates, we can use a Geographically Weighted Regression (GWR) model (Fotheringham et al., 2002).
- GWR permits the parameter estimates to vary locally, similar to a parameter drift for a time series model.
- GWR has been used primarily for exploratory data analysis, rather than hypothesis testing.

Geographically Weighted Regression (GWR)

- GWR rewrites the linear model in a slightly different form:

$$\mathbf{y}_i = \mathbf{X}\beta_i + \epsilon$$

where i is the location at which the local parameters are to be estimated.

- Parameter estimates are solved using a weighting scheme:

$$\beta_i = (\mathbf{X}'\mathbf{W}_i\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}_i\mathbf{y}$$

- where the weight w_{ij} for the j observation is calculated with a Gaussian function.

$$w_{ij} = e^{\left(\frac{-d_{ij}}{h}\right)^2}$$

where $d_{i,j}$ is the Euclidean distance between the location of observation i and location j, and h is the bandwidth.

- Bandwidth may be user-defined or selected by minimization of root mean square prediction error.

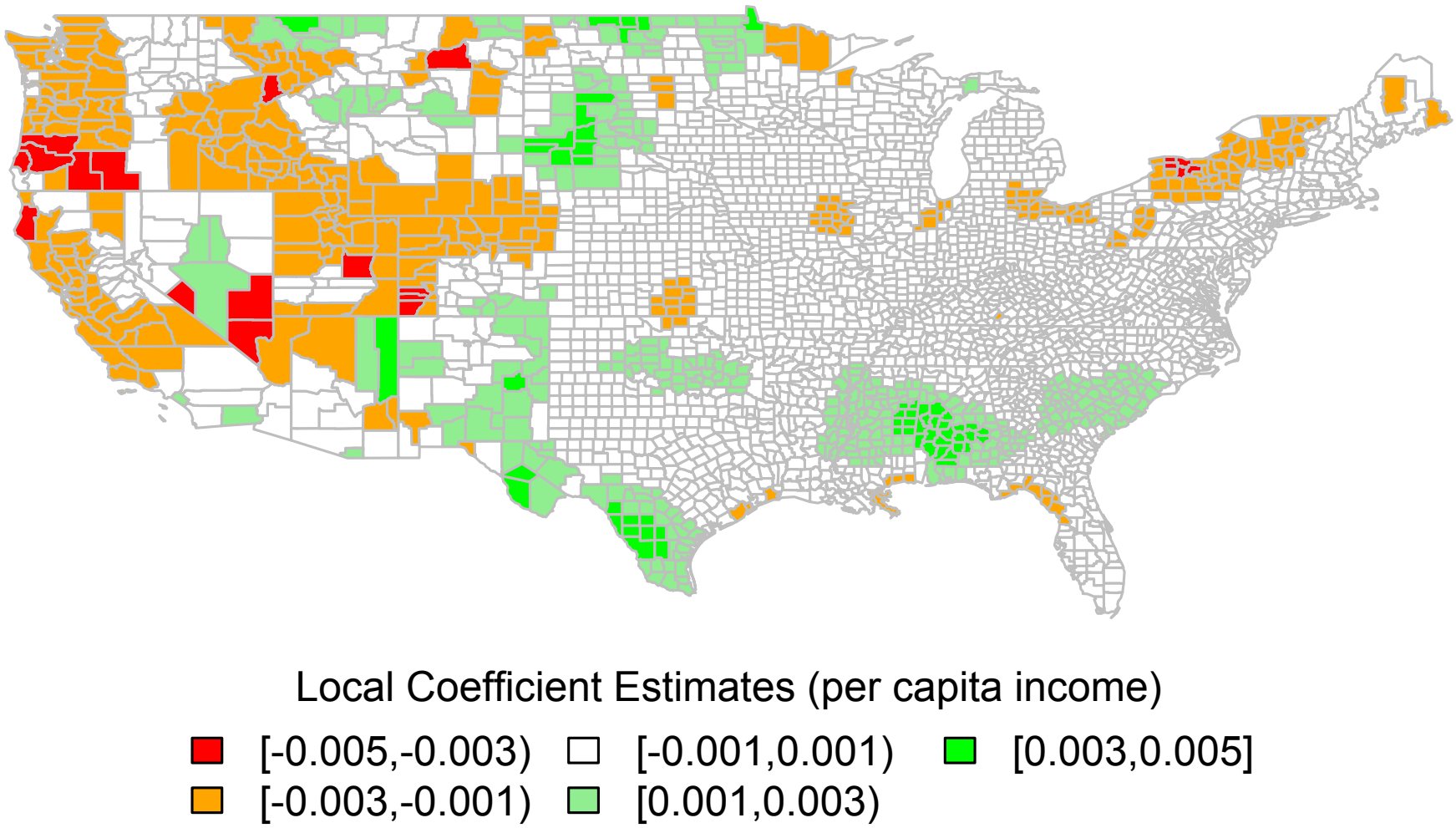
GWR Estimates

Let's try running the same election model as before with GWR:

	Geographically Weighted Regression				
	Global	Min	Mean	Max	S.E.
(Intercept)	63.4340	-26.02	59.95	185.36	(20.5262)
Per capita income	-0.0002	-0.0061	0.0001	0.0061	(0.0010)
Bandwidth	0.6649				
N	3,111				
Moran's $\mathcal{I}$ Residuals	0.0796				
Moran's $\mathcal{I}$ Std. Deviate	7.4239***				
	'$p \leq .1$, *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$				

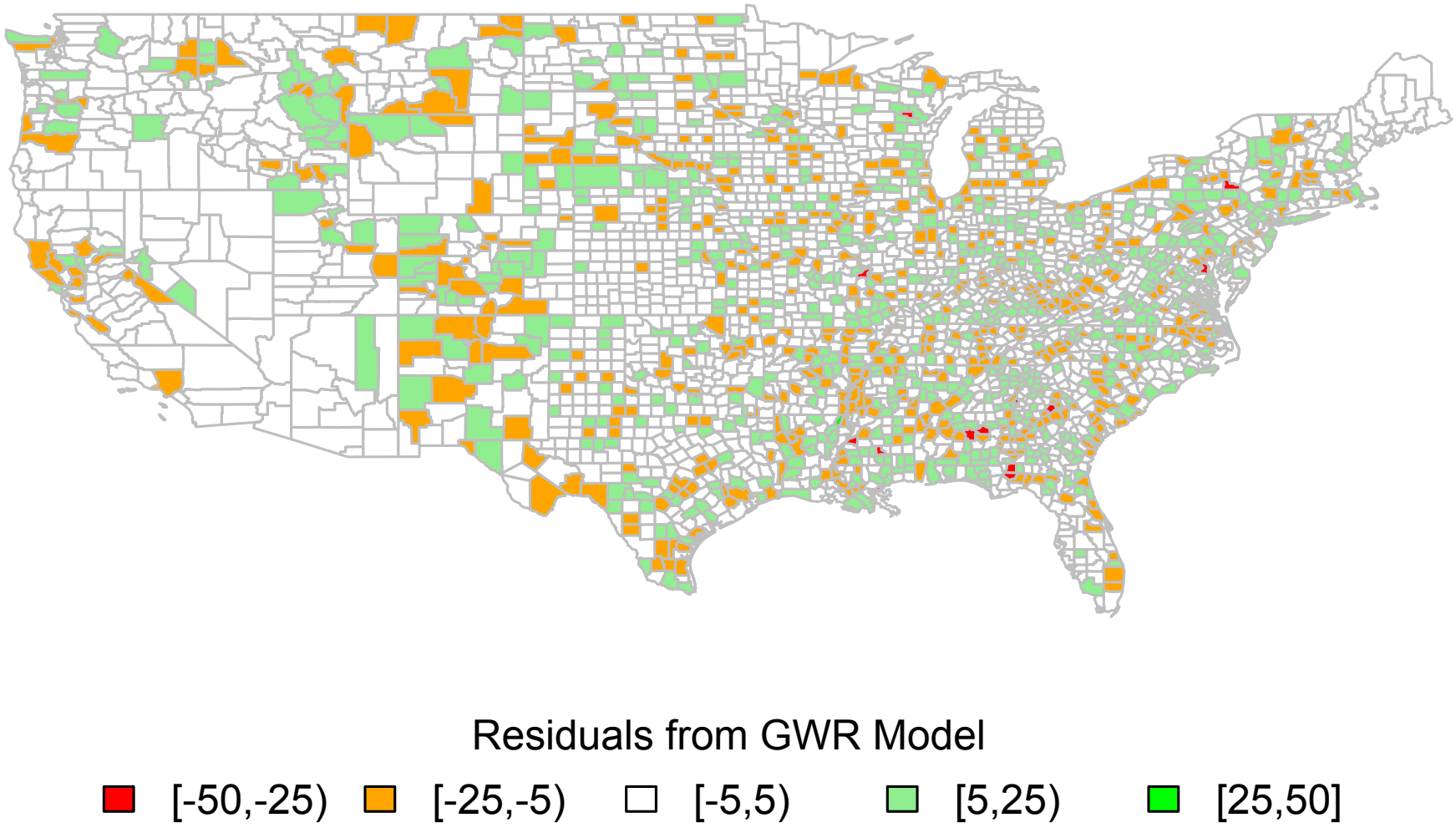
GWR Local Coefficient Estimates

Below is a map of local coefficients. The relationship between income and support for Bush is negative in red areas, and positive in green areas.



GWR Residuals

Below is a map of residuals from the GWR model.



Examples in R

Switch to R tutorial script. Section 6.b.

Spatial autologistic model

- Up to this point, we have only examined models which assume that the dependent variable is continuous and normally distributed.
- But what if we are interested in studying discrete events, measured categorically? (win/lose, war/peace, sick/healthy, Democrat/Republican, etc.)
- We may want to consider spatial dependence between observations with a conditional probability model, where the occurrence of an event $y = 1$ in neighboring units conditions the likelihood that unit i will itself experience the event.
- One option for such a task is the spatial autologistic model (Ward and Gleditsch 2002).

Spatial autologistic model

- The autologistic model states the conditional probability p_i that $y_i = 1$, given values y_j at units $(j \neq i)$:

$$p_i = P(y_i = 1 | \mathbf{W}y_i) = \frac{e^{\alpha + \mathbf{X}_i'\beta + \gamma \mathbf{W}y_i}}{1 + e^{\alpha + \mathbf{X}_i\beta + \gamma \mathbf{W}y_i}}$$

- where β is a vector of parameters for exogenous variables, γ is a scalar parameter for the spatial lag of y and $\mathbf{W}$ is a connectivity matrix.
- When $\gamma = 0$, this expression reduces to a standard logistic model and observations are considered independent of each other.
- When $\beta = 0$, this expression reduces to a pure autologistic model where unit-level covariates exert no independent influence on y once spatial dependence is taken into account.

Spatial autologistic model

- A maximum pseudo-likelihood estimator (MPE) for the unknown parameter vector $\theta = (\alpha \quad \beta \quad \gamma)$ is defined as the vector $\hat{\theta}$ which maximizes

$$\prod_{i=1}^{n} P(y_i = 1|\mathbf{W}y_i) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{y_i}$$

- An analytical form of the full likelihood is intractable because observations y_i are conditionally dependent on one another (Besag, 1974).
- Two solutions have been proposed:
  1. Maximum pseuso-likelihood estimation (MPLE).
  2. MCMC techniques.

Spatial autologistic model: MPLE approach

- Maximum pseudo-likelihood estimation maximizes the function obtained by multiplying together the logit likelihoods represented by equation on the previous slide (Besag 1977).
- This is equivalent to a maximum likelihood fit for a logit regression model with independent observations y_i.
- This procedure has been shown to provide consistent estimates of model parameters (Cressie, 1993).
- However, the standard errors of the estimated parameters are not directly applicable because they assume independence of the observations.
- The inefficiency of MPLE increases when the strength of spatial interaction is high (Huffer and Wu 1998).

Spatial autologistic model: MCMC approach

- Markov Chain Monte Carlo estimation can yield approximations closer to the full likelihood function (Geyer and Thompson 1992, Ward and Gleditsch 2002).
- One approach uses a probabilistic random map generated from the autologistic model, defined by parameters θ and sufficient statistics $s(y)$.

$$s(y) = \left(\sum_{i=1}^{n} y_i, \quad \sum_{i=1}^{n} X_i y_i, \quad \frac{1}{2} \sum_{i=1}^{n} \mathbf{W} y_i \right)$$

- A statistic $s(y)$ is sufficient for y if it contains all the information about y that is available in the sample.

Spatial autologistic model: MCMC

- A Gibbs sampler is used to generate a set of m sampled simulated maps with sufficient statistics $(y_{l \in \{1,\ldots,m\}})$
- The samples are conditioned on the vector of parameters ψ, the initial values for which are typically pseudolikelihood estimates for $\hat{\theta}$.
- The idea is to find the values of $\hat{\theta}$ that yield the sufficient statistics $s(y)$ for the observed data.
- MCMC maximum likelihood is obtained by solving the score equation

$$s(y) = \frac{\sum_{l=1}^{m} s(y_m) e^{(\hat{\theta}-\psi)' s(y_m)}}{\sum_{j=1}^{m} e^{(\hat{\theta}-\psi)' s(y_m)}}$$

Spatial autologistic model: MPLE vs. MCMC

- Let's try running the autologistic with some real data.
- Ward and Gleditsch (2002) estimate a simple model of war, where the probability of war in country i is conditioned on the number of neighboring countries experiencing war and the level democracy in those countries:

$$P(War_i = 1 | War_i' \mathbf{W}) = \frac{e^{\alpha + \beta_1 Dem_i + \beta_2 Dem_i' \mathbf{W^s} + \gamma War_i' \mathbf{W}}}{1 + e^{\alpha + \beta_1 Dem_i + \beta_2 Dem_i' \mathbf{W^s} + \gamma War_i' \mathbf{W}}}$$

- where Dem_i is a country i's Polity score (scaled $-10 : 10$ from least to most democratic), $\mathbf{W^s}$ is a row-standardized weights matrix and $\mathbf{W}$ is a binary contiguity matrix.

Spatial autologistic model: MPLE vs. MCMC

MPLE ($\hat{\psi}$) and MCMC ($\hat{\theta}$) estimates for the model are shown below:

	MPLE		MCMC	
	Coef	S.E.	Coef	S.E.
(Intercept)	-1.87	(0.33)	-1.53	(0.09)
Democracy	-0.02	(0.03)	-0.06	(0.01)
Spatial Lag of Democracy	0.01	(0.05)	-0.02	(0.01)
Spatial Lag of War	0.31	(0.13)	0.21	(0.02)

So, parameter estimates are generally similar, but the standard errors from MCMC are much smaller.

Examples in R

Switch to R tutorial script. Section 6.c.

Partial Adjustment Model

- Recall that with cross-sectional data, we often assume that observations represent an equilibrium outcome of a spatiotemporal process working over time.
- Here we will examine how spatiotemporal models relate to models used for CS data (SAR, SEM).
- For simplicity, we assume that:
  - Units are influenced by their own and their neighbors' history (no simultaneous dependence).
  - $\mathbf{W}$ is symmetric.
  - No structural change over time.
  - The matrix $\mathbf{X}$ – which may include spatial lags of explanatory variables – is constant or deterministically growing with respect to time.

Partial Adjustment Model

A simple modeling framework for space-time data is the Partial Adjustment Model (PAM).

- Like a conventional temporal model, PAM allows the dependent variable for each unit y_t to depend on that unit's own past values $y_{t-1}, \ldots, y_0$ and $X_{t-1}, \ldots, X_0$.
- This framework is extended to allow for spatial dependence on other regions by incorporating spatial lags of temporal lags $\mathbf{W}y_{t-1}, \ldots, \mathbf{W}y_0$ and $\mathbf{W}X_{t-1}, \ldots, \mathbf{W}X_0$.
- For development of this model, see Greene (1997) and LeSage and Pace (2009).

Partial Adjustment Model

The spatial PAM is formally defined below:

$$\mathbf{y}_t = (\tau \mathbf{I}_n - \rho \mathbf{W})\mathbf{y}_{t-1} + \mathbf{X}_\mathbf{t}^* \beta + \mathbf{u}_t$$
$$\mathbf{X}_t^* = \psi^t \mathbf{X}_0^* = \psi^t [\mathbf{X}_0 \quad \mathbf{W}\mathbf{X}_0 \quad \iota_n]$$
$$\mathbf{u}_t = \mathbf{X}_t^* \gamma + r + \epsilon_t$$

- Where τ governs dependence between each region at time t and $t-1$, ρ governs spatial dependence between each region at time t and neighboring regions at $t-1$, ψ is the growth rate parameter for $\mathbf{X}_0^*$ ($\psi = 1$ implies no growth, $\psi > 1$ implies growth; assume $\psi > \tau$).
- As in the SDM, we allow for potential dependence between omitted variables and exogenous variables, such that the error term $\mathbf{u}_t$ is partitioned into an endogenous component $\mathbf{X}_t^* \gamma$, an independent and time-invariant component $r \sim N(0, \sigma_r^2 \mathbf{I}_n)$, and independent noise $\epsilon_t \sim N(0, \sigma_\epsilon^2 \mathbf{I}_n)$ which is allowed to vary with time.

Partial Adjustment Model

This dynamic process implies a cross-sectional steady state characterized by simultaneous spatial interaction. To demonstrate this, we can use the recursive relation implied in the PAM:

$$\mathbf{y}_{t-1} = (\tau \mathbf{I}_n - \rho \mathbf{W})\mathbf{y}_{t-2} + \mathbf{X}^*_{\mathbf{t-1}}\beta + \mathbf{u}_{t-1}$$

The state of this dynamic system after the passage of t time periods is:

$$\begin{aligned}
\mathbf{y}_t =& (\tau \mathbf{I}_n - \rho \mathbf{W})^t \mathbf{y}_0 \\
&+ \left(\mathbf{I}_n \psi^t + (\tau \mathbf{I}_n - \rho \mathbf{W})\psi^{t-1} + , \ldots, + (\tau \mathbf{I}_n - \rho \mathbf{W})^{t-1}\psi\right)\mathbf{X}^*_{\mathbf{0}}\beta + \tilde{\mathbf{u}}_t \\
\tilde{\mathbf{u}}_t =& \widetilde{\mathbf{X}^*_t \gamma} + \tilde{r} + \tilde{\epsilon}_t \\
\widetilde{\mathbf{X}^*_t \gamma} =& \left(\mathbf{I}_n \psi^t + (\tau \mathbf{I}_n - \rho \mathbf{W})\psi^{t-1} + , \ldots, + (\tau \mathbf{I}_n - \rho \mathbf{W})^{t-1}\psi\right)\mathbf{X}^*_{\mathbf{0}}\gamma \\
\tilde{r} =& \left(\mathbf{I}_n + (\tau \mathbf{I}_n - \rho \mathbf{W}) + , \ldots, + (\tau \mathbf{I}_n - \rho \mathbf{W})^{t-1}\right) r \\
\tilde{\epsilon}_t =& \epsilon_t + (\tau \mathbf{I}_n - \rho \mathbf{W})\epsilon_{t-1} + (\tau \mathbf{I}_n - \rho \mathbf{W})^2 \epsilon_{t-2} + , \ldots, + (\tau \mathbf{I}_n - \rho \mathbf{W})^{t-1}\epsilon_1
\end{aligned}$$

Partial Adjustment Model

Taking the expectation of $\mathbf{y}_t$ yields:

$$\begin{aligned}
\mathbb{E}[\mathbf{y}_t] &\approx \left(\mathbf{I}_n\psi^t + (\tau\mathbf{I}_n - \rho\mathbf{W})\psi^{t-1} + , \ldots, + (\tau\mathbf{I}_n - \rho\mathbf{W})^{t-1}\psi\right)\mathbf{X}_\mathbf{0}^*(\beta+\gamma) \\
&\approx \left(\mathbf{I}_n + (\tau\mathbf{I}_n - \rho\mathbf{W})\psi^{-1} + , \ldots, + (\tau\mathbf{I}_n - \rho\mathbf{W})^{t-1}\psi^{-(t-1)}\right)\psi^t\mathbf{X}_\mathbf{0}^*(\beta+\gamma) \\
&\approx \left(\mathbf{I}_n - \frac{\rho}{\psi-\tau}\mathbf{W}\right)^{-1}\left(\frac{\psi}{\psi-\tau}\right)\mathbf{X}_\mathbf{t}^*(\beta+\gamma) \\
&\approx (\mathbf{I}_n - \rho^*\mathbf{W})^{-1}\mathbf{X}_\mathbf{t}^*\beta^*
\end{aligned}$$

where $\rho^* = \frac{\rho}{\psi-\tau}$ and $\beta^* = \frac{\psi(\beta+\gamma)}{\psi-\tau}$. This implies the familiar expression

$$\mathbf{y}_t = \rho^*\mathbf{W}\mathbf{y}_t + \mathbf{X}_\mathbf{t}^*\beta^* + v_t$$

where v_t are the disturbances.

Partial Adjustment Model

Let's consider the properties of $\mathbf{y}_t = \rho^* \mathbf{W}\mathbf{y}_t + \mathbf{X}^*_\mathbf{t}\beta^* + \nu_t$.

- The spatial autoregressive parameter ρ is amplified by $\psi - \tau$, so that values of $\psi > 1$ (implying growth in $\mathbf{X}$) reduce the estimated spatial dependence of the system as measured by ρ^*. This gives more weight to the present.
- Lower values of $\psi < 1$ (similarly, higher values of the temporal parameter τ) increase the role of the past, allowing more time for spatial influences to develop.
- $\therefore$ even correctly-specified cross-sectional and spatiotemporal models could yield very different estimates of spatial dependence:
  - Cross-sectional samples place more emphasis on a long-run equilibrium result of a spatiotemporal process (i.e.: high spatial dependence).
  - Longitudinal samples place more emphasis on the temporal dependence parameters (i.e.: low spatial dependence).
  - But a process with low spatial dependence and high temporal dependence may still imply a long-run equilibrium with high spatial dependence.