An Introduction to Spatial Autoregressive Modeling

Data Expeditions

January 18th, 2020

Jonathan Holt

Reza Momenifar

Outline

Introduction

Spatial autocorrelation

Case Study

Introduction

Task: Predict student test scores using a linear regression model

Task: Predict student test scores using a linear regression model

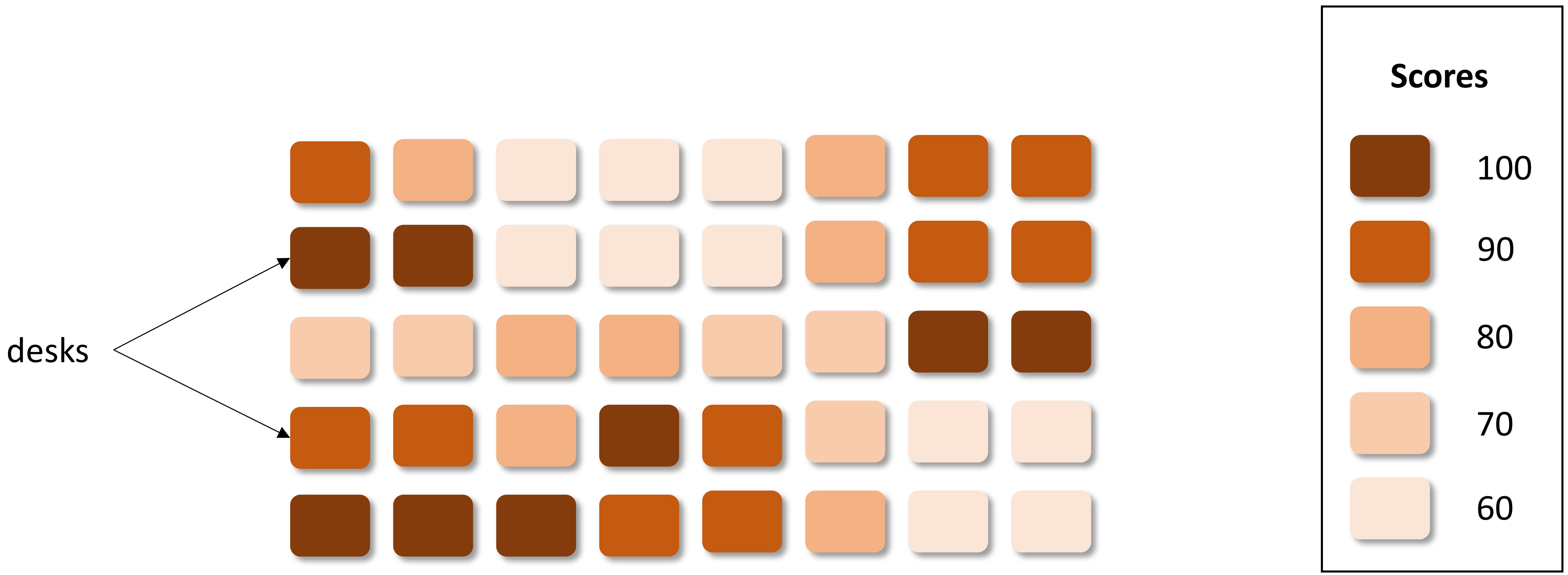
You are responsible for developing a linear regression model that predicts student test scores.

Activity: For 3 minutes, discuss with your neighbor the the most important predictor variables for your model.

Now, let’s look at the actual data

Imagine a classroom at UNC…

Students just received their final exams.



Front of classroom

Now, can you think of a better model to predict student test scores?

- Activity: For 3 minutes, discuss with your neighbor any changes that you would make to your original model.

How about:

$$score \sim \beta_0 + (\beta_1 * IQ) + (\beta_2 * hours\ studied) + (\beta_3 * score\ of\ neighbor) + \varepsilon$$

Notice that the response variable (score) is on **both** sides of the equation.

Task: Predict height of children using a linear regression model

Task: Predict height of children using a linear regression model

You are responsible for developing a regression model that predicts the height of a child.

Question: Name 3 predictors that you should include in your model.

How about:

$$height \sim \beta_0 + (\beta_1 * age) + (\beta_2 * \, gender) + (\beta_3 * \, height \; of \; child \; last \; year) + \varepsilon$$

Notice that the response variable (height) is on **both** sides of the equation.

“Autoregressive”

What do these two models have in common?

$$score \sim \beta_0 + (\beta_1 * IQ) + (\beta_2 * hours\ studied) + (\beta_3 * score\ of\ neighbor) + \varepsilon$$

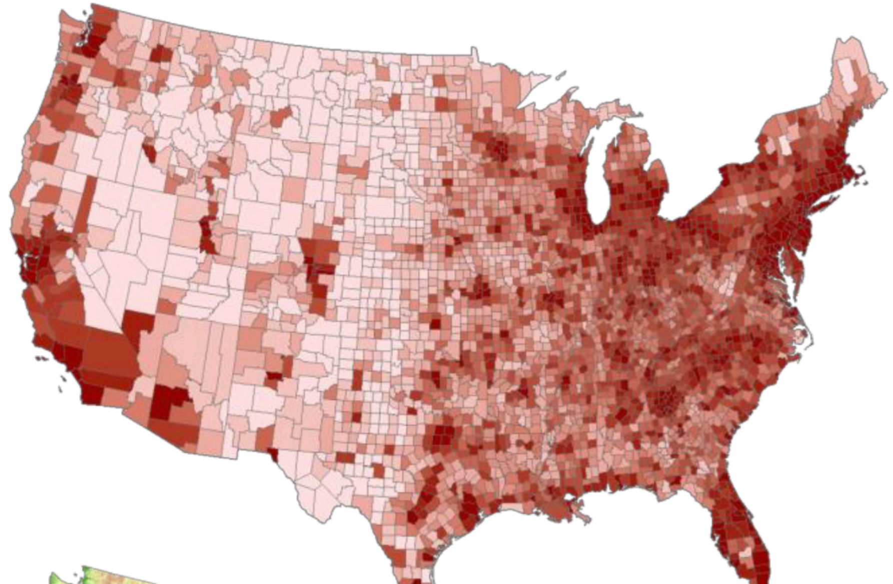
$$height \sim \beta_0 + (\beta_1 * age) + (\beta_2 * gender) + (\beta_3 * height\ of\ child\ last\ year) + \varepsilon$$

“Autoregressive”

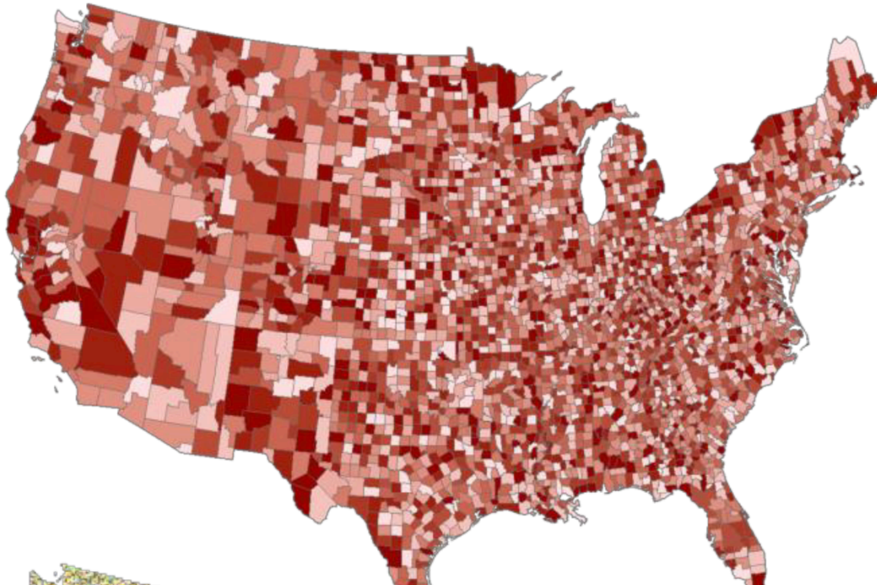
Spatial Autocorrelation

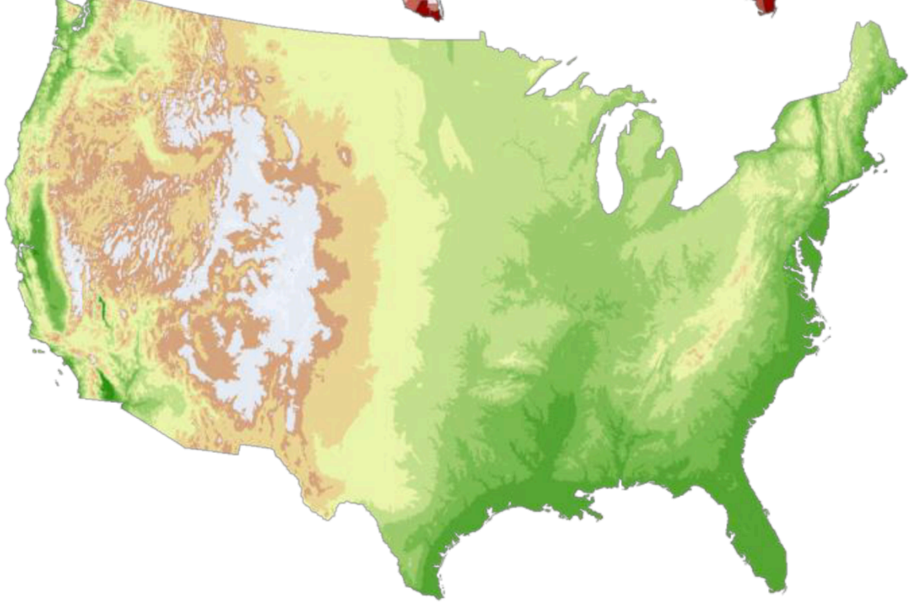
“The first law of geography: Everything is related to everything else, but near things are more related than distant things.” Waldo R. Tobler (Tobler 1970)

If features were randomly distributed ...

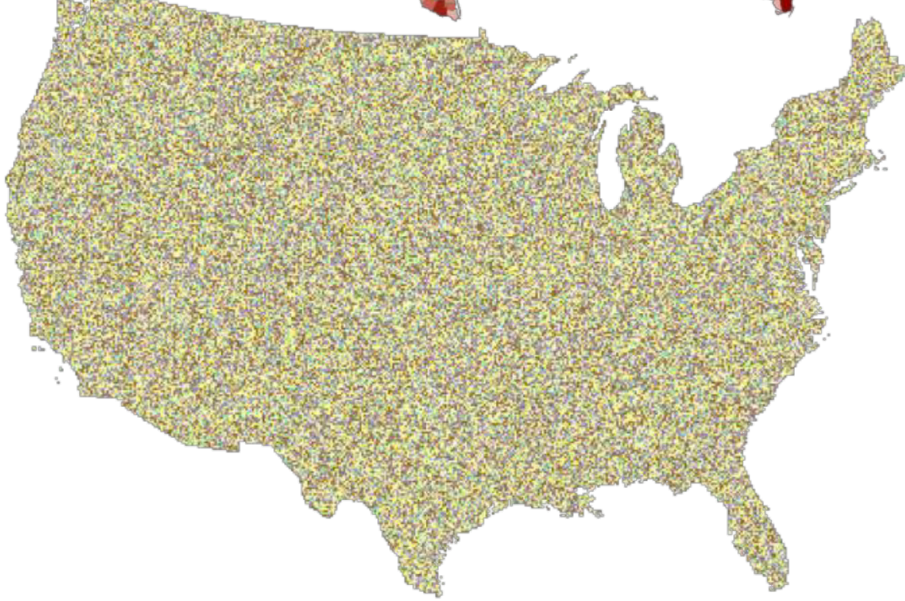


... population density map of the US would look like this





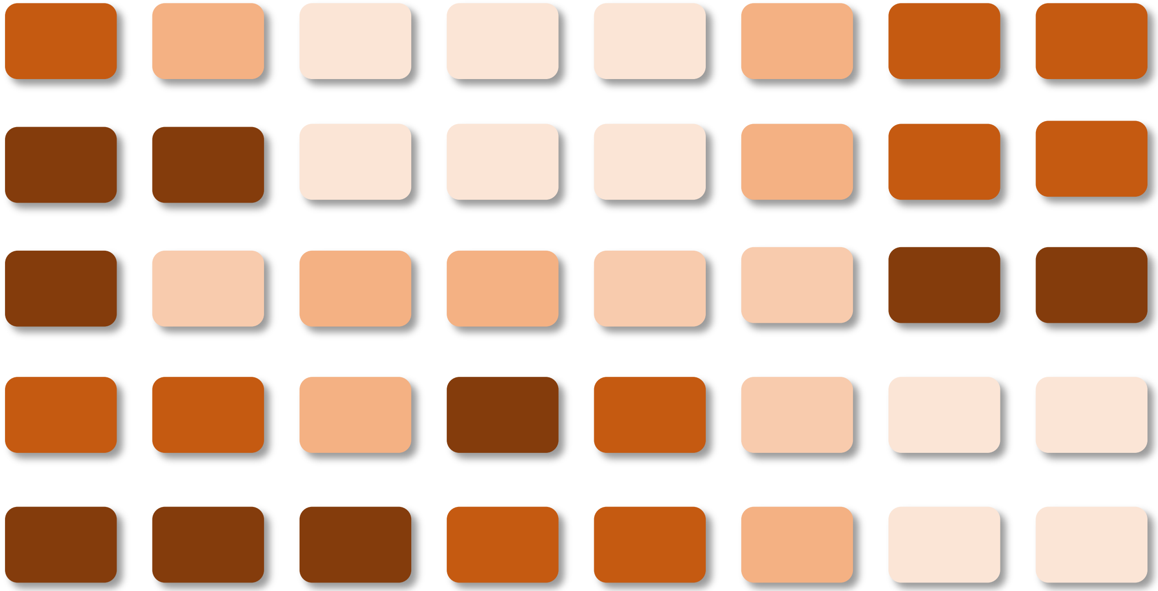
... elevation map of the US would look like this



Credit: Manuel Gimond

Spatial autoregressive modeling

- Spatial autoregressive models are models that account for **spatial autocorrelation** among observations (i.e., the response variable is not randomly distributed in space).



Vocabulary:

Correlation is between two **different** variables.

Autocorrelation is between the **same** variable at different spaces or times.

Examples of data with spatial autocorrelation

- Political elections
- Contaminant transfer
- Disease spread
- Housing market
- Weather

Recall the similarities between spatial and temporal autocorrelation

- How would you model the height of a growing child?

$$height \sim \beta_0 + (\beta_1 * age) + (\beta_2 * sex) + (\beta_3 * height\ previous\ year) + \varepsilon$$

Similar to

$$score \sim \beta_0 + (\beta_1 * IQ) + (\beta_2 * hours\ studied) + (\beta_3 * score\ of\ neighbor) + \varepsilon$$

In fact, many types of data are **spatially** *and* **temporally** autocorrelated

- Political elections
- Contaminant transfer
- Disease spread
- Housing market
- Weather

$$Rain\ in\ Durham\ at\ 2pm \sim \beta_0 + (\beta_1 * rain\ at\ 1pm) + (\beta_2 * rain\ in\ Hillsborough) + \cdots$$

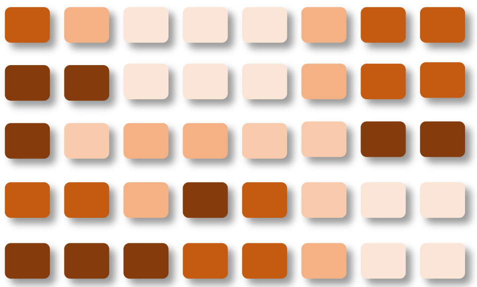
How do I know if my data are spatially autocorrelated?

- *Moran's I* test measures the spatial autocorrelation for continuous data

- $I = \frac{N}{W} \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}$

- *N* is the number of spatial units indexed by *i* and *j*
- *x* is the variable of interest; $\bar{x}$ is the mean of *x*
- *w* is a matrix of spatial weights
- W is the sum of all w_{ij}

Practice with the classroom test-score data

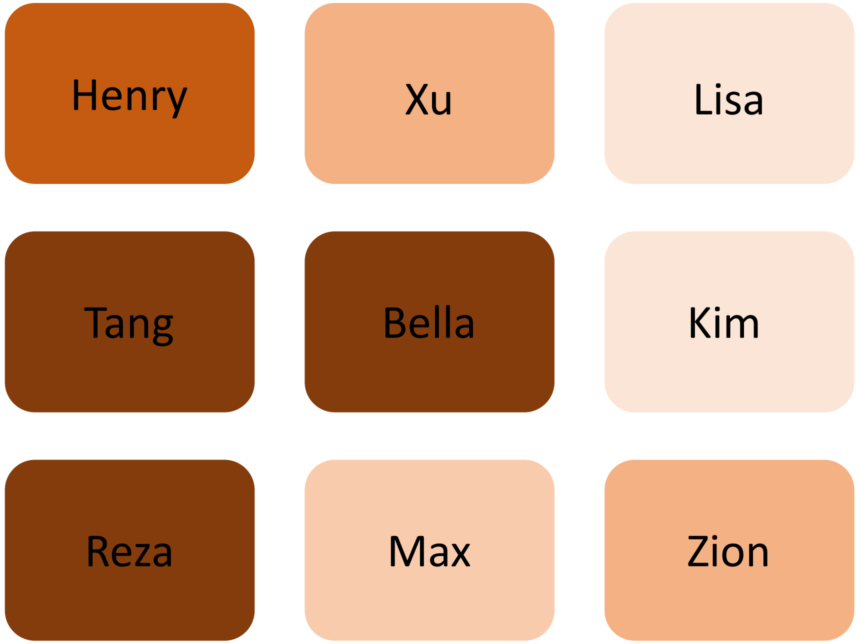


Henry 90	Xu 80	Lisa 60
Tang 100	Bella 100	Kim 60
Reza 100	Max 70	Zion 80

$\bar{x}$ = 82.22

For simplicity, consider these 9 students

Spatial weights matrix *w*

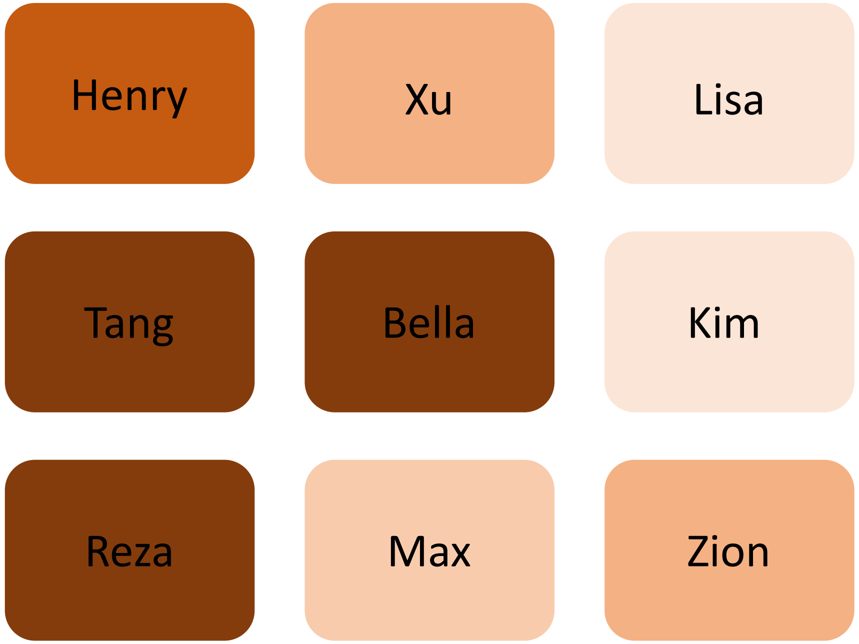


Henry	Xu	Lisa
Tang	Bella	Kim
Reza	Max	Zion

1 = adjacent
0 = not adjacent

	Henry	Xu	Lisa	Tang	Bella	Kim	Reza	Max	Zion
Henry									
Xu									
Lisa									
Tang									
Bella									
Kim									
Reza									
Max									
Zion									

Spatial weights matrix *w*

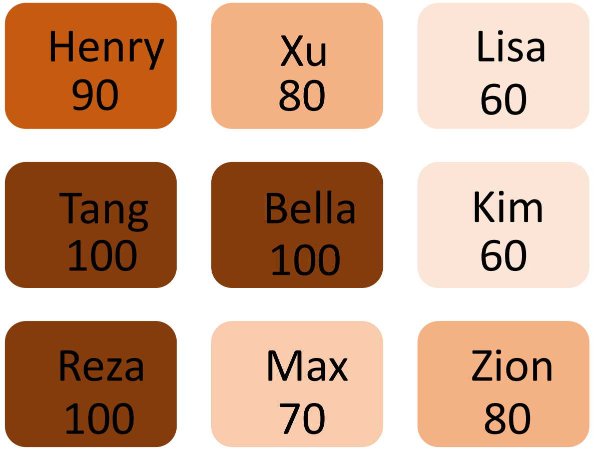


Henry	Xu	Lisa
Tang	Bella	Kim
Reza	Max	Zion

1 = adjacent
0 = not adjacent

	Henry	Xu	Lisa	Tang	Bella	Kim	Reza	Max	Zion
Henry	0	1	0	1	1	0	0	0	0
Xu	1	0	1	1	1	1	0	0	0
Lisa	0	1	0	0	1	1	0	0	0
Tang	1	1	0	0	1	0	1	1	0
Bella	1	1	1	1	0	1	1	1	1
Kim	0	1	1	0	1	0	0	1	1
Reza	0	0	0	1	1	0	0	1	0
Max	0	0	0	1	1	1	1	0	1
Zion	0	0	0	0	1	1	0	1	0

Putting it all together



Henry 90	Xu 80	Lisa 60
Tang 100	Bella 100	Kim 60
Reza 100	Max 70	Zion 80

$$I = \frac{N}{W}\frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

i

j

	Henry	Xu	Lisa	Tang	Bella	Kim	Reza	Max	Zion
Henry	0	1	0	1	1	0	0	0	0
Xu	1	0	1	1	1	1	0	0	0
Lisa	0	1	0	0	1	1	0	0	0
Tang	1	1	0	0	1	0	1	1	0
Bella	1	1	1	1	0	1	1	1	1
Kim	0	1	1	0	1	0	0	1	1
Reza	0	0	0	1	1	0	0	1	0
Max	0	0	0	1	1	1	1	0	1
Zion	0	0	0	0	1	1	0	1	0

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

Expected value for the null hypothesis

$$z = \frac{I - E(I)}{var(I)}$$

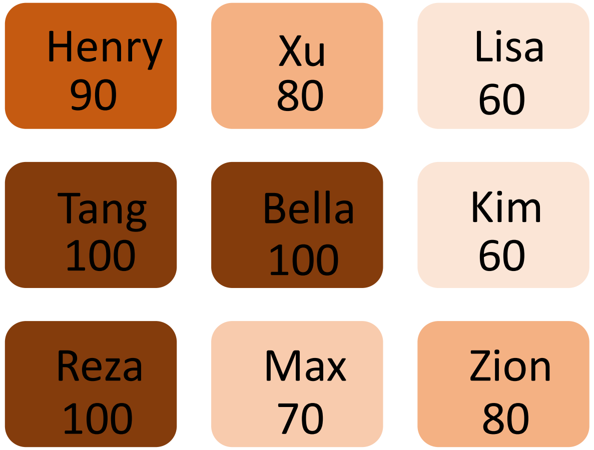
Z-score to test whether to reject null hypothesis

Interpreting Moran’s I

- In general,
  - ~ 1 means strong positive autocorrelation
  - ~ -1 means strong negative autocorrelation
  - ~ 0 means no autocorrelation
- We can do a hypothesis test to be sure… but we’ll use software for that.
  - Null hypothesis: I is (approximately) zero
  - Alternative hypothesis: I is greater or less than zero

R

Putting it all together



Henry 90	Xu 80	Lisa 60
Tang 100	Bella 100	Kim 60
Reza 100	Max 70	Zion 80

$$I = \frac{N}{W} \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

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

i

j

	Henry	Xu	Lisa	Tang	Bella	Kim	Reza	Max	Zion
Henry	0	1	0	1	1	0	0	0	0
Xu	1	0	1	1	1	1	0	0	0
Lisa	0	1	0	0	1	1	0	0	0
Tang	1	1	0	0	1	0	1	1	0
Bella	1	1	1	1	0	1	1	1	1
Kim	0	1	1	0	1	0	0	1	1
Reza	0	0	0	1	1	0	0	1	0
Max	0	0	0	1	1	1	1	0	1
Zion	0	0	0	0	1	1	0	1	0

I = 0.12
E(I) = -0.125

Alternative hypothesis = True
P-value = 0.03

Conclusion

- Our data is spatially autocorrelated.
- We still don’t know what to do about it…

Case Study: economic impact of green spaces in Zillow neighborhoods

Dataset

Zillow median neighborhood home price

Socio-demographics and home characteristics from the American Community Survey

- Median household income
- Number of rooms
- Etc.

Environmental attributes

- Land surface temperature
- Tree cover
- Etc.

The full study includes many variables

Variable definition, Unit	Min	Max	Mean	Std. dev.
ZHVI; median price per ft^2 (dollars)	12.1	1957.9	232.8	217.9
structural variables				
median number of rooms	2.1	9.0	6.3	1.0
median age of home (yrs)	5.0	78.0	50.1	19.0
demographic variables				
median age of residents (yrs)	16.3	77.1	38.7	7.1
population density (people/m^2)	0.0002	0.01	0.001	0.001
proportion of white residents (%)	0	1.0	0.7	0.2
proportion obtained bachelor’s degree (%)	0	0.62	0.24	0.11
proportion obtained master’s degree (%)	0	0.49	0.11	0.07
median household income (dollars)	10,940	250,000	73,000	34,500
community features				
categorical: majority road type	secondary road = 4149, tertiary road = 2110			
slope (degrees)	1.2	18.2	3.6	1.9
proportion impervious surfaces (%)	0	0.94	0.42	0.16
binary: 1 = college or university present			0.17	
binary: 1 = k-12 school present			0.78	
binary: 1 = highway present			0.51	
categorical: mode aspect	NE = 4690, NW = 265, SW = 1292			
categorical: mode development intensity	medium = 2346, high = 412, low = 3501			
categorical: U.S. state				
environmental attributes				
binary: 1 = golf course present			0.06	
binary: 1 = cemetery present			0.20	
binary: 1 = park present			0.74	
proportion park area (%)	0	0.55	0.03	0.05
binary: 1 = lake/pond present			0.22	
binary: 1 = stream/river present			0.10	
binary: 1 = swamp/marsh present			0.03	
land surface temperature, Celsius	0.20	44.7	27.4	6.7
tree canopy cover (%)	0	0.58	0.12	0.09
NDVI (-1 – 1)	0	0.47	0.25	0.08
proportion open space (%)	0	0.50	0.15	0.12

Task

Model the median neighborhood home price as a function of socio-demographics, home characteristics, and environmental attributes.

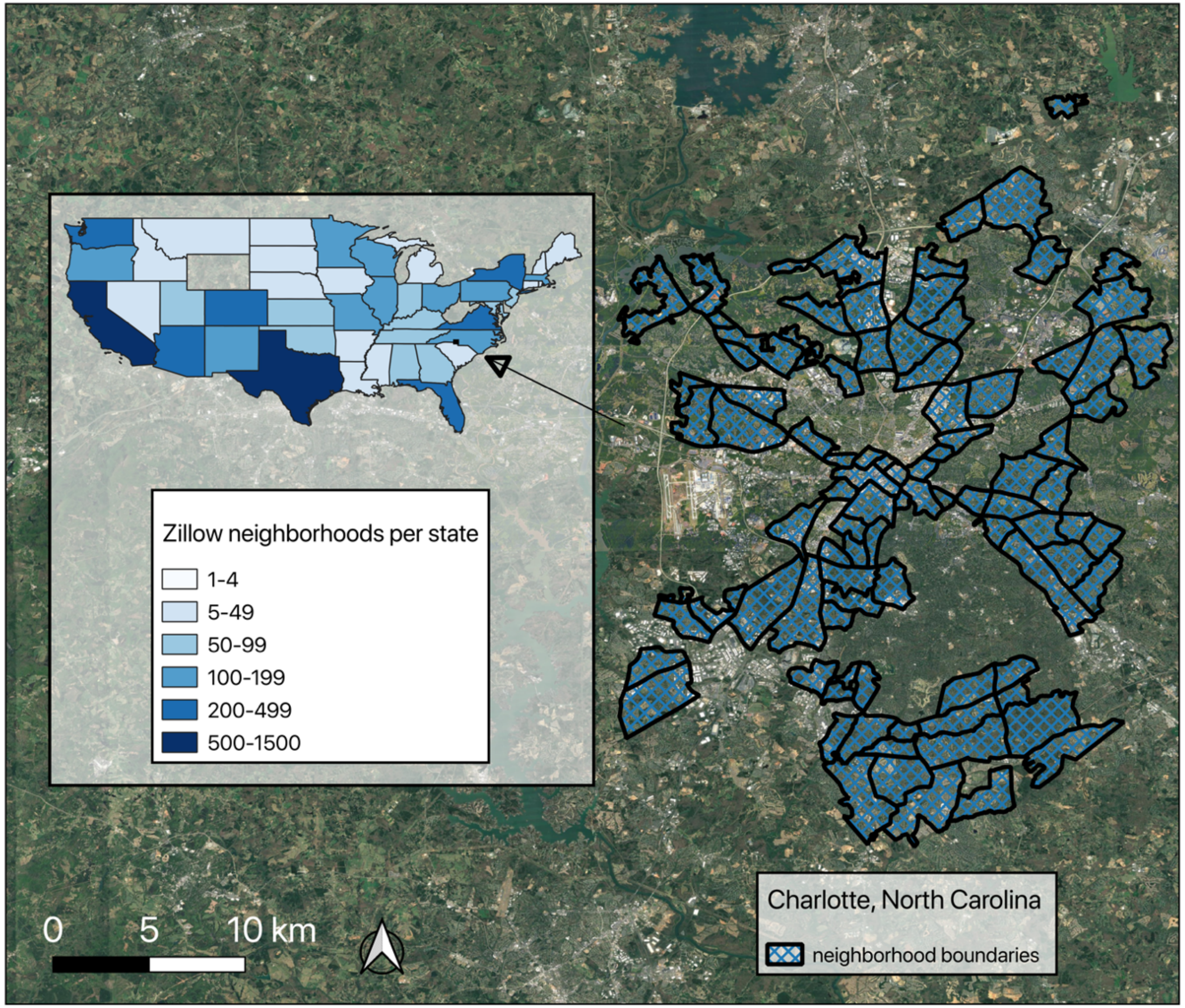
This is called a hedonic pricing analysis.

The least squares model looks like this:

$$price \sim \beta_0 + (\beta_1 * income) + (\beta_2 * age\ of\ home) + \cdots + (\underset{\uparrow}{\beta_3} * tree\ cover) + \varepsilon$$

This is what we are interested in

Zillow neighborhoods are spatially distributed, so we need to consider spatial autocorrelation.



What does Mr. Moran say?

“Reject the null hypothesis!”

Building a spatial autoregressive model

Original model (ordinary least squares)

$$P_i = \beta_0 + \beta_1 \underset{\overset{\uparrow}{\text{Structural features}}}{S_i} + \beta_2 \underset{\overset{\uparrow}{\text{Demographics}}}{D_i} + \beta_3 \underset{\overset{\uparrow}{\text{Community features}}}{A_i} + \beta_4 \underset{\overset{\uparrow}{\text{Environmental attributes}}}{E_i} + \varepsilon_i$$

Building a spatial autoregressive model

Spatial lag model

λ is an Estimated parameter (just like β)

$$P_i \sim \beta_0 + \lambda W P_i + \beta_1 S_i + \beta_2 D_i + \beta_3 A_i + \beta_4 E_i + \varepsilon_i$$

Structural features

demographics

Community features

Environmental attributes

W

	Walltown	Trinity Heights	Forest Hills
Walltown	0	1	0
Trinity Heights	1	0	0
Forest Hills	0	0	0

Question: 3 minutes

Link: https://bit.ly/38AAVnj

Building a spatial autoregressive model

Spatial error model

$$P_i = \beta_0 + \beta_1 S_i + \beta_2 D_i + \beta_3 A_i + \beta_4 E_i + \boxed{\rho W \mu_i} + \varepsilon_i$$

Structural features

demographics

Community features

Environmental attributes

W

	Walltown	Trinity Heights	Forest Hills
Walltown	0	1	0
Trinity Heights	1	0	0
Forest Hills	0	0	0

Building a spatial autoregressive model

Spatial lag AND error model

$$P_i = \beta_0 + \boxed{\lambda W P_i} + \underset{\substack{\uparrow \\ \text{Structural features}}}{\beta_1 S_i} + \underset{\substack{\uparrow \\ \text{demographics}}}{\beta_2 D_i} + \underset{\substack{\uparrow \\ \text{Community features}}}{\beta_3 A_i} + \underset{\substack{\uparrow \\ \text{Environmental attributes}}}{\beta_4 E_i} + \boxed{\rho W \mu_i} + \varepsilon_i$$

W

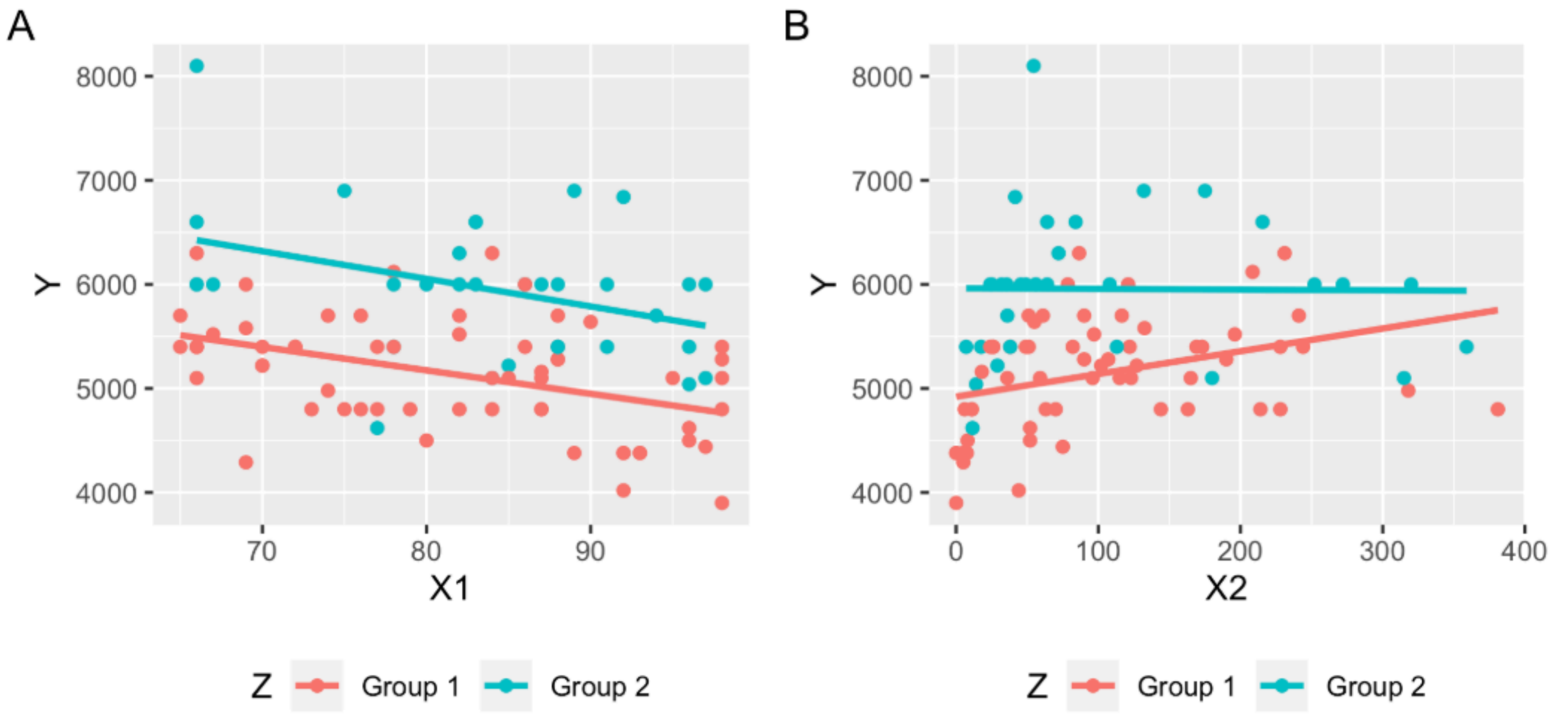
	Walltown	**Trinity Heights**	**Forest Hills**
Walltown	0	1	0
Trinity Heights	1	0	0
Forest Hills	0	0	0

Actual model estimates for the spatial lag + spatial error model

Interaction term!

Variable	coeff.
spatial lag for price (λ)	0.03***
spatial error (ρ)	0.72***
intercept	0.45**
environmental attribute variables	
park = 1	0.05***
park = 1 * park area	0.005**
stream/river = 1	-0.02**
ln(land surface temperature)	0.23***
(ln(land surface temperature))^2	-0.04***
ln(percent tree canopy cover)	0.05***
ln(NDVI)	-0.17***
ln(open space)	-0.007***
R^2	0.90
log-likelihood	275
AIC	-392

Recall interactions from Monday's lecture



Include interaction terms in the Zillow model

“main effects” “interaction effects”

$$P_i = \beta_0 + \beta_1(temperature) + \beta_2(tree\ cover) + \beta_3(temperature * tree\ cover)$$

$$= (\beta_3 * temperature)(tree\ cover)$$

$$= (\beta_3 * tree\ cover)(temperature)$$

How to interpret interaction coefficient β_3?

- β_3 positive
  - “As temperature increases, the effect of tree cover on price becomes more positive”
  - And vice versa
- β_3 negative
  - “As temperature increases, the effect of tree cover on price becomes more negative”
  - And vice versa

$$\beta_1(temperature) + \beta_2(tree\ cover) + \beta_3(open\ space) + \beta_4(temperature * tree\ cover) + \beta_5(temperature * open\ space) + \beta_6(tree\ cover * open\ space)$$

	interaction effects		**main effects**
	Tree canopy cover	open space	
temperature	0.18***	-0.06***	0.24***
tree canopy cover		-0.002	-0.56***
open space			0.17***

Question: 4 minutes. https://bit.ly/2Hy7VRd

$$\beta_1(temperature) + \beta_2(tree\ cover) + \beta_3(NDVI) + \beta_4(open\ space) + \beta_5(temperature * income) + \beta_6(tree\ cover * income) + \beta_7(NDVI * income) + \beta_8(open\ space * income)$$

	interaction effects with income	main effects
temperature	-0.15***	1.78***
tree canopy cover	0.04***	-0.39***
NDVI	0.03	-0.46
open space	0.01	-0.07

Question: 4 minutes. https://bit.ly/325OfOd