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

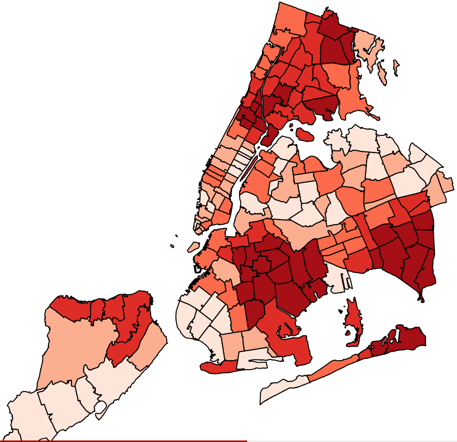
- Spatial data come in *shapefiles* which are really mini-databases
- ORDBMS - Linking spatial and attribute data

Prepping our data

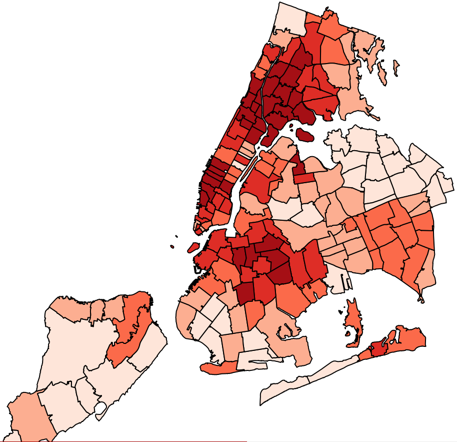
- Spatial data come in *shapefiles* which are really mini-databases
- ORDBMS - Linking spatial and attribute data
- Six parts, all combine to create a map to represent data

Loading Our data

Percent Black



AIDs Rate per 1000 people



Measuring Spatial Autocorrelation

- What is spatial autocorrelation?

Measuring Spatial Autocorrelation

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

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:

Formula for Moran's I

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

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_stuff/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

Moran I statistic	Expectation	Variance	p-value
0.68	-0.01	0.00	0.00

Where are these clusters?

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

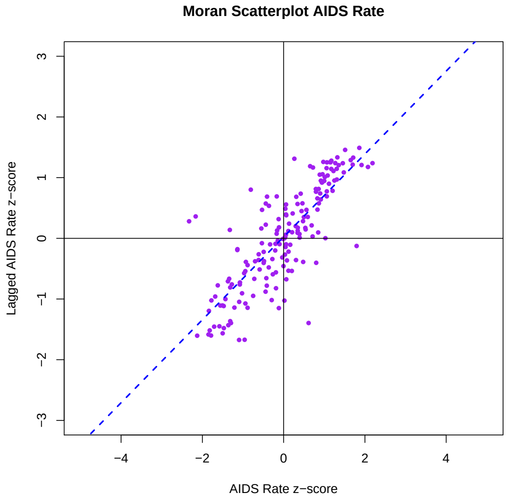
Where are these clusters?

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

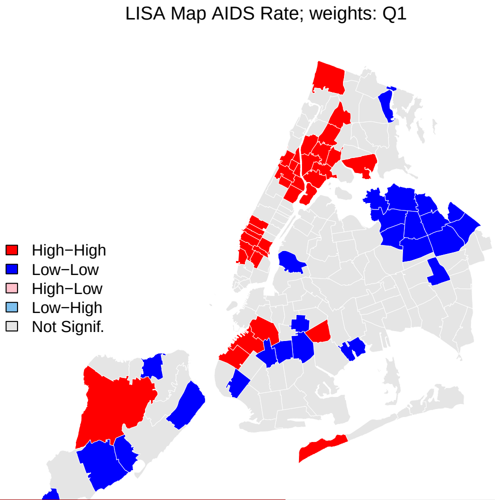
Where are these clusters?

- 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?



Where are these clusters?



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

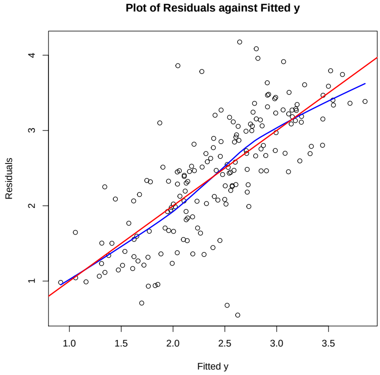
Plain OLS

	Model 1
(Intercept)	-3.05^{***}
	(0.80)
PctWht	-0.01^{***}
	(0.00)
PctHisp	0.02^{***}
	(0.00)
Gini	7.97^{***}
	(0.71)
PctHSEd	0.02^{**}
	(0.01)
PctFemHH	0.01
	(0.01)
R^2	0.57
Adj. R^2	0.55
Num. obs.	174
RMSE	0.54

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table: Statistical models

Did we model out our autocorrelation?

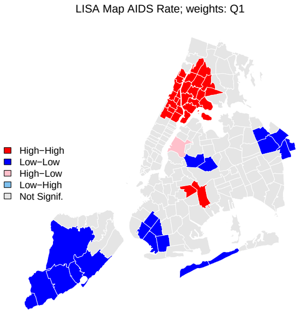


Did we model out our spatial autocorrelation?

Moran I statistic	Expectation	Variance	p-value
0.63	-0.01	0.00	0.00

No! This statistic didn't change much from before we ran our regression: .68 vs .63

Where is the residual autocorrelation?



Can we Model this Spatial Autocorrelation

- One option is to use fixed effects

Can we Model this Spatial Autocorrelation

- One option is to use fixed effects
- But this removes something interesting – the spatial relationship

Can we Model this Spatial Autocorrelation

- One option is to use fixed effects
- But this removes something interesting – the spatial relationship
- Another option is to explicitly model the spatial relationship

Spatial Autoregressive Model

	Model 1
(Intercept)	-1.52^{*}
	(0.63)
PctWht	-0.01^{***}
	(0.00)
PctHisp	0.01^{**}
	(0.00)
Gini	4.48^{***}
	(0.64)
PctHSEd	0.01
	(0.01)
PctFemHH	0.00
	(0.01)
ρ	0.54^{***}
	(0.06)
Num. obs.	174
Parameters	8
Log Likelihood	-101.93
AIC (Linear model)	285.72
AIC (Spatial model)	219.85
LR test: statistic	67.86
LR test: p-value	0.00

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table: Statistical models

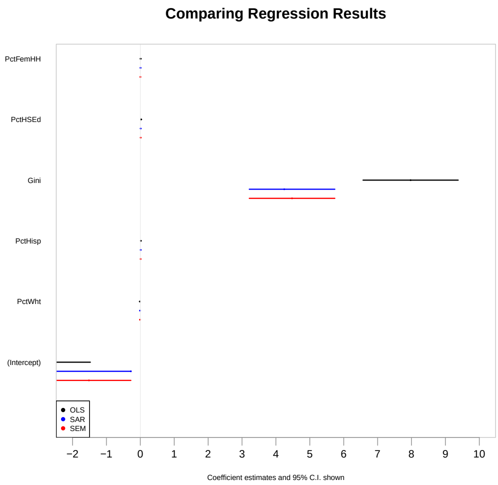
Spatial Error Model

	Model 1
(Intercept)	-0.28
	(0.88)
PctWht	-0.01^{**}
	(0.00)
PctHisp	0.01^{**}
	(0.00)
Gini	4.25^{***}
	(0.90)
PctHSEd	0.01
	(0.01)
PctFemHH	0.01
	(0.01)
λ	0.66^{***}
	(0.06)
Num. obs.	174
Parameters	8
Log Likelihood	-107.85
AIC (Linear model)	285.72
AIC (Spatial model)	231.70
LR test: statistic	56.01
LR test: p-value	0.00

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table: Statistical models

Comparing Findings - OLS



Did These Resolve our Spatial Autocorrelation?

Moran I statistic	Expectation	Variance	p-value
-0.10	-0.01	0.00	0.08

YES!

Conditional Autoregressive Model (CAR)

- Model the spatial relationship according to a Markov Random Field (MRF)

Conditional Autoregressive Model (CAR)

- Model the spatial relationship according to a Markov Random Field (MRF)
- Each observation depends only on values at neighboring locations, not glocal values

Conditional Autoregressive Model (CAR)

- Model the spatial relationship according to a Markov Random Field (MRF)
- Each observation depends only on values at neighboring locations, not glocal values
- Written out:

Conditional Autoregressive Model (CAR)

- Model the spatial relationship according to a Markov Random Field (MRF)
- Each observation depends only on values at neighboring locations, not glocal values
- Written out:
- c_{ij} is the weight matrix

Conditional Autoregressive Model (CAR)

- Model the spatial relationship according to a Markov Random Field (MRF)
- Each observation depends only on values at neighboring locations, not glocal values
- Written out:
- c_{ij} is the weight matrix
- $$E(y_i|y_{-i}) = x_i\beta + \sum_{j=1}^{n} +c_{ij}[y_j - x_i\beta]$$

Conditional Autoregressive Model (CAR)

- Model the spatial relationship according to a Markov Random Field (MRF)
- Each observation depends only on values at neighboring locations, not glocal values
- Written out:
- c_{ij} is the weight matrix
- $$E(y_i|y_{-i}) = x_i\beta + \sum_{j=1}^{n} + c_{ij}[y_j - x_i\beta]$$
- $$\text{var}(y_i|y_{-i}) = \sigma_i^2$$

Conditional Autoregressive Model (CAR) (cont'd)

- This defines a joint multivariate normal distribution with variance

Conditional Autoregressive Model (CAR) (cont'd)

- This defines a joint multivariate normal distribution with variance
- $$\sum_{Y} = (I - C)^{-1} \sum_{C}$$

Conditional Autoregressive Model (CAR) (cont'd)

- This defines a joint multivariate normal distribution with variance
- $$\sum_{Y} = (I - C)^{-1} \sum_{C}$$
- $$\sum_{CAR} = \sigma^2 (I - C)^{-1} V_C$$

Conditional Autoregressive Model (CAR) (cont'd)

- This defines a joint multivariate normal distribution with variance
- $$\sum_Y = (I - C)^{-1} \sum_C$$
- $$\sum_{CAR} = \sigma^2 (I - C)^{-1} V_C$$
- $$= \sigma^2 V_{CAR}$$

SAR vs. CAR

- SAR:

$$y_i \sim N(0, (I - W)^{-1} D^{\sim} (I - W)'^{-1})$$

$$\sum_{SAR} = \sigma^2 (I - W)^{-1} V_{\epsilon} (I - W')^{-1}$$

SAR vs. CAR

- SAR:

$$y_i \sim N(0, (I - W)^{-1} D^{\sim} (I - W)'^{-1})$$

$$\sum_{SAR} = \sigma^2 (I - W)^{-1} V_{\epsilon} (I - W')^{-1}$$

- CAR:

$$y_i \sim N(0, (I - C)^{-1} D)$$

$$\sum_{CAR} = \sigma^2 (I - C)^{-1} V_C$$

SAR vs. CAR

- CAR only requires the variance to estimate, SAR requires the full variance-covariance matrix

SAR vs. CAR

- CAR only requires the variance to estimate, SAR requires the full variance-covariance matrix
- CAR is computationally less intensive

SAR vs. CAR

- CAR only requires the variance to estimate, SAR requires the full variance-covariance matrix
- CAR is computationally less intensive
- CAR: local/first order dependency

SAR vs. CAR

- CAR only requires the variance to estimate, SAR requires the full variance-covariance matrix
- CAR is computationally less intensive
- CAR: local/first order dependency
- CAR: fit with MCMC – can fit GLMs

SAR vs. CAR

- CAR only requires the variance to estimate, SAR requires the full variance-covariance matrix
- CAR is computationally less intensive
- CAR: local/first order dependency
- CAR: fit with MCMC – can fit GLMs
- SAR: global/higher order dependency

SAR vs. CAR

- CAR only requires the variance to estimate, SAR requires the full variance-covariance matrix
- CAR is computationally less intensive
- CAR: local/first order dependency
- CAR: fit with MCMC – can fit GLMs
- SAR: global/higher order dependency
- SAR: fit with MLE

Fitting a CAR (1)

	Median	2.5%	97.5%	n.sample	% accept	n.effective	Geweke.diag
(Intercept)	-0.60	-2.35	1.08	18000.00	100.00	2417.70	0.50
PctWht	-0.01	-0.02	-0.01	18000.00	100.00	2973.70	1.40
PctHisp	0.01	0.01	0.02	18000.00	100.00	2592.00	-0.80
Gini	3.98	2.10	5.97	18000.00	100.00	1628.90	-0.30
PctHSEd	0.02	0.00	0.03	18000.00	100.00	3252.90	-0.90
PctFemHH	0.01	-0.01	0.03	18000.00	100.00	2006.10	0.20
nu2	0.01	0.00	0.06	18000.00	100.00	1599.50	0.70
tau2	0.50	0.34	0.65	18000.00	100.00	3907.90	-0.80
rho	0.81	0.61	0.94	18000.00	44.10	5196.60	0.90

Table: uninformative priors

Fitting a CAR (2)

	Median	2.5%	97.5%	n.sample	% accept	n.effective	Geweke.diag
(Intercept)	-0.81	-2.48	0.86	18000.00	100.00	11033.20	1.40
PctWht	-0.01	-0.02	-0.01	18000.00	100.00	13600.70	0.20
PctHisp	0.01	0.01	0.02	18000.00	100.00	13780.40	0.30
Gini	4.03	2.12	5.93	18000.00	100.00	7508.30	-2.00
PctHSEd	0.02	0.00	0.03	18000.00	100.00	15248.50	-0.70
PctFemHH	0.01	-0.01	0.03	18000.00	100.00	10575.60	1.30
nu2	0.08	0.06	0.13	18000.00	100.00	10732.90	-1.30
tau2	0.25	0.16	0.38	18000.00	100.00	10628.20	2.70
rho	0.91	0.75	0.98	18000.00	43.80	10169.70	0.50

Table: weakly informative priors

Geographically Weighted Regression

- A non-parametric method for inference with non-stationary spatial data

Geographically Weighted Regression

- A non-parametric method for inference with non-stationary spatial data
- Non-stationarity: spatial variation is not constant across the data

Geographically Weighted Regression

- A non-parametric method for inference with non-stationary spatial data
- Non-stationarity: spatial variation is not constant across the data
- Use kernel regression to model local spatial variation

Geographically Weighted Regression

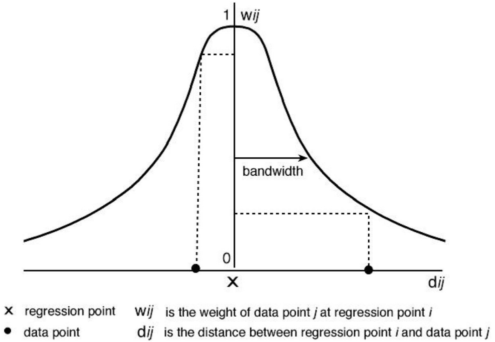
- A non-parametric method for inference with non-stationary spatial data
- Non-stationarity: spatial variation is not constant across the data
- Use kernel regression to model local spatial variation
- A weighted moving window regression:

$$Y(x) = \alpha(u_i, v_i) + \sum_k \beta_k(u_i, v_i)x_k + e_i$$

Geographically Weighted Regression

- A Kernel as a way of weighting observations which are closer together

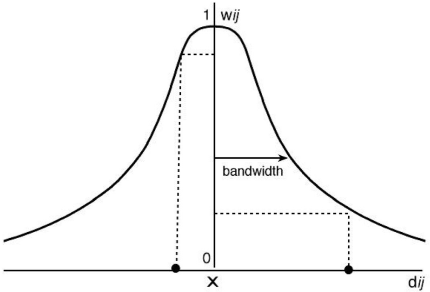
The concept of a kernel



Geographically Weighted Regression

- A Kernel as a way of weighting observations which are closer together
- Usually, fit adaptively – different "sizes" depending on the density of the data

The concept of a kernel



× regression point *wij* is the weight of data point *j* at regression point *i*

• data point *dij* is the distance between regression point *i* and data point *j*

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

Other tools

- Spatial/Spatio-temporal scan statistics

Other tools

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

Other tools

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

Other tools

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

Other tools

- 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 Jan Box-Steffensmeier for this opportunity, and to Elisabeth Root, whose code I drew heavily from.

References