Spatial Data and Analysis in R

A PRISM Workshop

Adam Lauretig

The Ohio State University

Outline

Introduction

Spatial Data Prep

Spatial Autocorrelation

Regression

Spatial Regression

Discussion

Why Are We Here?

- Tobler's first law of geography:

Why Are We Here?

- Tobler's first law of geography:
- "Everything is related to everything else, but near things are more related than distant things"

Why Are We Here?

- Tobler's first law of geography:
- "Everything is related to everything else, but near things are more related than distant things"
- We want to quantify how the spatial relationship between our observations affect our inferences

A Caveat

- There are entire disciplines which study these issues (one of them is downstairs)

A Caveat

- There are entire disciplines which study these issues (one of them is downstairs)
- I will be introducing *spatial statistics* with a touch of *GIS*

A Caveat

- There are entire disciplines which study these issues (one of them is downstairs)
- I will be introducing *spatial statistics* with a touch of *GIS*
- I will not be discussing *GIS* in depth, nor will I discuss remote sensing *at all*

What are Spatial Data?

- Information (attributes) associated with a location

What are Spatial Data?

- Information (attributes) associated with a location
- Many kinds of spatial data: Points, Lines, Polygons, Raster data

What are Spatial Data?

- Information (attributes) associated with a location
- Many kinds of spatial data: Points, Lines, Polygons, Raster data
- Today, we are working with polygon data

Prepping our data

- Spatial data come in *shapefiles* which are really mini-databases

Prepping our data

- Spatial data come in *shapefiles* which are really mini-databases
- ORDBMS - Linking spatial and attribute data

Prepping our data

- Spatial data come in *shapefiles* which are really mini-databases
- ORDBMS - Linking spatial and attribute data
- Six parts, all combine to create a map to represent data

Loading Our data

Percent Black

```
## Error in eval(expr, envir, enclos):  could not find function "choropleth"
```

AIDs Rate per 1000 people

```
## Error in eval(expr, envir, enclos):  could not find function "choropleth"
```

Measuring Spatial Autocorrelation

- What is spatial autocorrelation?

I's formula is:

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

where w_{ij} is the weight between observation i and j

Measuring Spatial Autocorrelation

- What is spatial autocorrelation?
- Observations with more similar values tend to occur more closely together

I's formula is:

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

where w_{ij} is the weight between observation i and j

Measuring Spatial Autocorrelation

- What is spatial autocorrelation?
- Observations with more similar values tend to occur more closely together
- Most common test: Moran's I:

I's formula is:

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

where w_{ij} is the weight between observation i and j

The Weights Matrix W

- In Moran's I, there was this thing w_{ij}

The Weights Matrix W

- In Moran's I, there was this thing w_{ij}
- This is the *weights matrix*

The Weights Matrix W

- In Moran's I, there was this thing w_{ij}
- This is the *weights matrix*
- It allows us to measure the effect neighboring observations j have on our observation of interest i

The Weights Matrix W

- In Moran's I, there was this thing w_{ij}
- This is the *weights matrix*
- It allows us to measure the effect neighboring observations j have on our observation of interest i
- Can be specified in a variety of ways, the simplest of which is binary ("contiguity"): 1 if observations share a boundary, 0 if they do not

The Weights Matrix W

- In Moran's I, there was this thing w_{ij}
- This is the *weights matrix*
- It allows us to measure the effect neighboring observations j have on our observation of interest i
- Can be specified in a variety of ways, the simplest of which is binary (”contiguity”): 1 if observations share a boundary, 0 if they do not
- The default in R is "row standardized," where $w_{ij} = \frac{i}{\sum j}$

Creating a weights matrix in *R*

```
library(rgdal)
library(spdep)
library(sp)
library(spatstat)
file_path <- "/Users/adamlauretig/data/prism_presentation/NYAIDS_data"
ny <- readOGR(dsn = file_path,
              layer = "NYAIDS", verbose=FALSE)
nygal <- poly2nb(ny) #Create the neighborhood object
nyQ1.gal <- nb2listw(nygal, zero.policy=T) #Create the weights object
```

Running the Moran's I

```
## Error in ogrInfo(dsn = dsn, layer = layer, encoding = encoding, use_iconv = use_iconv, :  Cannot open data source


##
##  Moran I test under randomisation
##
## data:  log1p(ny$Rate1000)
## weights: nyQ1.gal
##
## Moran I statistic standard deviate = 13.044, p-value < 2.2e-16
## alternative hypothesis: two.sided
## sample estimates:
## Moran I statistic       Expectation          Variance
##       0.682998774      -0.005780347       0.002788498
```

Where are these clusters?

- We can calculate this using a *Local Indicator of Spatial Autocorrelation* (LISA)

Where are these clusters?

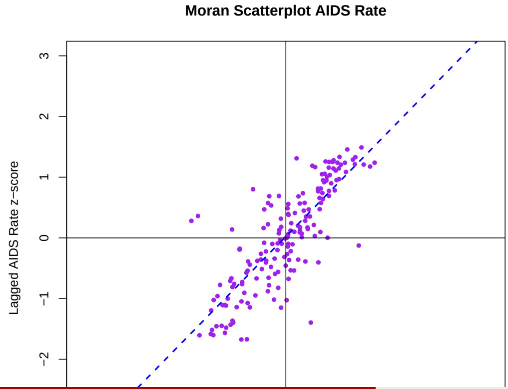
- We can calculate this using a *Local Indicator of Spatial Autocorrelation* (LISA)
- Measure how similar a value is compared to neighboring values

Where are these clusters?

- We can calculate this using a *Local Indicator of Spatial Autocorrelation* (LISA)
- Measure how similar a value is compared to neighboring values
- While the Moran's I detects clustering, the LISA detects *clusters*

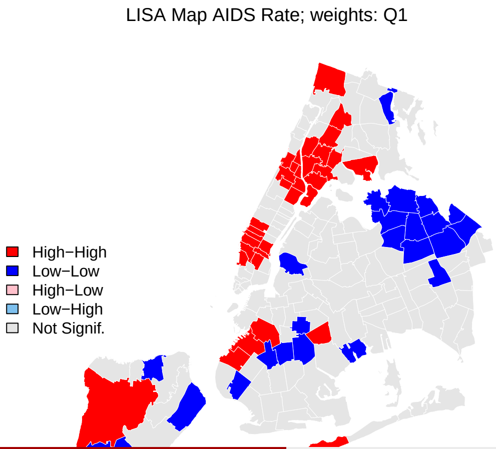
Where are these clusters?

```
## Error in ogrInfo(dsn = dsn, layer = layer, encoding = encoding,
use_iconv = use_iconv, :  Cannot open data source
```



Where are these clusters?

```
## Error in ogrInfo(dsn = dsn, layer = layer, encoding = encoding, use_iconv = use_iconv, :  Cannot open data source
```



What About Regression with Spatial Data?

- Remember, a key regression assumption is that our observations are IID

What About Regression with Spatial Data?

- Remember, a key regression assumption is that our observations are IID
- If there is spatial autocorrelation in our data, we have fewer independent data points than we initially supposed

What About Regression with Spatial Data?

- Remember, a key regression assumption is that our observations are IID
- If there is spatial autocorrelation in our data, we have fewer independent data points than we initially supposed
- This shrinks our standard errors, making variables appear significant

What About Regression with Spatial Data?

- Remember, a key regression assumption is that our observations are IID
- If there is spatial autocorrelation in our data, we have fewer independent data points than we initially supposed
- This shrinks our standard errors, making variables appear significant
- Once we adjust our variance-covariance matrix, previously significant covariates might lose their significance

What About Regression with Spatial Data?

- Remember, a key regression assumption is that our observations are IID
- If there is spatial autocorrelation in our data, we have fewer independent data points than we initially supposed
- This shrinks our standard errors, making variables appear significant
- Once we adjust our variance-covariance matrix, previously significant covariates might lose their significance
- Addtionally, with spatial autocorrelation, our coefficients may be biased

What About Regression with Spatial Data?

- Remember, a key regression assumption is that our observations are IID
- If there is spatial autocorrelation in our data, we have fewer independent data points than we initially supposed
- This shrinks our standard errors, making variables appear significant
- Once we adjust our variance-covariance matrix, previously significant covariates might lose their significance
- Addtionally, with spatial autocorrelation, our coefficients may be biased
- Two ways of handling this: Spatial Error Models, and Spatial Autoregressive models

The Math: Spatial Error Model

- Normally: $y_i = x_i\beta + e_i$, where $e = I(Y - X\beta)$

The Math: Spatial Error Model

- Normally: $y_i = x_i\beta + e_i$, where $e = I(Y - X\beta)$
- But, we want to model spatial dependence in the residuals

The Math: Spatial Error Model

- Normally: $y_i = x_i\beta + e_i$, where $e = I(Y - X\beta)$
- But, we want to model spatial dependence in the residuals
- $e_i = \sum_{j=1}^{n} w_{ij} + e_j + \varepsilon_i$, where $w_{ii} = 0$

The Math: Spatial Error Model

- Normally: $y_i = x_i\beta + e_i$, where $e = I(Y - X\beta)$
- But, we want to model spatial dependence in the residuals
- $e_i = \sum_{j=1}^{n} w_{ij} + e_j + \varepsilon_i$, where $w_{ii} = 0$
- Basically, we regress the error e_i on the surrounding errors

The Math: Spatial Error Model

- Normally: $y_i = x_i\beta + e_i$, where $e = I(Y - X\beta)$
- But, we want to model spatial dependence in the residuals
- $e_i = \sum_{j=1}^{n} w_{ij} + e_j + \varepsilon_i$, where $w_{ii} = 0$
- Basically, we regress the error e_i on the surrounding errors
- We wind up with $e = (I - W)(Y - X\beta)$

The Math: Spatial Error Model

- Normally: $y_i = x_i\beta + e_i$, where $e = I(Y - X\beta)$
- But, we want to model spatial dependence in the residuals
- $e_i = \sum_{j=1}^{n} w_{ij} + e_j + \varepsilon_i$, where $w_{ii} = 0$
- Basically, we regress the error e_i on the surrounding errors
- We wind up with $e = (I - W)(Y - X\beta)$
- ε is the residual of residuals, with $\sum_{\varepsilon} = \sigma^2 I$

The Math: Spatial Error Model

- Normally: $y_i = x_i\beta + e_i$, where $e = I(Y - X\beta)$
- But, we want to model spatial dependence in the residuals
- $e_i = \sum_{j=1}^{n} w_{ij} + e_j + \varepsilon_i$, where $w_{ii} = 0$
- Basically, we regress the error e_i on the surrounding errors
- We wind up with $e = (I - W)(Y - X\beta)$
- ε is the residual of residuals, with $\sum_{\varepsilon} = \sigma^2 I$
- The full model: $y_i = x_i\beta + \sum_{j=1}^{n} w_{ij}e_j + \varepsilon_i$

The Math: Spatial Autoregressive Model

- Option 2: the Spatial autoregressive model

The Math: Spatial Autoregressive Model

- Option 2: the Spatial autoregressive model
- Instead of lagging the error term, lag y, the DV

The Math: Spatial Autoregressive Model

- Option 2: the Spatial autoregressive model
- Instead of lagging the error term, lag y, the DV
- $y_i = x_i\beta + \sum_{j=1}^{n} w_{ij}y_j + \varepsilon_i$

The Math: Spatial Autoregressive Model

- Option 2: the Spatial autoregressive model
- Instead of lagging the error term, lag y, the DV
- $y_i = x_i\beta + \sum_{j=1}^{n} w_{ij}y_j + \varepsilon_i$
- SAR vs. SEM

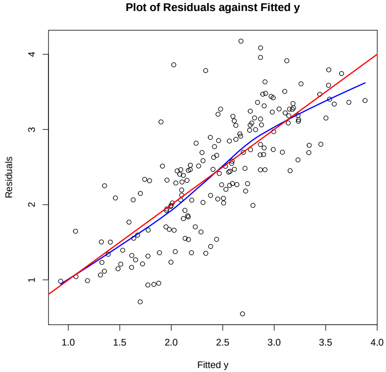
Plain OLS

```
## Error in ogrInfo(dsn = dsn, layer = layer, encoding = encoding, use_iconv = use_iconv, :  Cannot open data source


##
## Call:
## lm(formula = lrate ~ PctWht + PctHisp + Gini + PctHSEd + PctFemHH,
##     data = ny)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.07597 -0.31021 -0.02123  0.29746  1.81376
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.047530   0.796873  -3.824 0.000185 ***
## PctWht      -0.014813   0.002953  -5.016 1.33e-06 ***
## PctHisp      0.016042   0.003569   4.495 1.29e-05 ***
## Gini         7.973273   0.710959  11.215  < 2e-16 ***
## PctHSEd      0.023190   0.007328   3.165 0.001844 **
## PctFemHH     0.008062   0.008182   0.985 0.325885
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5376 on 168 degrees of freedom
## Multiple R-squared:  0.5678,Adjusted R-squared:  0.5549
## F-statistic: 44.14 on 5 and 168 DF,  p-value: < 2.2e-16
```

Did we model out our autocorrelation?

```
## Error in ogrInfo(dsn = dsn, layer = layer, encoding = encoding, use_iconv = use_iconv, :  Cannot open data source
```



Spatial Autocorrelation in our Residuals?

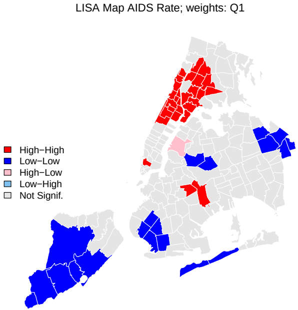
```
## Error in ogrInfo(dsn = dsn, layer = layer, encoding = encoding, use_iconv = use_iconv, :  Cannot open data source


##
##  Moran I test under randomisation
##
## data:  ny$resid
## weights: nyQ1.gal
##
## Moran I statistic standard deviate = 12.028, p-value < 2.2e-16
## alternative hypothesis: two.sided
## sample estimates:
## Moran I statistic       Expectation          Variance
##       0.637575327      -0.005847953       0.002861601
```

YES! Quite a bit of it

Where is the autocorrelation?

```
## Error in ogrInfo(dsn = dsn, layer = layer, encoding = encoding, use_iconv = use_iconv, :  Cannot open data source
```



Picking a model

```
## Error in ogrInfo(dsn = dsn, layer = layer, encoding = encoding, use_iconv = use_iconv, :  Cannot open data source

##
##  Lagrange multiplier diagnostics for spatial dependence
##
## data:
## model: lm(formula = lrate ~ PctWht + PctHisp + Gini + PctHSEd +
## PctFemHH, data = ny)
## weights: nyQ1.gal
##
## LMerr = 52.978, df = 1, p-value = 3.374e-13
##
##
##  Lagrange multiplier diagnostics for spatial dependence
##
## data:
## model: lm(formula = lrate ~ PctWht + PctHisp + Gini + PctHSEd +
## PctFemHH, data = ny)
## weights: nyQ1.gal
##
## LMlag = 83.543, df = 1, p-value < 2.2e-16
```

Spatial Autoregression Model

```
## Error in ogrInfo(dsn = dsn, layer = layer, encoding = encoding, use_iconv = use_iconv, :  Cannot open data source


##
## Call:lagsarlm(formula = lrate ~ PctWht + PctHisp + Gini + PctHSEd +
##     PctFemHH, data = ny, listw = nyQ1.gal, method = "eigen",
##     quiet = TRUE, control = (pre_eig = W.eig))
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.0425965 -0.2328207  0.0019989  0.1662031  1.6535569
##
## Type: lag
## Coefficients: (asymptotic standard errors)
##               Estimate Std. Error z value  Pr(>|z|)
## (Intercept) -1.6099231  0.5363501 -3.0016  0.002685
## PctWht      -0.0090690  0.0019540 -4.6414 3.461e-06
## PctHisp      0.0070588  0.0024109  2.9278  0.003413
## Gini         4.3050352  0.5759844  7.4742 7.772e-14
## PctHSEd      0.0094072  0.0048753  1.9296  0.053660
## PctFemHH     0.0012643  0.0052510  0.2408  0.809733
##
## Rho: 0.61126, LR test value: 94.796, p-value: < 2.22e-16
## Asymptotic standard error: 0.048384
##     z-value: 12.634, p-value: < 2.22e-16
## Wald statistic: 159.61, p-value: < 2.22e-16
##
## Log likelihood: -68.52996 for lag model
## ML residual variance (sigma squared): 0.11787, (sigma: 0.34333)
```

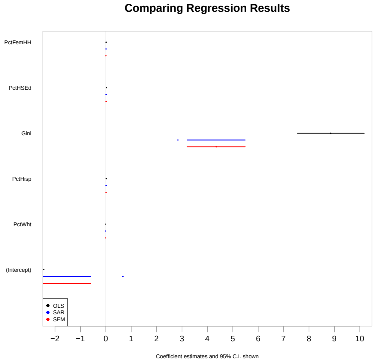
Spatial Error Model

```
## Error in ogrInfo(dsn = dsn, layer = layer, encoding = encoding, use_iconv = use_iconv, :  Cannot open data source

##
## Call:errorsarlm(formula = lrate ~ PctWht + PctHisp + Gini + PctHSEd +
##     PctFemHH, data = ny, listw = nyQ1.gal, method = "eigen",
##     quiet = TRUE)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.810183 -0.210778 -0.010702  0.183664  1.483237
##
## Type: error
## Coefficients: (asymptotic standard errors)
##               Estimate Std. Error z value  Pr(>|z|)
## (Intercept)  0.6482015  0.7293475  0.8887 0.3741419
## PctWht      -0.0094809  0.0026792 -3.5387 0.0004021
## PctHisp      0.0100471  0.0034434  2.9178 0.0035251
## Gini         2.8574821  0.7939676  3.5990 0.0003195
## PctHSEd      0.0061382  0.0059573  1.0304 0.3028343
## PctFemHH     0.0089076  0.0084238  1.0574 0.2903141
##
## Lambda: 0.80134, LR test value: 84.601, p-value: < 2.22e-16
## Asymptotic standard error: 0.040685
##     z-value: 19.696, p-value: < 2.22e-16
## Wald statistic: 387.94, p-value: < 2.22e-16
##
## Log likelihood: -73.03395 for error model
## ML residual variance (sigma squared): 0.1112, (sigma: 0.33347)
```

Comparing Findings - OLS

```
## Error in ogrInfo(dsn = dsn, layer = layer, encoding = encoding, use_iconv = use_iconv, :  Cannot open data source
## Warning in data.row.names(row.names, rowsi, i):  some row.names duplicated:  7,8,9,10,11,12,13,14,15,16,17,18 --> row.names NOT used
```



Did These Resolve our Spatial Autocorrelation?

```
## Error in ogrInfo(dsn = dsn, layer = layer, encoding = encoding, use_iconv = use_iconv, :  Cannot open data source


##
##  Moran I test under randomisation
##
## data:  ny.err.eig$resid
## weights: nyQ1.gal
##
## Moran I statistic standard deviate = -1.0922, p-value = 0.2748
## alternative hypothesis: two.sided
## sample estimates:
## Moran I statistic       Expectation          Variance
##      -0.065777589      -0.006060606       0.002989709
```

YES!

Workflow for Spatial Regression

- In addition to normal EDA, do some ESDA (Exploratory Spatial Data Analysis), mapping out variables

Workflow for Spatial Regression

- In addition to normal EDA, do some ESDA (Exploratory Spatial Data Analysis), mapping out variables
- Check for spatial autocorrelation

Workflow for Spatial Regression

- In addition to normal EDA, do some ESDA (Exploratory Spatial Data Analysis), mapping out variables
- Check for spatial autocorrelation
- Run your normal regression, with the variables you think are necessary

Workflow for Spatial Regression

- In addition to normal EDA, do some ESDA (Exploratory Spatial Data Analysis), mapping out variables
- Check for spatial autocorrelation
- Run your normal regression, with the variables you think are necessary
- Check once more for spatial autocorrelation, in your residuals

Workflow for Spatial Regression

- In addition to normal EDA, do some ESDA (Exploratory Spatial Data Analysis), mapping out variables
- Check for spatial autocorrelation
- Run your normal regression, with the variables you think are necessary
- Check once more for spatial autocorrelation, in your residuals
- If there's still autocorrelation, run a spatial model

Workflow for Spatial Regression

- In addition to normal EDA, do some ESDA (Exploratory Spatial Data Analysis), mapping out variables
- Check for spatial autocorrelation
- Run your normal regression, with the variables you think are necessary
- Check once more for spatial autocorrelation, in your residuals
- If there's still autocorrelation, run a spatial model
- One final check for autocorrelation in your residuals

Where Can we Apply These Methods in Political Science

- Voting and Political Behavior patterns (Data available at the Census Tract level (or less))

Where Can we Apply These Methods in Political Science

- Voting and Political Behavior patterns (Data available at the Census Tract level (or less))
- Agricultural/industrial data (economic output)

Where Can we Apply These Methods in Political Science

- Voting and Political Behavior patterns (Data available at the Census Tract level (or less))
- Agricultural/industrial data (economic output)
- Conflict/Political Violence data (ex: ACLED)

Where Can we Apply These Methods in Political Science

- Voting and Political Behavior patterns (Data available at the Census Tract level (or less))
- Agricultural/industrial data (economic output)
- Conflict/Political Violence data (ex: ACLED)
- Anything with an address or lon/lat coordinates can be georeferenced

Where Can we Apply These Methods in Political Science

- Voting and Political Behavior patterns (Data available at the Census Tract level (or less))
- Agricultural/industrial data (economic output)
- Conflict/Political Violence data (ex: ACLED)
- Anything with an address or lon/lat coordinates can be georeferenced
-

Other tools

- CAR Models

Other tools

- CAR Models
- Spatial/Spatio-temporal scan statistics

Other tools

- CAR Models
- Spatial/Spatio-temporal scan statistics
- Geographically Weighted Regression

Other tools

- CAR Models
- Spatial/Spatio-temporal scan statistics
- Geographically Weighted Regression
- Kernel Density Estimation

Other tools

- CAR Models
- Spatial/Spatio-temporal scan statistics
- Geographically Weighted Regression
- Kernel Density Estimation
- Kriging/Geostatistics

Other tools

- CAR Models
- Spatial/Spatio-temporal scan statistics
- Geographically Weighted Regression
- Kernel Density Estimation
- Kriging/Geostatistics
- Spatio-temporal approaches

Resources

- Yuri Zhukov's Spatial Workshop:
  http://www.people.fas.harvard.edu/~zhukov/spatial.html

Resources

- Yuri Zhukov's Spatial Workshop: `http://www.people.fas.harvard.edu/~zhukov/spatial.html`
- Brunsdon, Chris, and Lex Comber. *An introduction to R for spatial analysis & mapping*. Sage, 2015.

Resources

- Yuri Zhukov's Spatial Workshop: `http://www.people.fas.harvard.edu/~zhukov/spatial.html`
- Brunsdon, Chris, and Lex Comber. *An introduction to R for spatial analysis & mapping.* Sage, 2015.
- Bivand, Roger S., and Edzer J. Pebesma. *Applied Spatial Data Analysis with R.* Springer, 2013.

Resources

- Yuri Zhukov's Spatial Workshop:
  `http://www.people.fas.harvard.edu/~zhukov/spatial.html`
- Brunsdon, Chris, and Lex Comber. *An introduction to R for spatial analysis & mapping.* Sage, 2015.
- Bivand, Roger S., and Edzer J. Pebesma. *Applied Spatial Data Analysis with R.* Springer, 2013.
- Waller, Lance A., and Carol A. Gotway. *Applied spatial statistics for public health data.* John Wiley & Sons, 2004.

Resources

- Yuri Zhukov's Spatial Workshop: `http://www.people.fas.harvard.edu/~zhukov/spatial.html`
- Brunsdon, Chris, and Lex Comber. *An introduction to R for spatial analysis & mapping.* Sage, 2015.
- Bivand, Roger S., and Edzer J. Pebesma. *Applied Spatial Data Analysis with R.* Springer, 2013.
- Waller, Lance A., and Carol A. Gotway. *Applied spatial statistics for public health data.* John Wiley & Sons, 2004.
- Cressie, Noel. *Statistics for spatial data.* John Wiley & Sons, 1993.

Acknowledgements

Thank you to Will Massengill and Drew Rosenberg for this opportunity, and to Elisabeth Root, whose code I drew heavily from

References