

Geospatial Analysis

C. Alex Simpkins Jr., Ph.D
UC San Diego, RDPRobotics LLC



Department of Cognitive Science
rdrobotics@gmail.com
csimpkinsjr@ucsd.edu

Lectures :http://casimpkinsjr.radiantdolphinpress.com/pages/cogs108_ss1_23/index.html

Course Reminders

- Complete D4 by Monday at midnight
- Complete Project checkpoint 1 by Sunday at midnight
 - Project proposal
 - Checkpoint 1: Data

Why Geospatial Analysis?

How a coastline 100 million years ago influences modern election results in Alabama

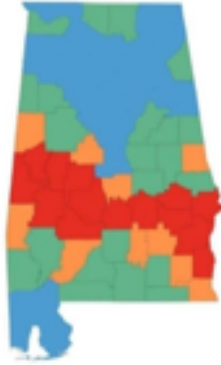
Cretaceous Sediments



Fertile Blackland Prairie Soil



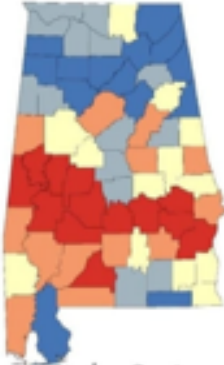
Average Farm Size, 1997



Slave Population, 1860

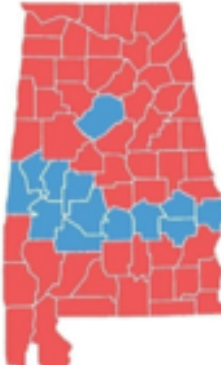


Black population, 2010



-Starkey Comics

Election Results, 2020

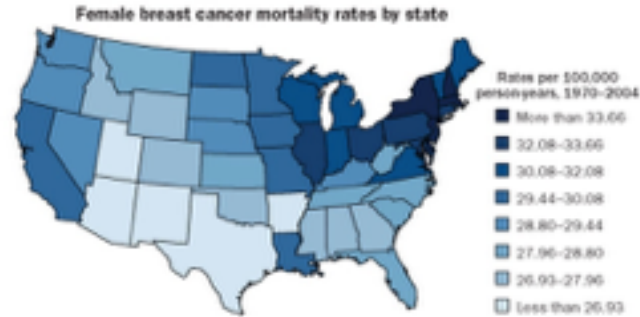
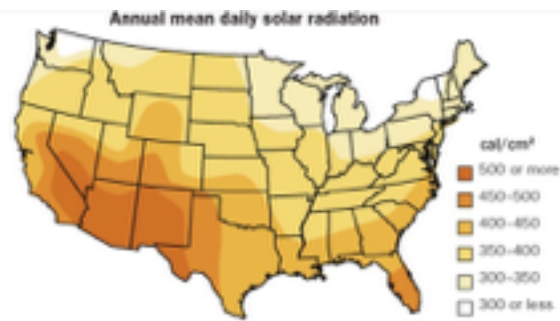


Why Geospatial Analysis?

“Everything is related to everything else, but near things are more related than distant things.” -Tobler 1979

“...the purpose of geographic inquiry is to examine relationships between geographic features collectively and to use the relationships to describe the real-world phenomena that map features represent” - Clarke 2001

Clearly visualizes
important differences in
disease distribution

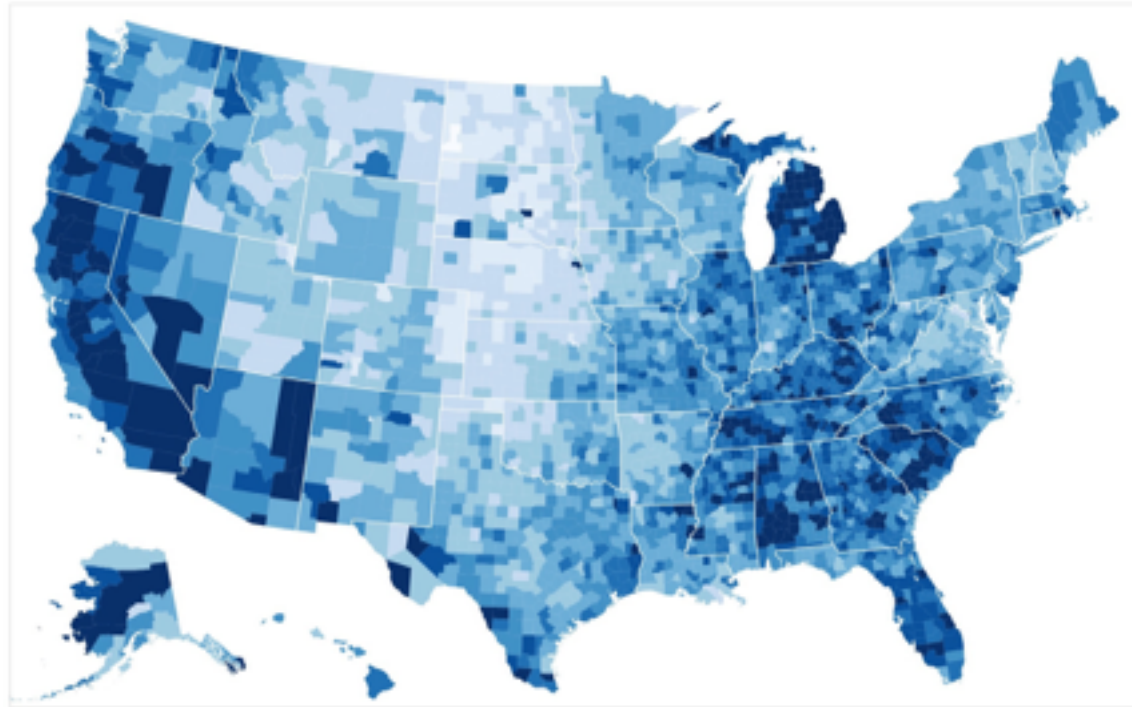


ON THE MAP Scientists who study vitamin D can't help but notice that a host of diseases seem to vary with latitude. Type 1 diabetes, multiple sclerosis and even some cancers appear to be more common in areas that get less sun -- meaning less opportunity for the body to produce vitamin D. The maps above illustrate the apparent link between solar radiation and breast cancer mortality rates.

SOURCE, FROM TOP: D. M. HARRIS AND V.L.W. GO // J. OF NUTRITION 2004; NATIONAL CANCER INSTITUTE

Visualizing Geospatial Data

Unemployment
rate by county
(August 2016)



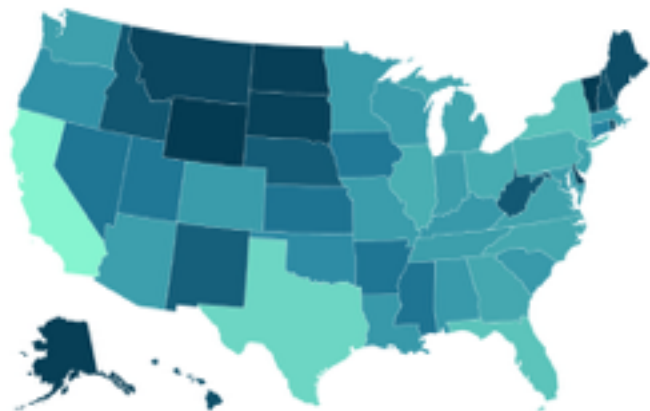
This choropleth encodes unemployment rates from 2008 with a [quantize scale](#) ranging from 0 to 15%. A [threshold scale](#) is a useful alternative for coloring arbitrary ranges.

[Open in a new window.](#)

Choropleth maps are useful for visualizing *clear regional patterns* in the data

Use light colors for low values. Dark colors for high values.

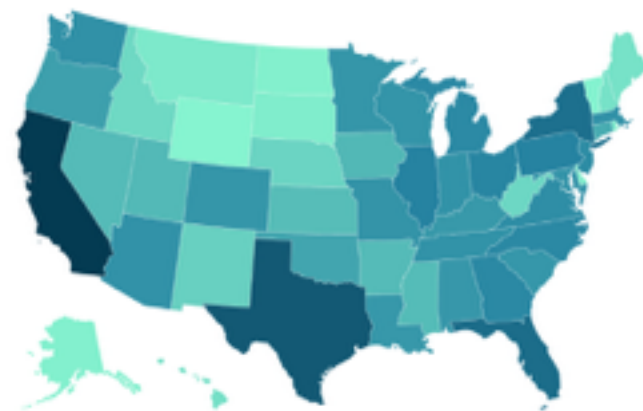
NOT IDEAL



LOW POPULATION HIGH



BETTER



LOW POPULATION HIGH

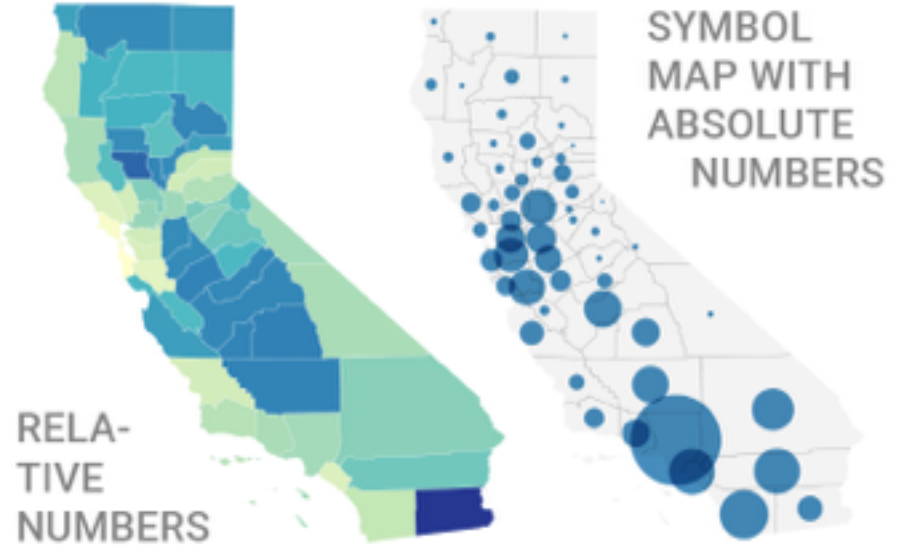


Choropleth should display relative differences, *not* absolute numbers

NOT IDEAL

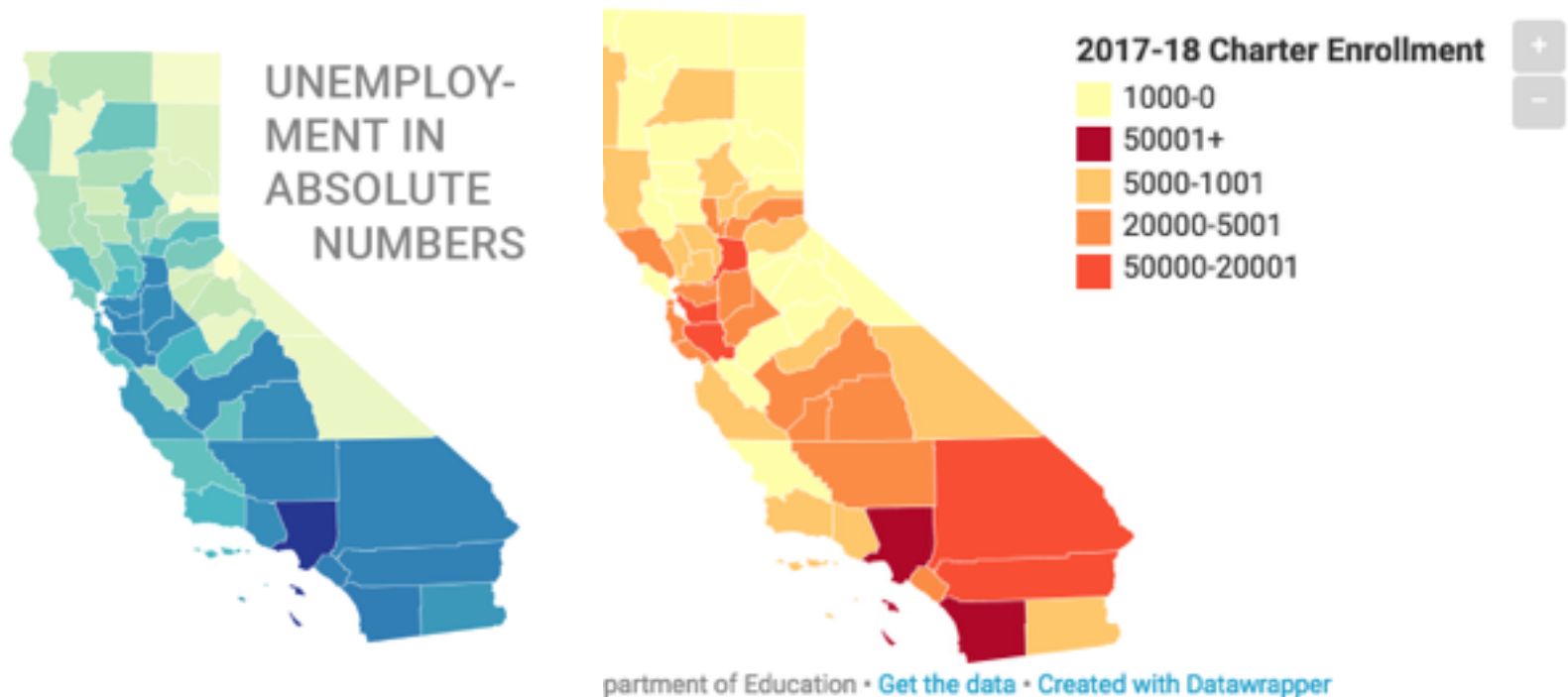


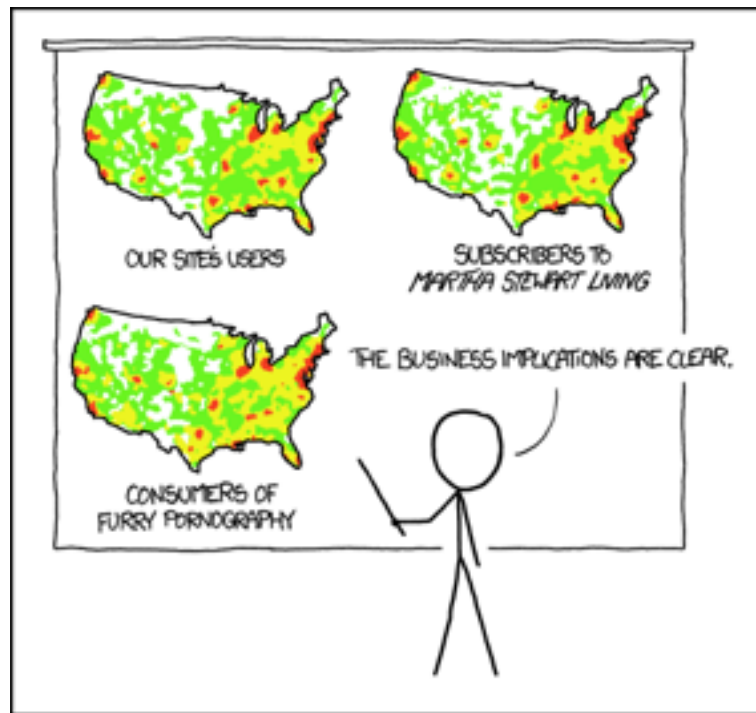
BETTER



Map: Where Are Students Attending Charter Schools?

The majority of California's charter school student population is concentrated in Los Angeles, San Diego and Bay Area counties. Hover through the counties on each map for more information on their



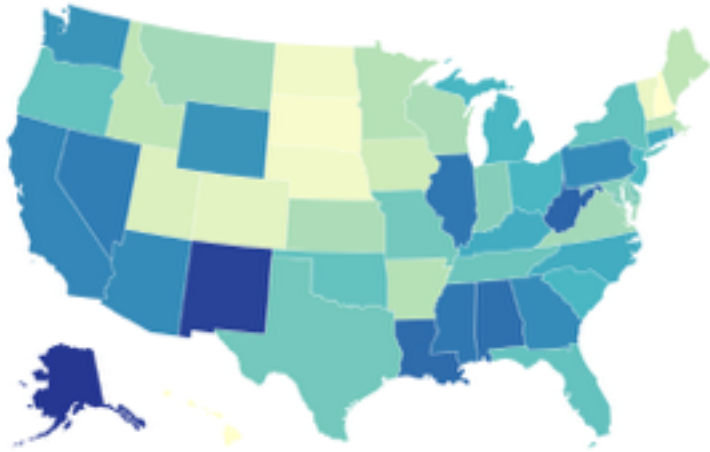


PET PEEVE #208:
GEOGRAPHIC PROFILE MAPS WHICH ARE
BASICALLY JUST POPULATION MAPS

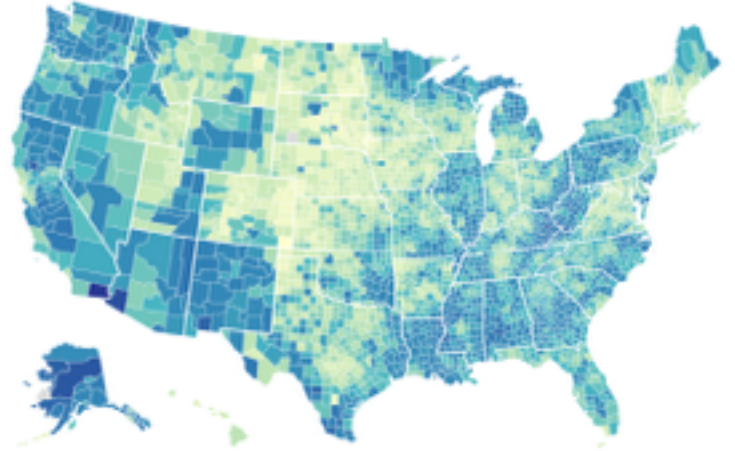
Choropleth maps can be misleading

Consider using the smallest unit possible
(but there are exceptions!)

NOT IDEAL

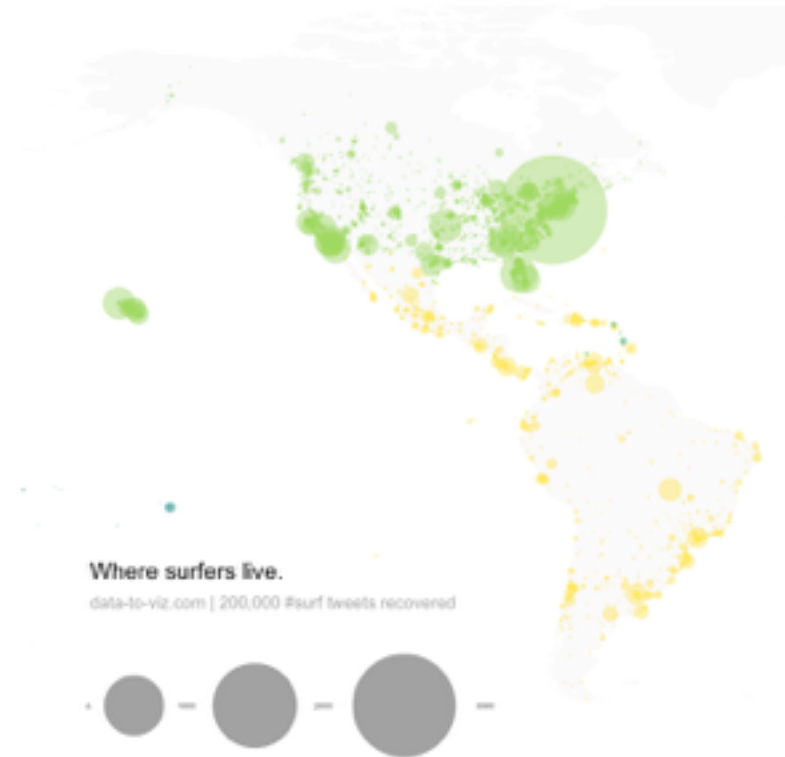


BETTER



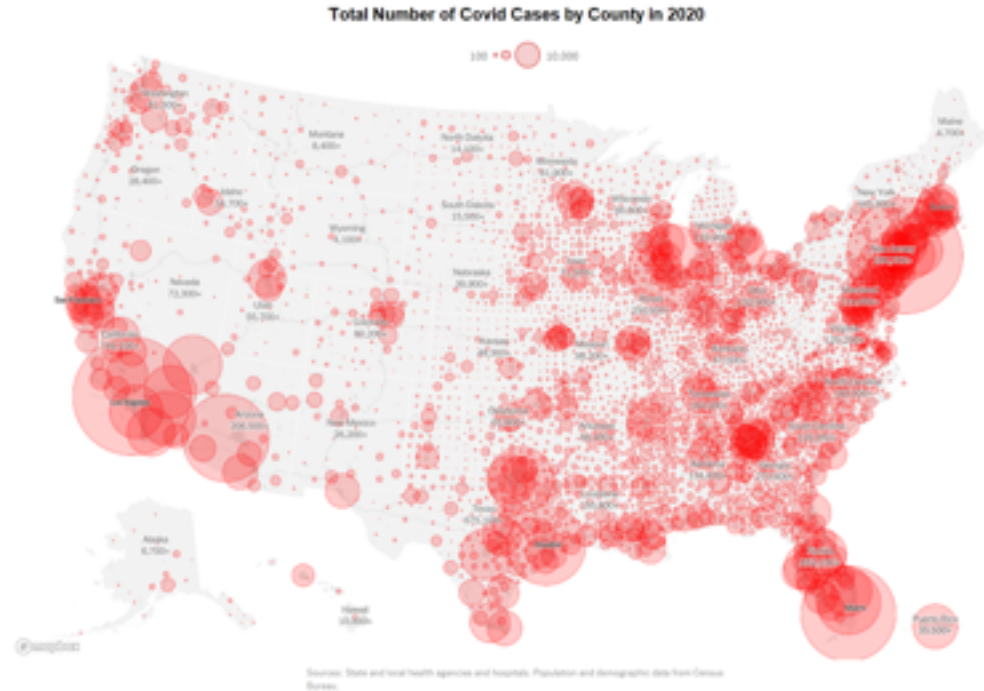
Bubble graphs can represent a third (or fourth!) dimension, giving a sense of relative contribution

- Useful for comparing data in 3 numeric-data dimensions
- A 4th dimension can be represented by color
- XY-data is like a scatter plot with the third axis a circle whose size is determined by the dimension
- Helps with relative comparisons
- Cannot be used to display a lot of data, difficult to express actual values



Bubble maps

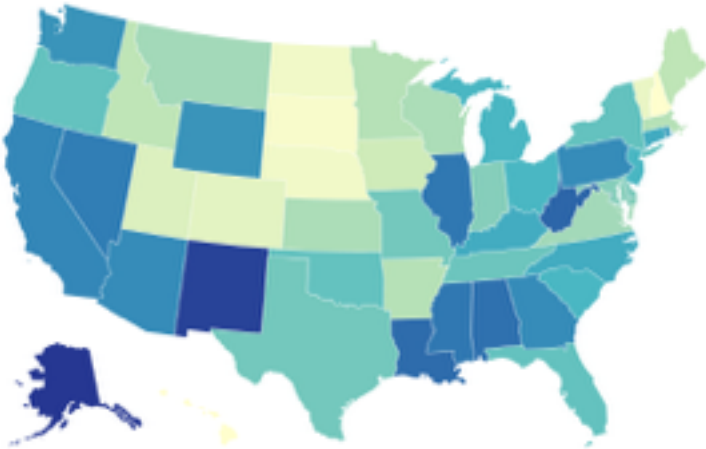
- Coordinates of latitude and longitude
- Bubble size is third axis, such as population density, COVID cases, etc
- Notes:
 - *Consider using area rather than radius to avoid exaggerating bubble sizes*
 - *Transparency for bubbles*
 - *Legend!*



Visualization Choices

Cartograms should be considered when displaying how many people were affected

NOT IDEAL



Choropleths answer “How much area was affected?”

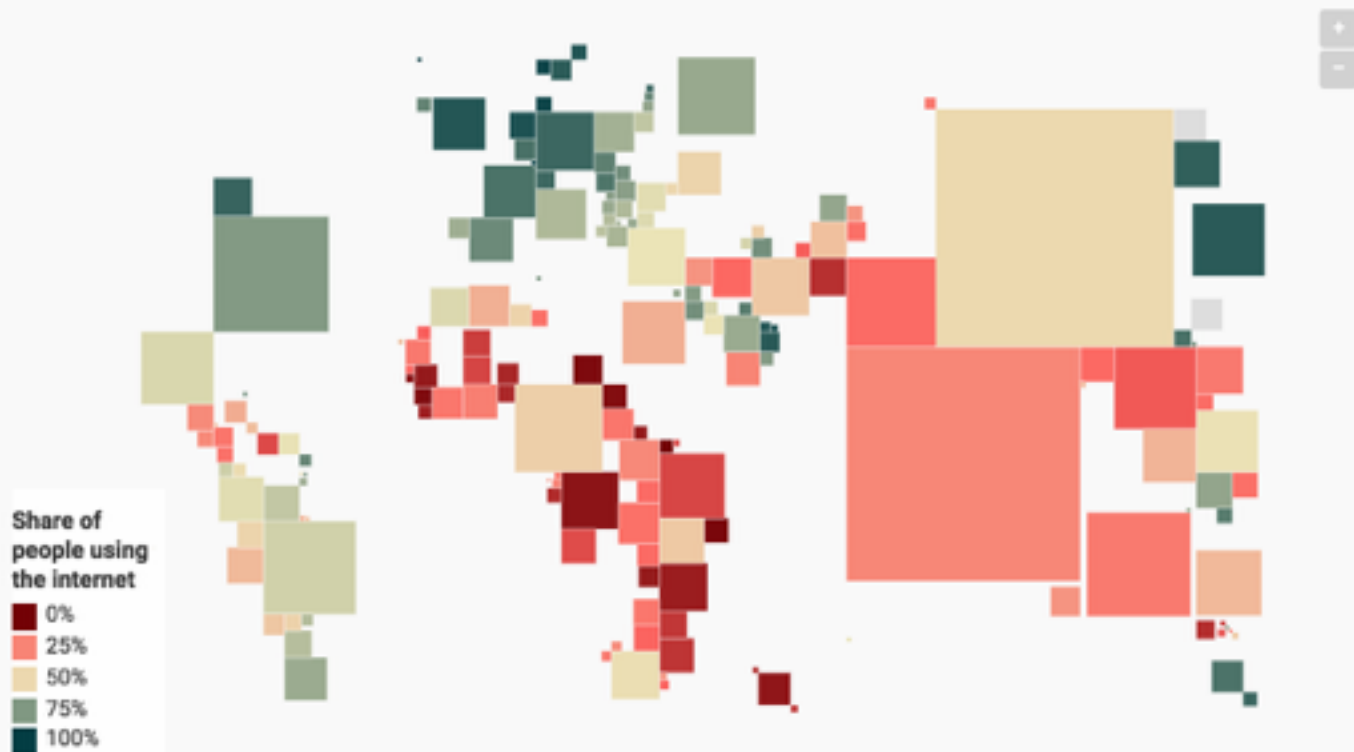
BETTER



Cartograms answer “How many people were affected?”

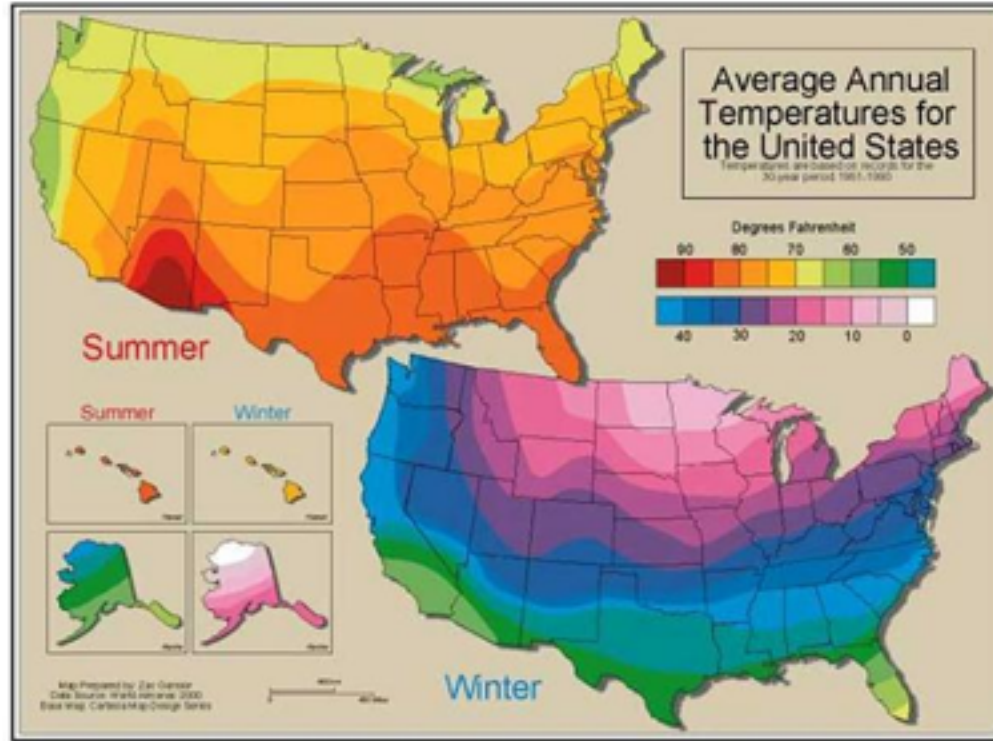
Share of individuals using the internet, 2015

Share of individuals using the internet, measured as the percentage of the population. Internet users are individuals who have used the Internet (from any location) in the last 3 months. The Internet can be used via a computer, mobile phone, personal digital assistant, games machine, digital TV etc.



Source: [Our World in Data](#) · [Get the data](#)

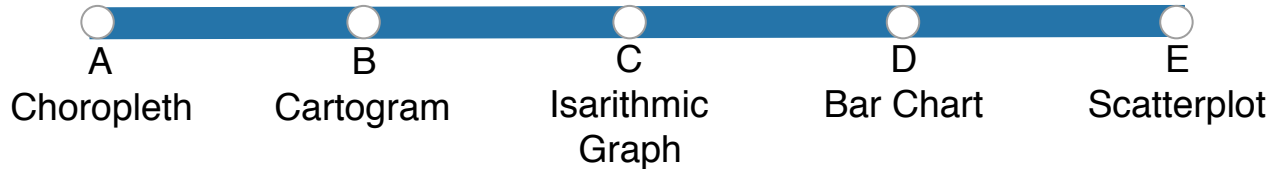
Isarithmic maps demonstrate smooth, continuous phenomena (temperature, elevation, rainfall, etc.)





You want to visualize how many people have been affected by COVID19 worldwide.

Best approach to visualize these data?



Spatial Statistics : The Why

Spatial Statistics

The statistical techniques we've discussed so far don't work well when considering spatial distributions...

Spatial Statistics

The statistical techniques we've discussed so far don't work well when considering spatial distributions...

...which means we have a chance to take a look at data and the relationship between the data in new and interesting ways (distance, adjacency, interaction, and neighbor)

Spatial data violate conventional statistics:

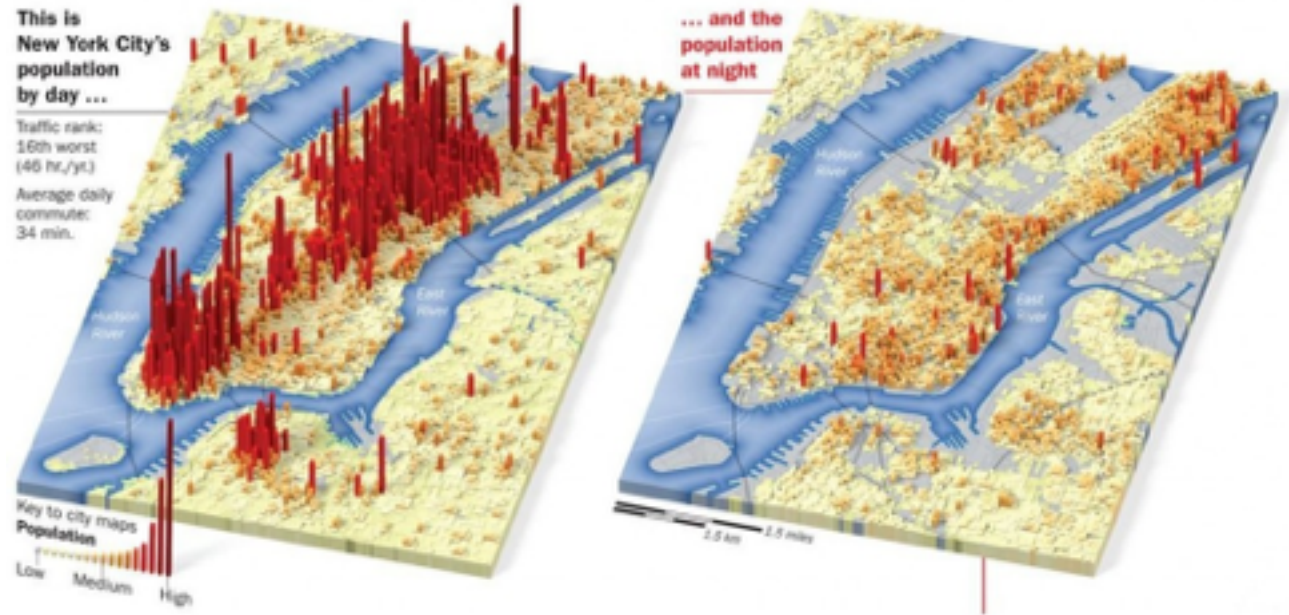
Violations of conventional statistics:

- Spatial autocorrelation
- Modifiable areal unit problem (MAUP)
- Edge effects (Boundary problem)
- Ecology fallacy
- Nonuniformity of space

Spatial Autocorrelation

Data from locations near one another in space are more likely to be similar than data from locations remote from one another:

- Housing market
- Elevation change
- Temperature



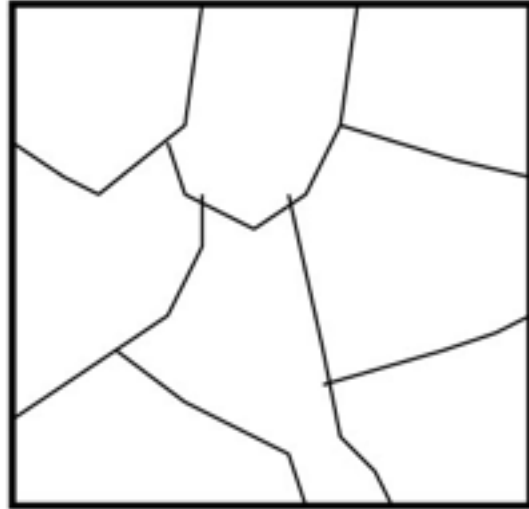
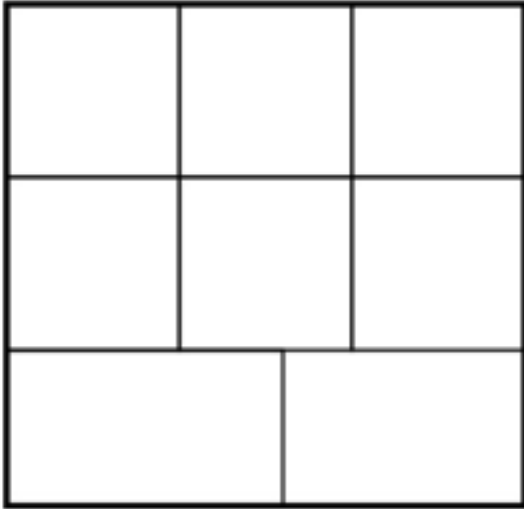
Modifiable Areal Unit Problem (MAUP)

The aggregation units used are arbitrary with respect to the phenomena under investigation, yet the aggregation units used will affect statistics determined on the basis of data reported in this way.

If the spatial units in a particular study were specified differently, we might observe very different patterns and relationships.

Modifiable Areal Unit Problem (MAUP)

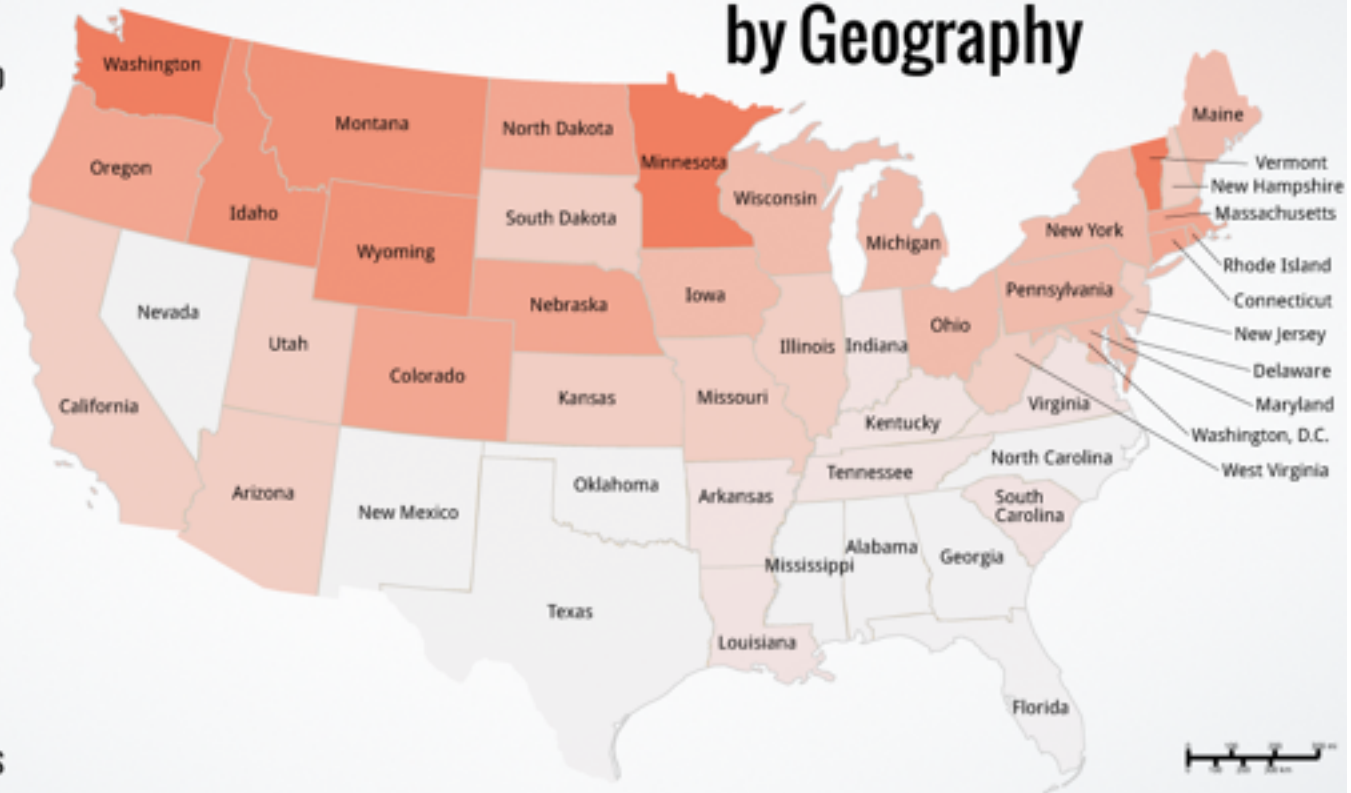
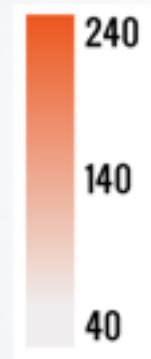
modifiable area: Units are arbitrarily defined and different organization of the units may create different analytical results.



Multiple Sclerosis by Geography

CASE-CONTROL RATIO OF MS

A higher ratio indicates greater prevalence



LEARN MORE AT
WWW.INVW.ORG/MS

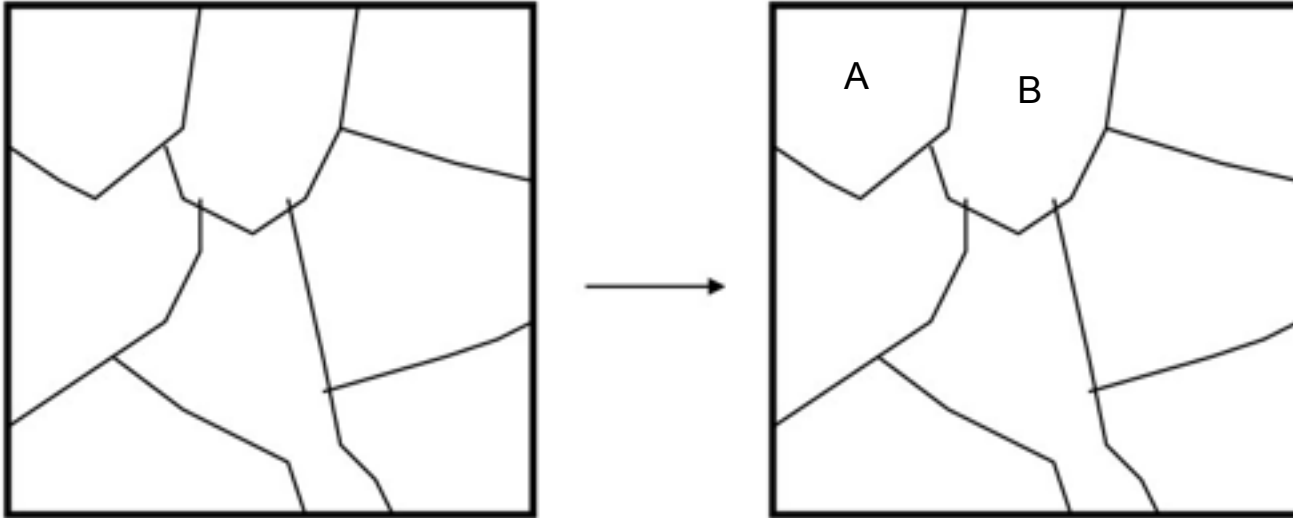
PRODUCED BY: JASON ALCOORN/INVESTIGATEWEST SOURCE: BERETICH AND BERETICH (2009)

Modifiable Areal Unit Problem (MAUP)

- Potential problems in almost every field that utilizes spatial data.
- One of the most stubborn problems in spatial analysis when spatially aggregated data is used

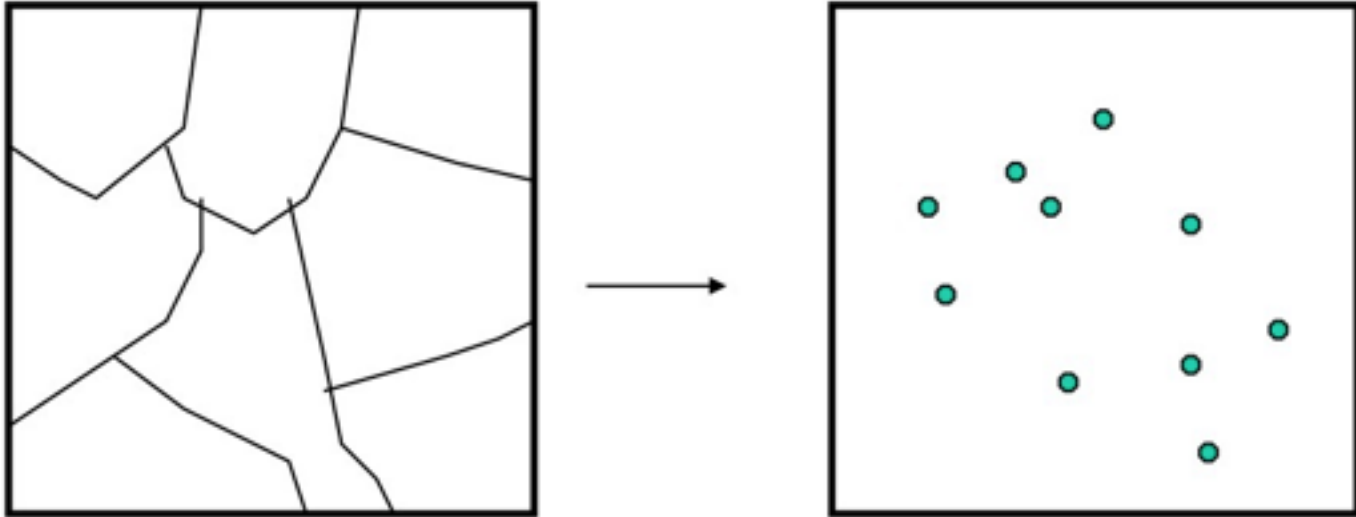
Edge Effects (The Boundary Problem)

Analyzing A vs B ignores similarities between the two based on their shared boundary



Ecological Fallacy

The Ecological Fallacy is a situation that can occur when a researcher or analyst makes an inference about an individual based on aggregate data for a group.



Ecological Fallacy

Example: we might observe a *strong relationship between income and crime at the county level*, with lower-income areas being associated with higher crime rate.

Conclusion:

- Lower-income persons are more likely to commit crime
- Lower-income areas are associated with higher crime rates
- Lower-income counties tend to experience higher crime rates

Ecological Fallacy

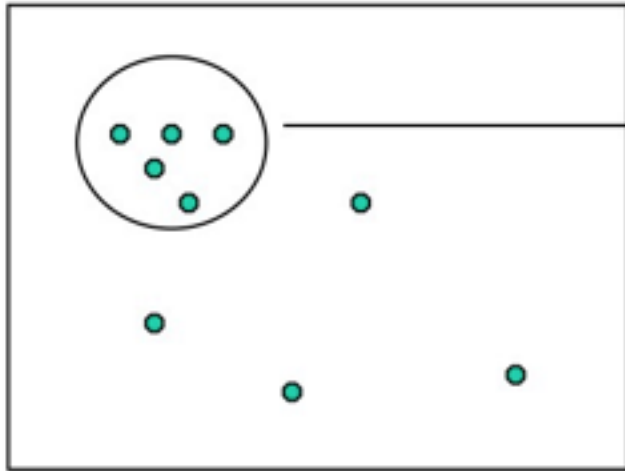
Issues:

Inferences drawn about associations between the characteristics of an aggregate population and the characteristics of sub-units within the population are wrong. That is: *results from aggregated data (e.g. counties) cannot be applied to individual people*

What should we do?

Be aware of the process of aggregating or disaggregating data may conceal the variations that are not visible at the larger aggregate level

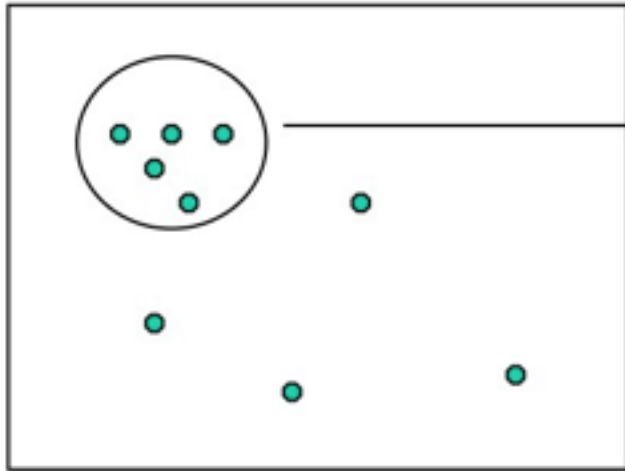
Nonuniformity



Area with high crime rates?

Crime locations

Nonuniformity

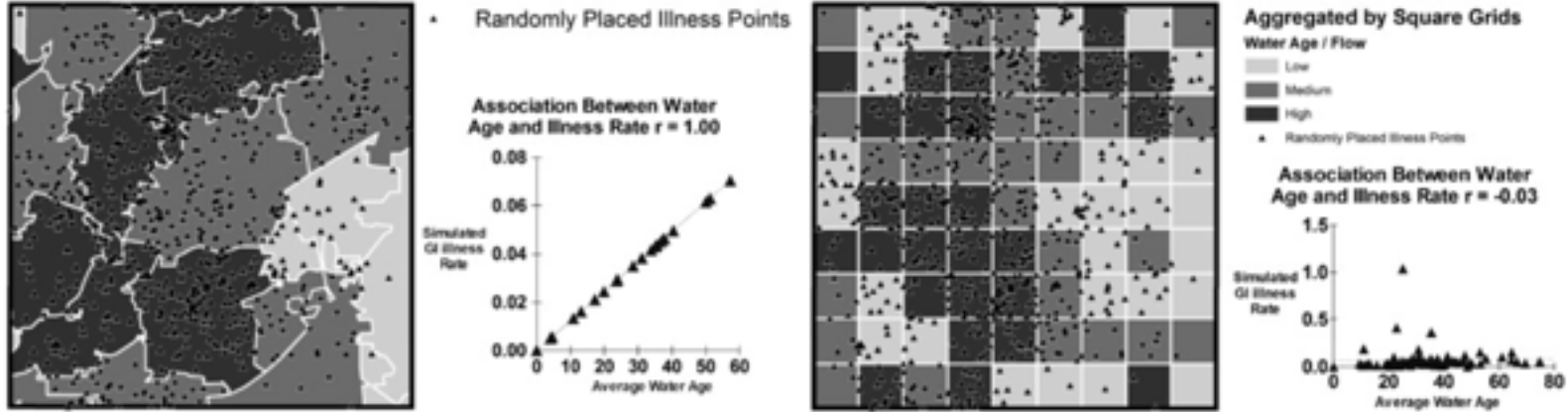


Area with high crime rates?

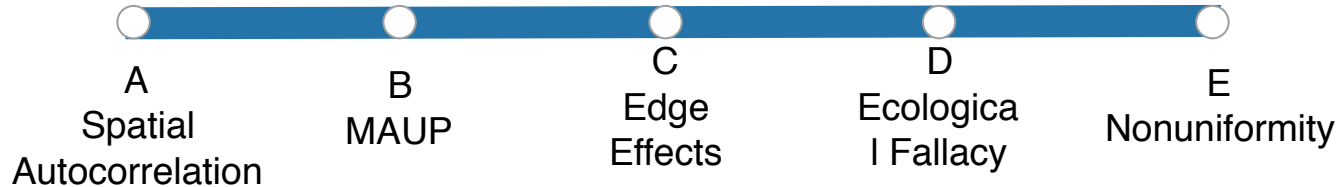
Conclusion: Bank robberies are clustered
....but only because banks are clustered!

Crime locations

Spatial Statistics



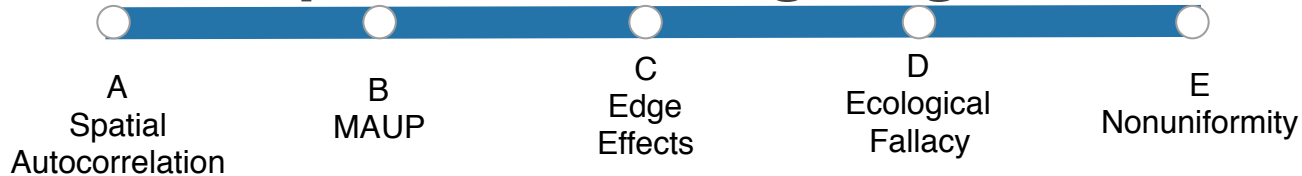
What explains what's going on here?





In Baltimore City, police spend more time in a few neighborhoods. Crime rates are higher in those neighborhoods.

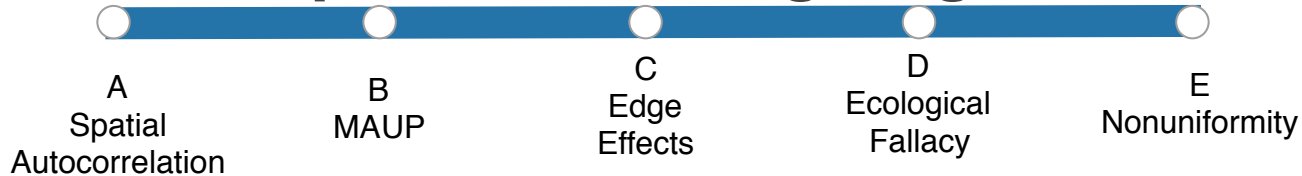
What explains what's going on here?





A Trader Joe's just opened in a new neighborhood. Nearby homes are now worth more money.

What explains what's going on here?



Spatial Statistics : The Basics

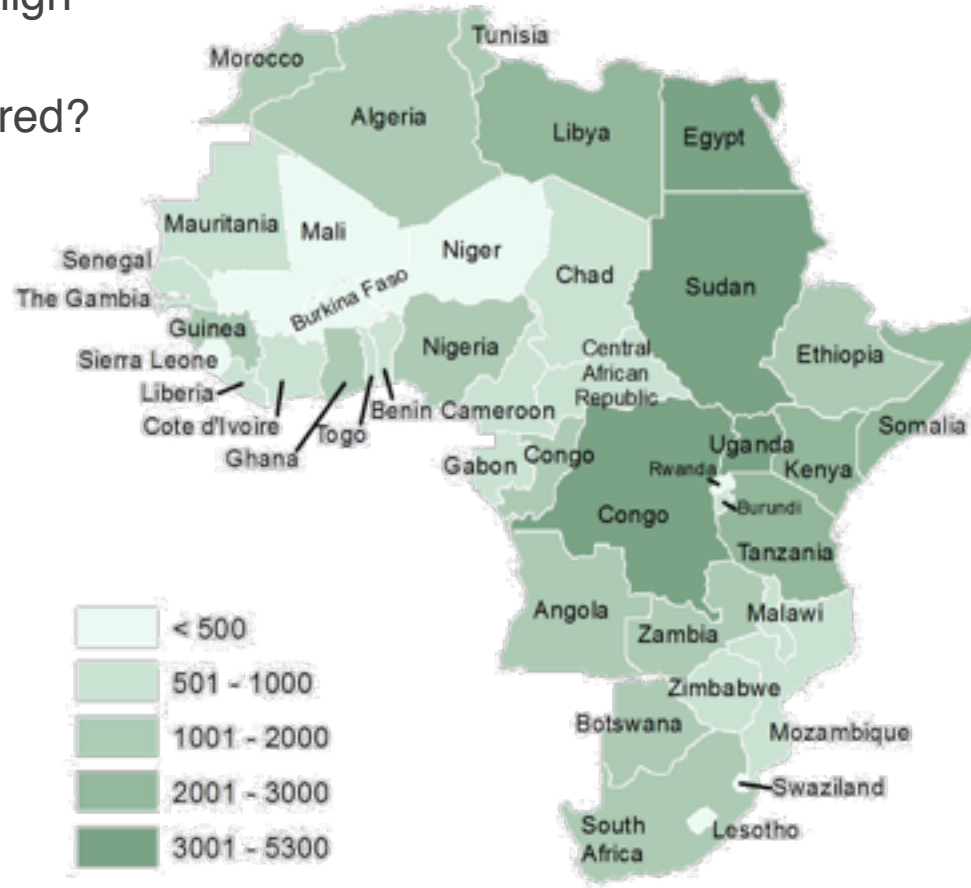
Table 1.1: Index of total African conflict for the 1966-78 period (Anselin and O'Loughlin 1992).

Country	Conflicts	Country	Conflicts
EGYPT	5246	LIBERIA	980
SUDAN	4751	SENEGAL	933
UGANDA	3134	CHAD	895
ZAIRE	3087	TOGO	848
TANZANIA	2881	GABON	824
LIBYA	2355	MAURITANIA	811
KENYA	2273	ZIMBABWE	795
SOMALIA	2122	MOZAMBIQUE	792
ETHIOPIA	1878	IVORY COAST	758
SOUTH AFRICA	1875	MALAWI	629
MOROCCO	1861	CENTRAL AFRICAN REPUBLIC	618
ZAMBIA	1554	CAMEROON	604

Data source: Anselin, L. and John O'Loughlin. 1992. *Geography of international conflict and cooperation: spatial dependence and regional context in Africa*. In *The New Geopolitics*, ed. M. Ward, pp. 39-75.

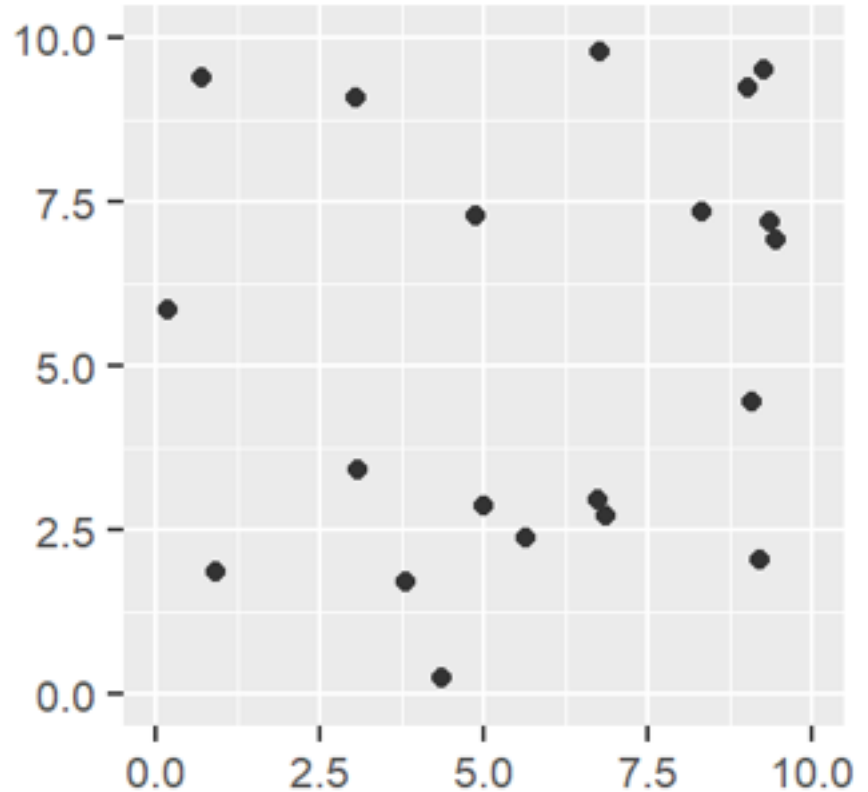
Are countries with a high conflict index score geographically clustered?

Are countries with a high conflict index score geographically clustered?



Global Point Density

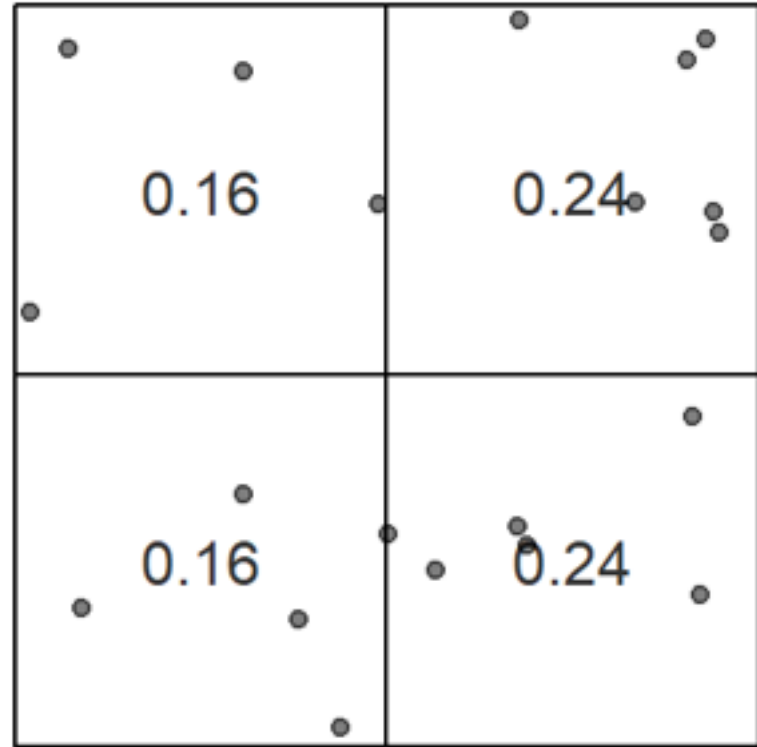
the ratio of observed
number of points to the
study region's surface area



Quadrat Density (local)

Surface is divided and then point density is calculated within quadrat

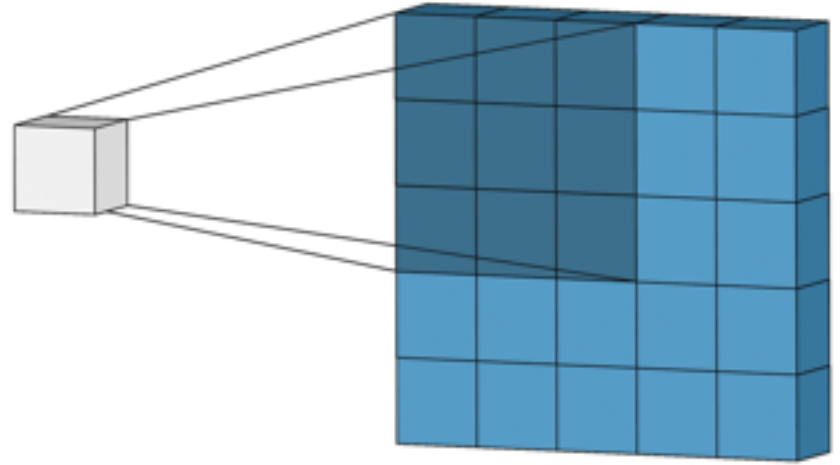
Note: quadrat number and shape will affect measurement estimate. Suffers from MAUP.



Kernel Density (local)

Point density is calculated within sliding windows (window size = kernel)

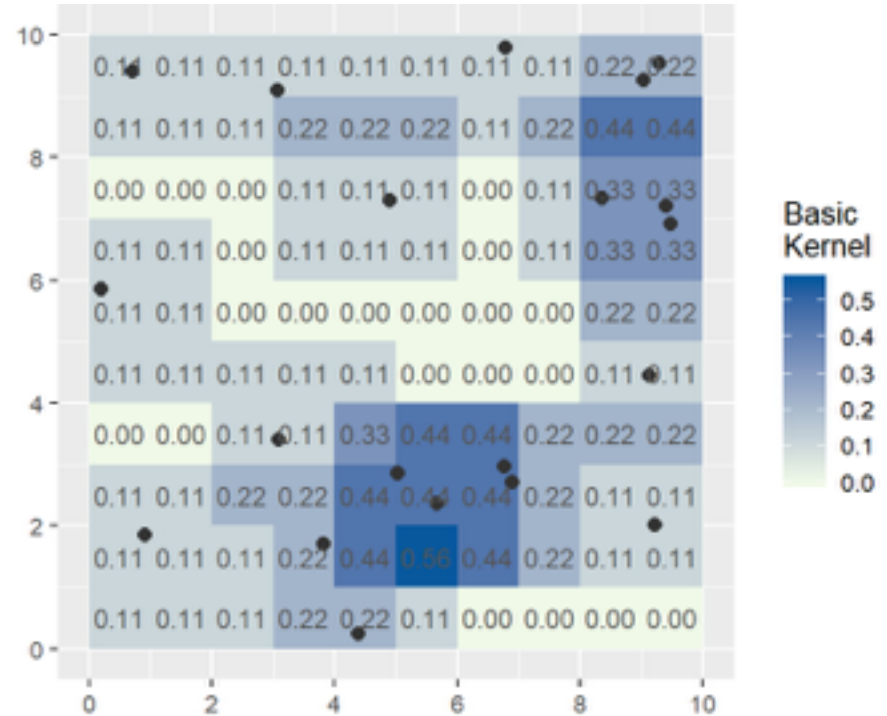
Note: kernel will affect measurement estimate, but this is less susceptible to MAUP.



Kernel Density (local)

Point density is calculated within sliding windows (window size = kernel)

Note: kernel will affect measurement estimate, but this is less susceptible to MAUP.



Modeling these data: Poisson Point Process

(Density-based Methods - - how the points are distributed relative to the study space)

$$\lambda(i) = e^{\alpha + \beta Z(i)}$$

$\lambda(i)$ is the modeled intensity at location i

e^α is the base intensity when the covariate is *zero*

e^β is the multiplier by which the intensity increases (or decreases) for each 1 unit increase in the covariate

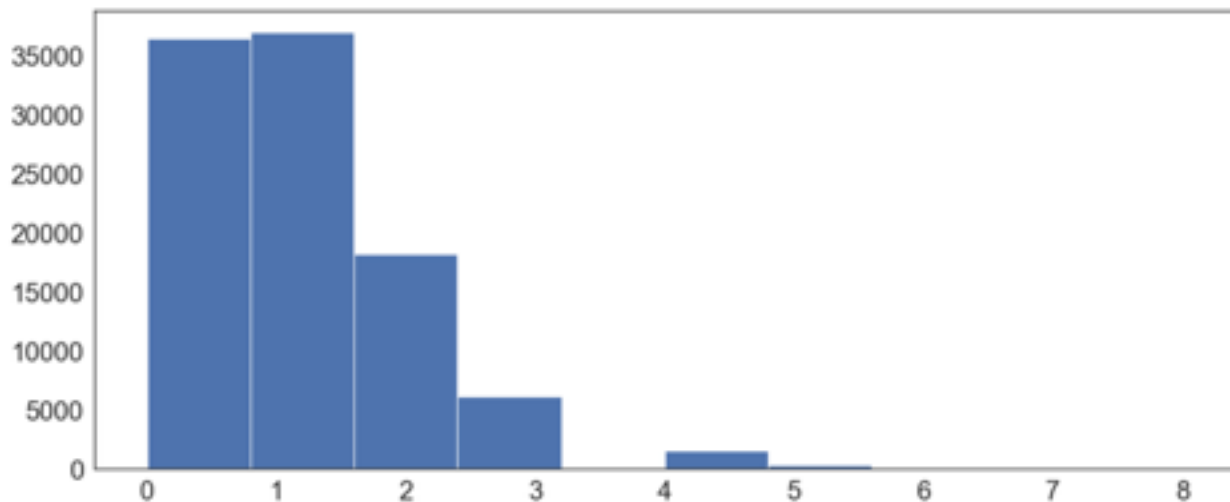
Poisson Distribution

The Poisson Distribution models events in fixed intervals of time, given a known average rate (and independent occurrences).

In [55]:

Slide Type Fragment

```
dat = poisson.rvs(mu=1, size=100000)
plt.hist(dat);
```



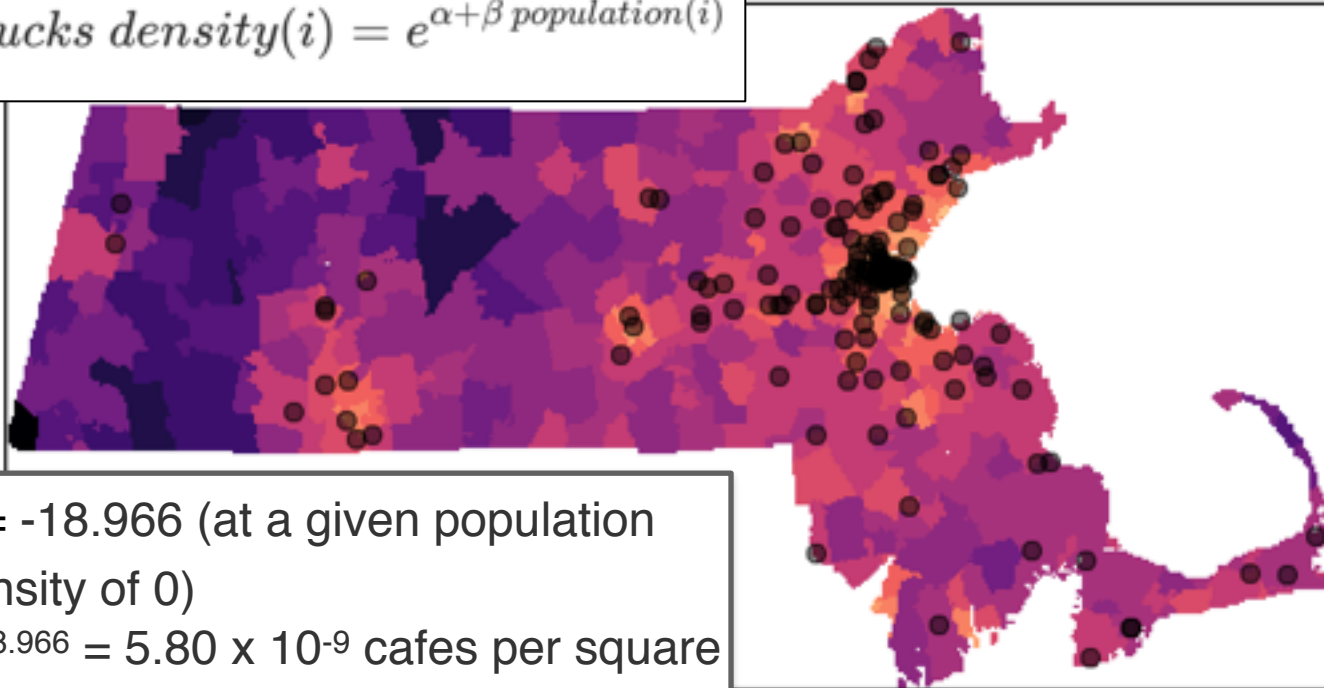
Slide Type Fragment

The number of visitors a fast food drive-through gets each minute follows a Poisson distribution. In this case, maybe the average is 3, but there's some variability around that number.

A Poisson distribution can help calculate the probability of various events related to customers going through the drive-through at a restaurant. It will predict lulls (0 customers) and flurry of activity (5+ customers), allowing staff to plan and schedule more precisely.

Location of Starbucks relative to population density in MA

$$\text{Starbucks density}(i) = e^{\alpha + \beta \text{ population}(i)}$$



$\alpha = -18.966$ (at a given population density of 0)

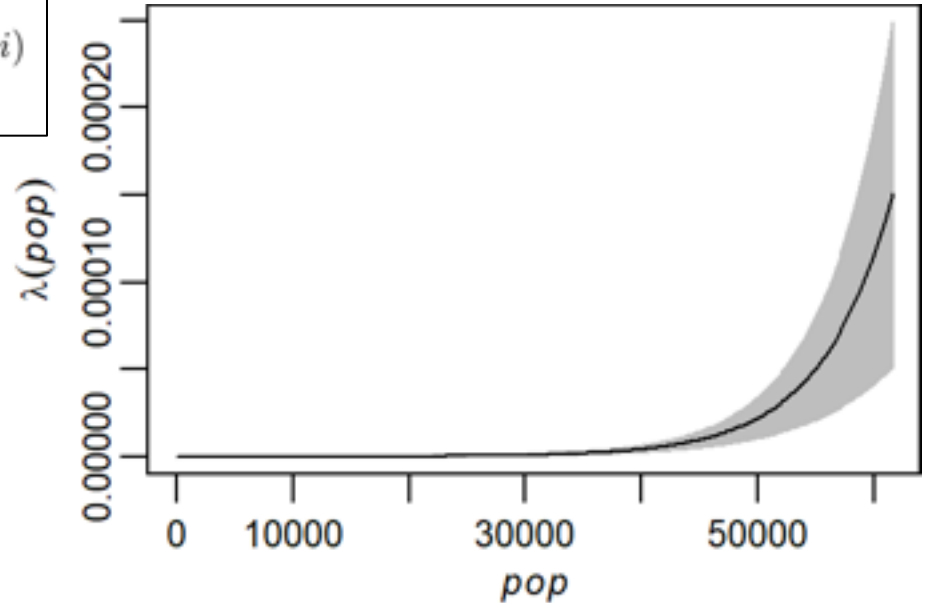
$e^{-18.966} = 5.80 \times 10^{-9}$ cafes per square meter

$\beta = 0.00017$; $e^{0.00017}$ or 1.00017

Location of Starbucks relative to population density in MA

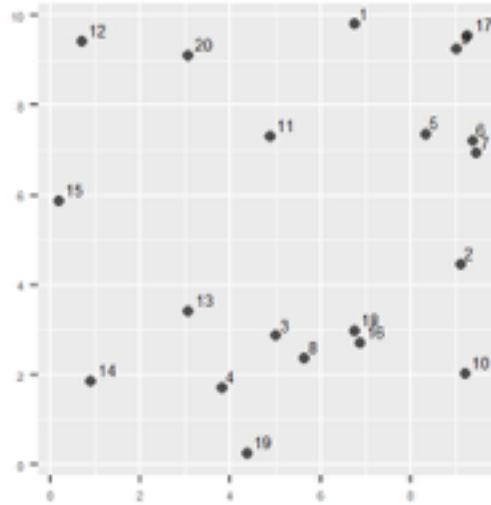
$$\text{Starbucks density}(i) = e^{\alpha + \beta \text{ population}(i)}$$

$\alpha = -18.966$ (at a given population density of 0)
 $e^{-18.966} = 5.80 \times 10^{-09}$ cafes per square meter
 $\beta = 0.00017$; $e^{0.00017}$ or 1.00017



Modeling these data: Average Nearest Neighbor

(Distance-based Methods - how the points are distributed relative to one another)



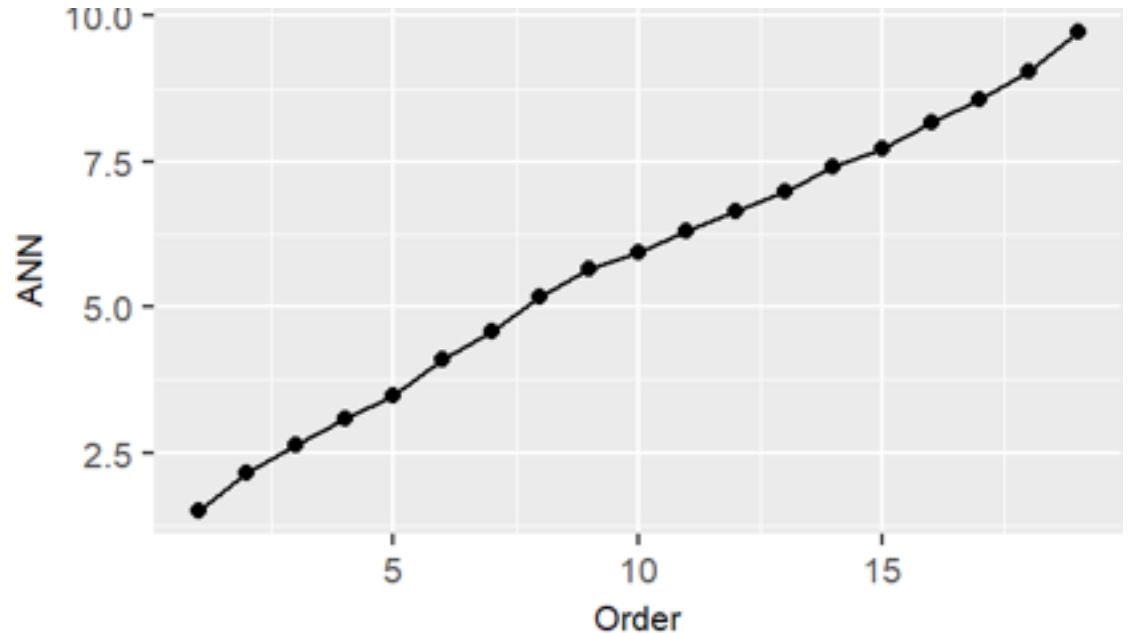
From	To	Distance	Froms	To	Distance
1	9	2.32	11	20	2.55
2	10	2.43	12	20	2.39
3	8	0.81	13	4	1.85
4	19	1.56	14	13	2.67
5	6	1.05	15	12	3.58
6	7	0.3	16	18	0.29
7	6	0.3	17	9	0.37
8	3	0.81	18	16	0.29
9	17	0.37	19	4	1.56
10	2	2.43	20	12	2.39

ANN = 1.52 units

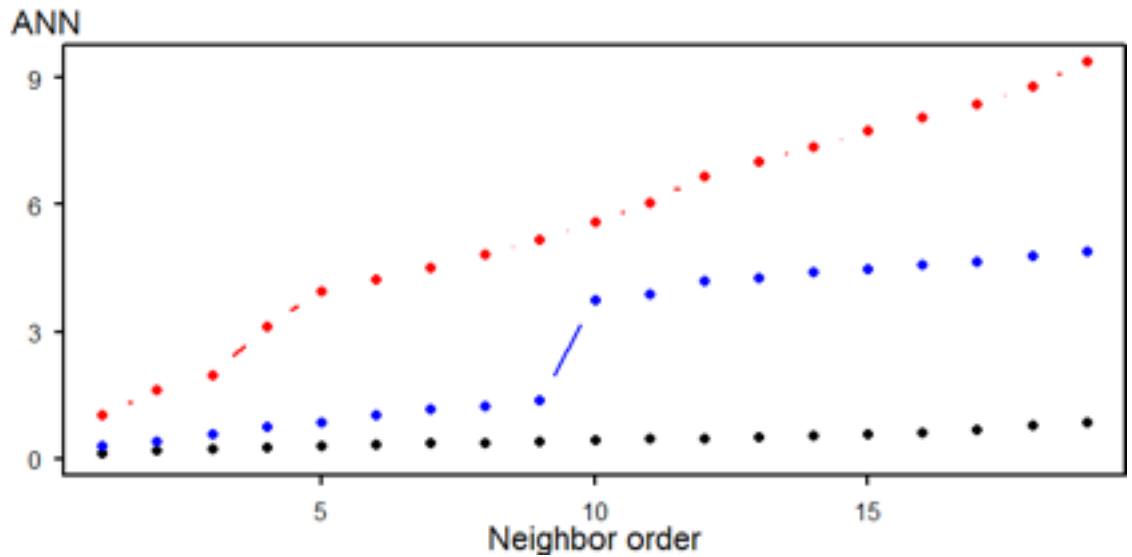
Modeling these data: Average Nearest Neighbor

(Distance-based Methods - how the points are distributed relative to one another)

plot the ANN values for different order neighbors, that is for the first closest point, then the second closest point, and so forth.



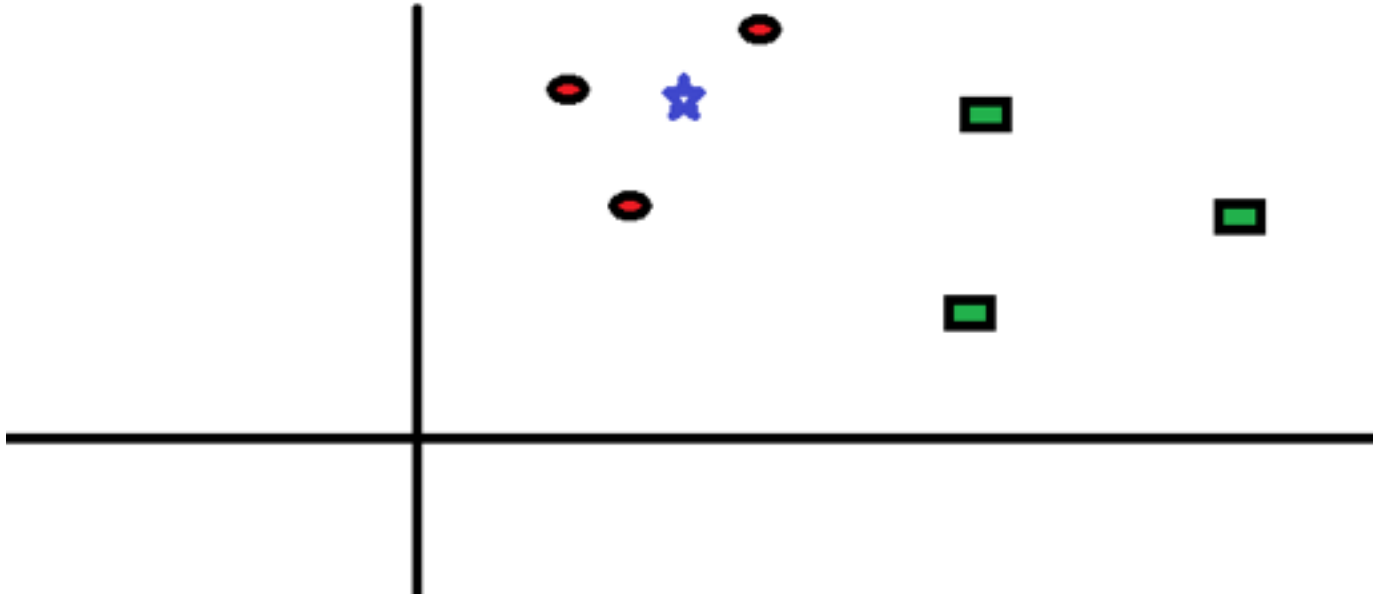
ANN vs neighbor order
offers insight into
underlying spatial
relationship



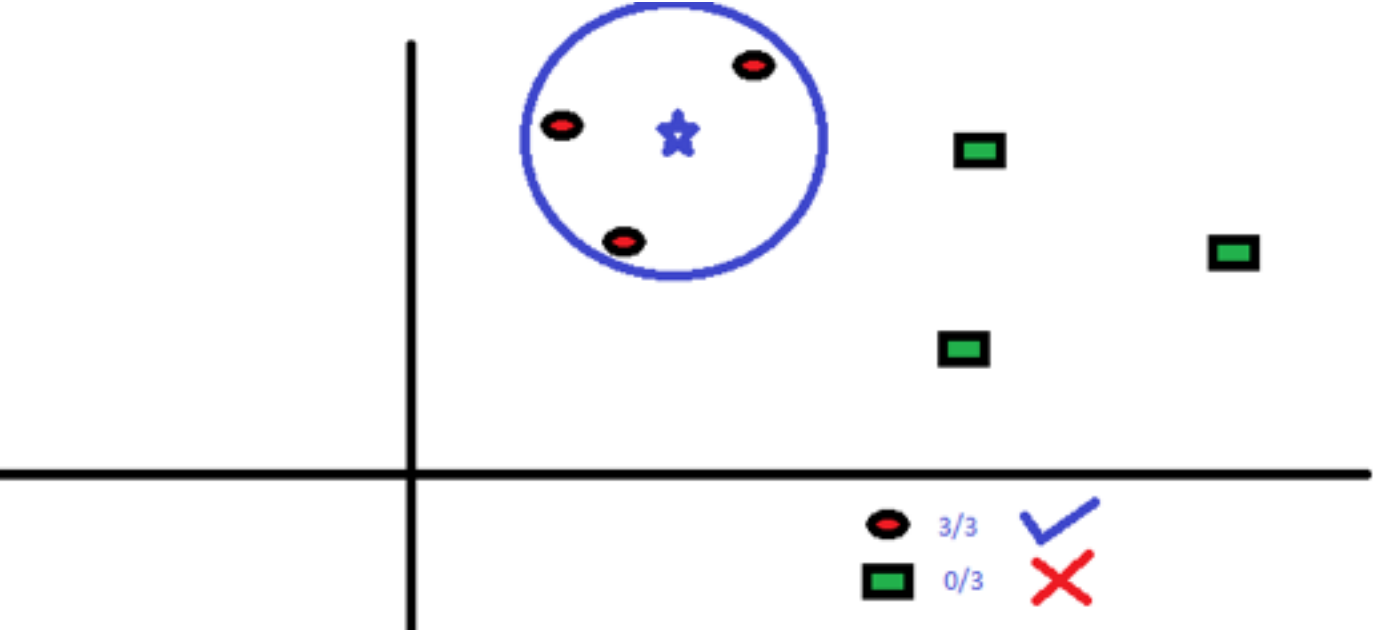
Note: study space definition
affects this measure



KNN: K Nearest Neighbor for Classification



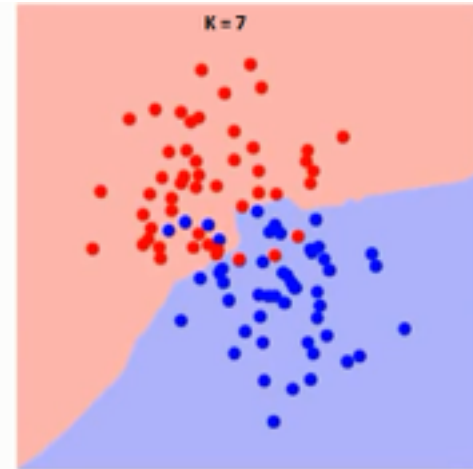
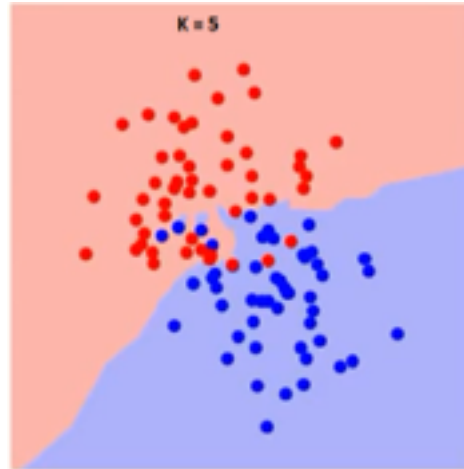
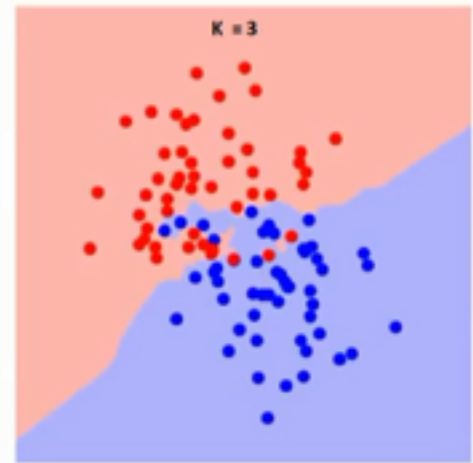
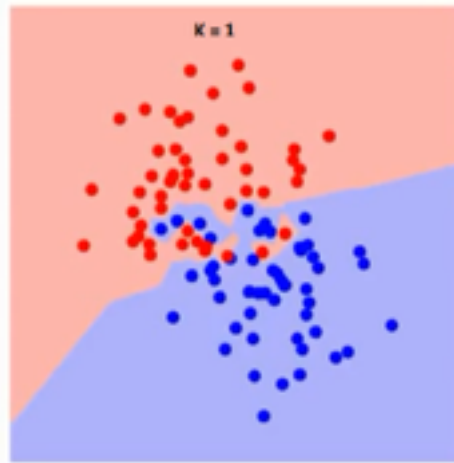
KNN: To which class does the blue star belong?



KNN: Choosing K

K specifies how many neighbors to consider.

Note that as more neighbors are considered, the boundary smooths out.



KNN: Pros & Cons

Pros:

- No assumptions about data (good for nonlinear)
- Simple and interpretable
- Relatively high accuracy
- Versatile (classification & regression)

Cons:

- Computationally intensive
- High Memory requirements
- Stores all (or most) of training data
- Prediction slow with large N
- Sensitive to outliers/irrelevant features

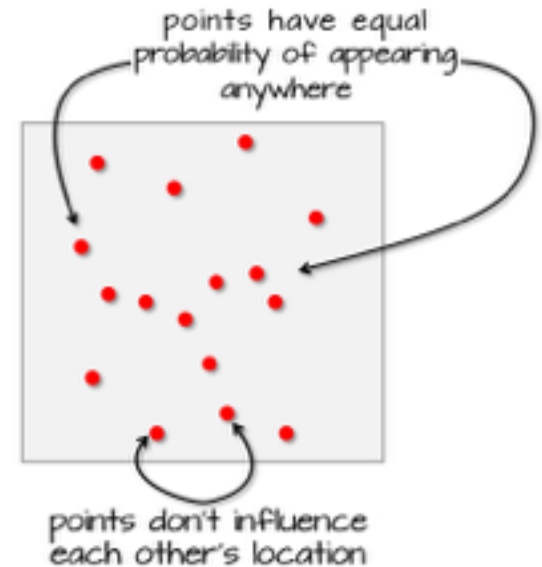
Hypothesis Testing: CSR/IPR

(Distance-based Methods - how the points are distributed relative to one another)

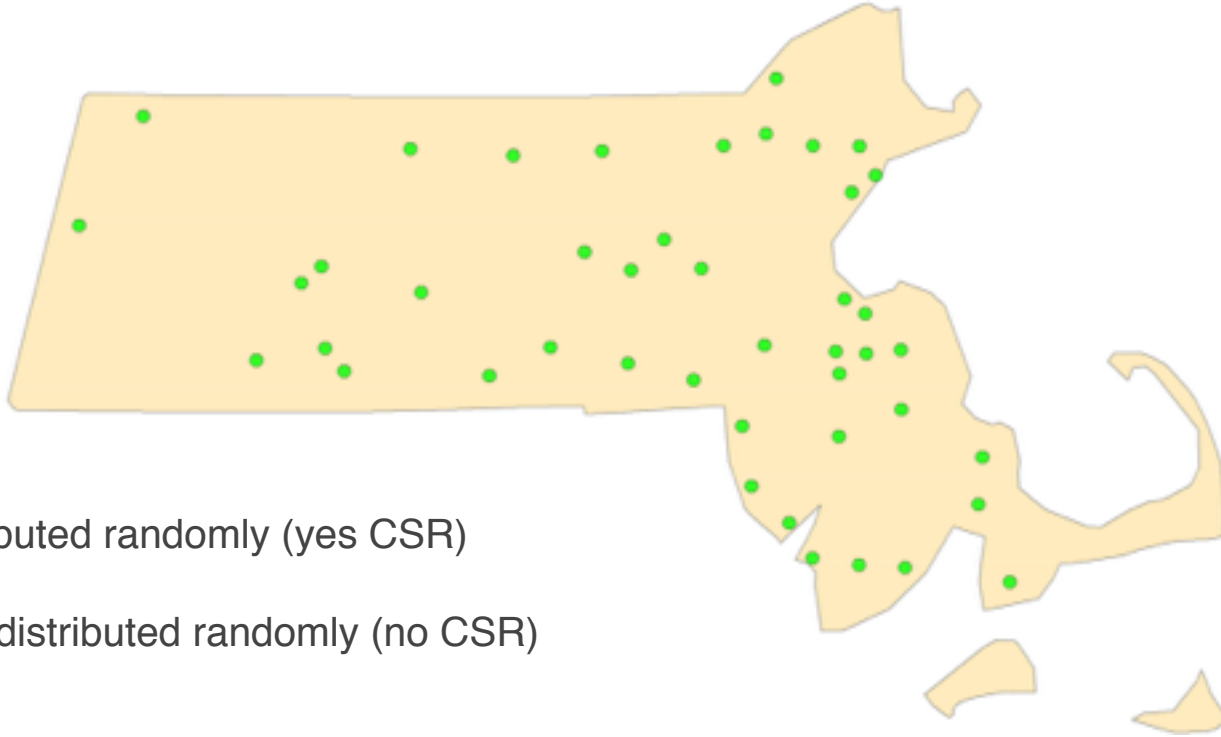
Compare observed point patterns to ones generated by an independent random process (IRP), aka complete spatial randomness (CSR).

CSR/IPR satisfy two conditions:

1. Any event has equal probability of being in any location, a 1st order effect.
2. The location of one event is independent of the location of another event, a 2nd order effect



Is this distribution of Walmarts in MA the result of CSR?



H_0 : Distributed randomly (yes CSR)

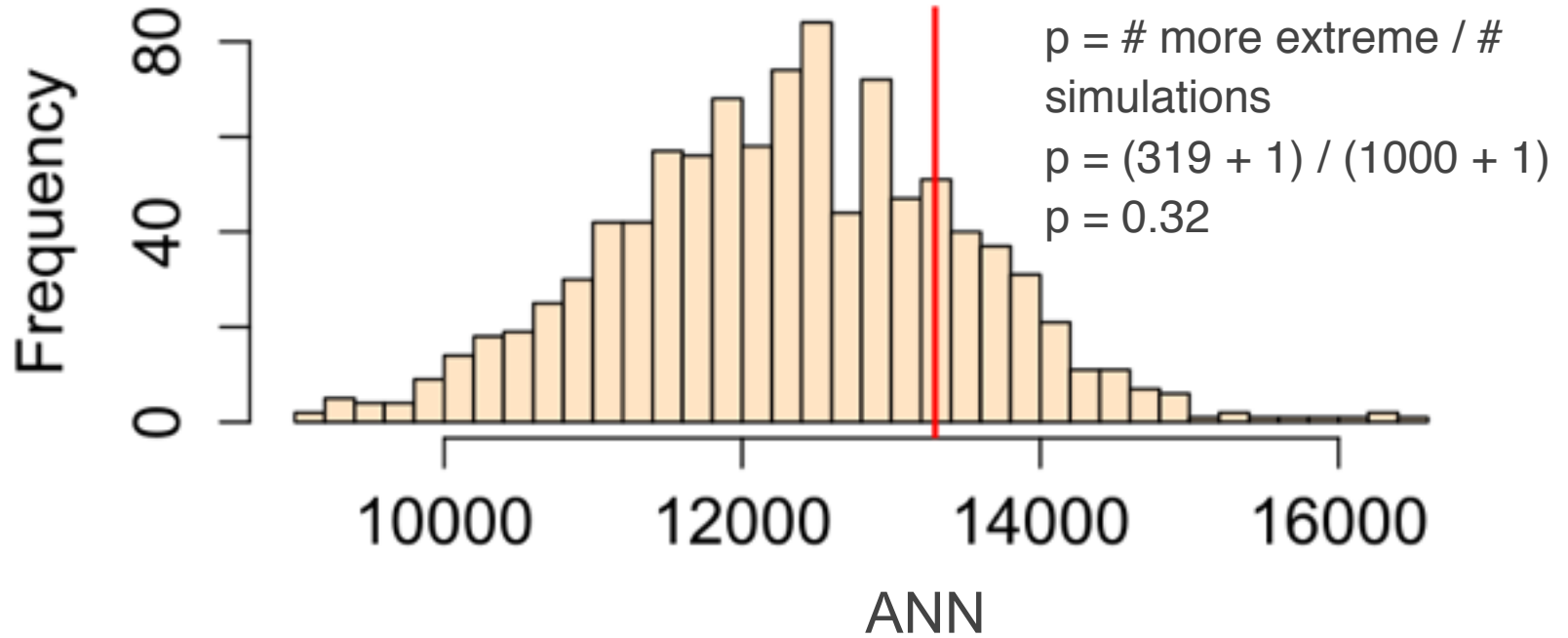
H_a : NOT distributed randomly (no CSR)

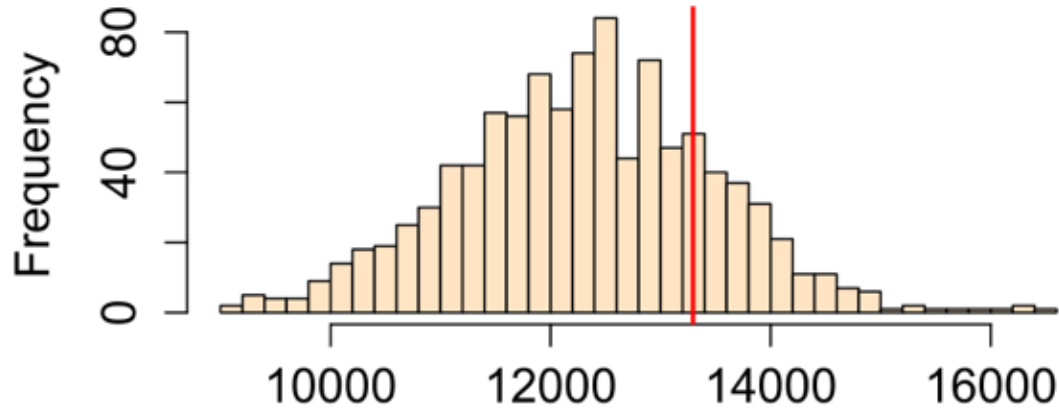
Hypothesis Testing: A Monte Carlo Test

1. First, we postulate a process—our null hypothesis, H_0 . For example, we hypothesize that the distribution of Walmart stores is consistent with a completely random process (CSR).
2. Next, we simulate many realizations of our postulated process and compute a statistic (e.g. ANN) for each realization.
3. Finally, we compare our observed data to the patterns generated by our simulated processes and assess (via a measure of probability) if our pattern is a likely realization of the hypothesized process.

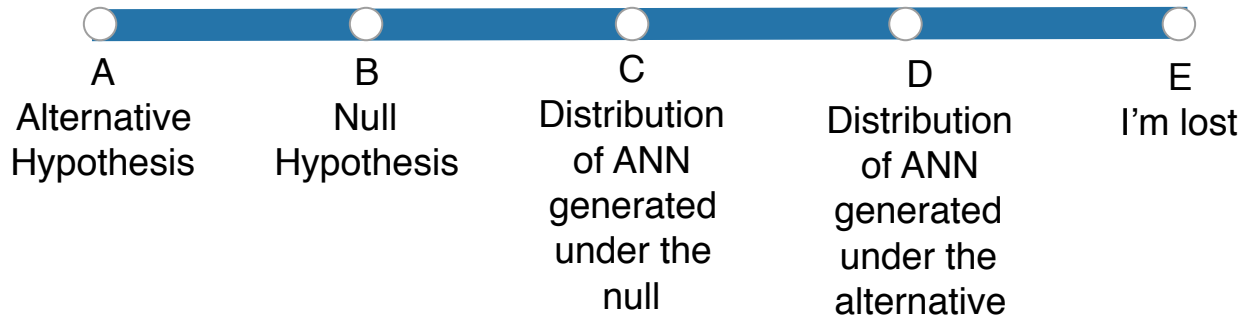


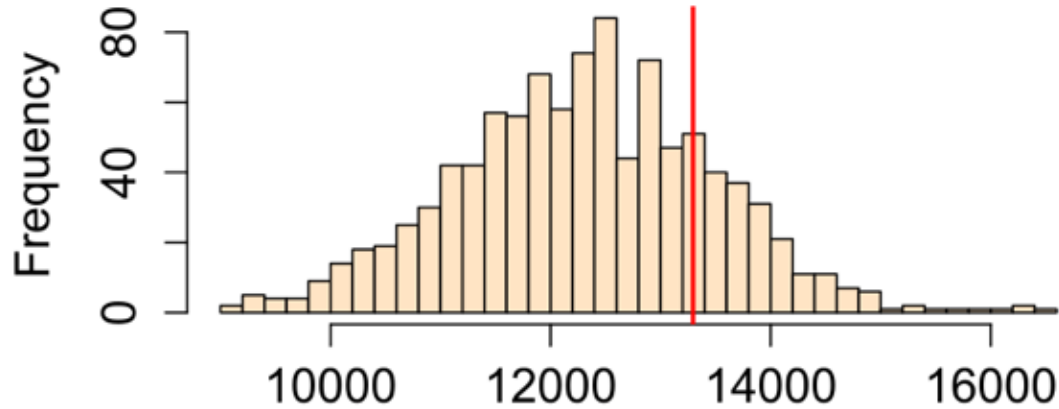
This is an example of bootstrapping!



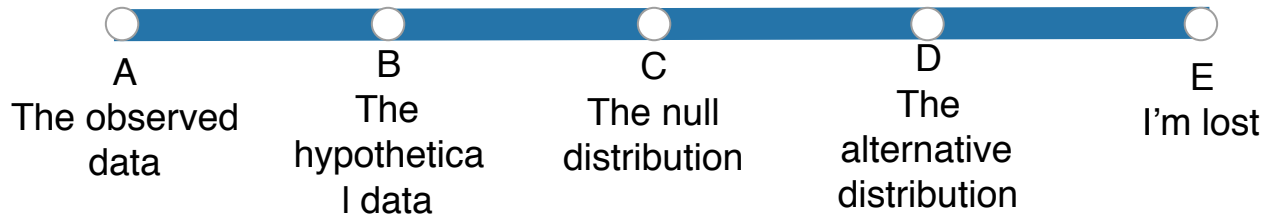


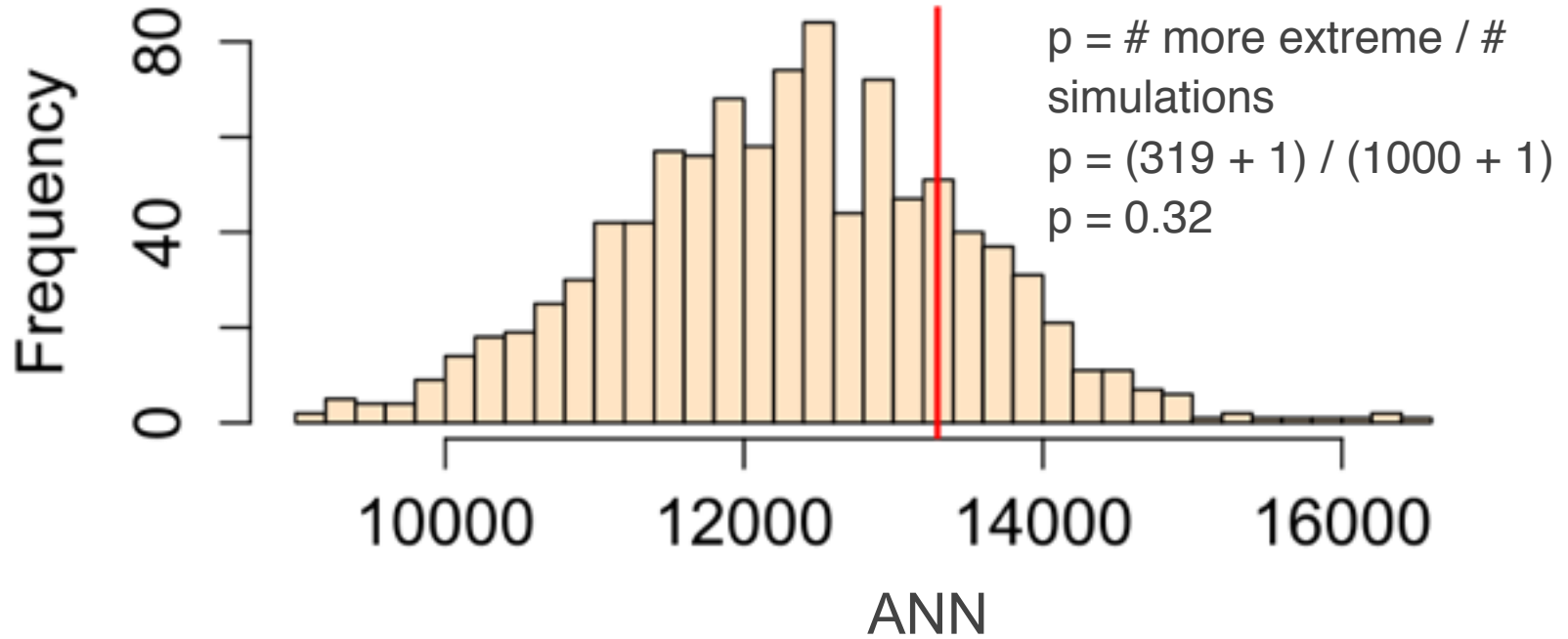
What does the histogram represent in this image?





What does the red line represent?

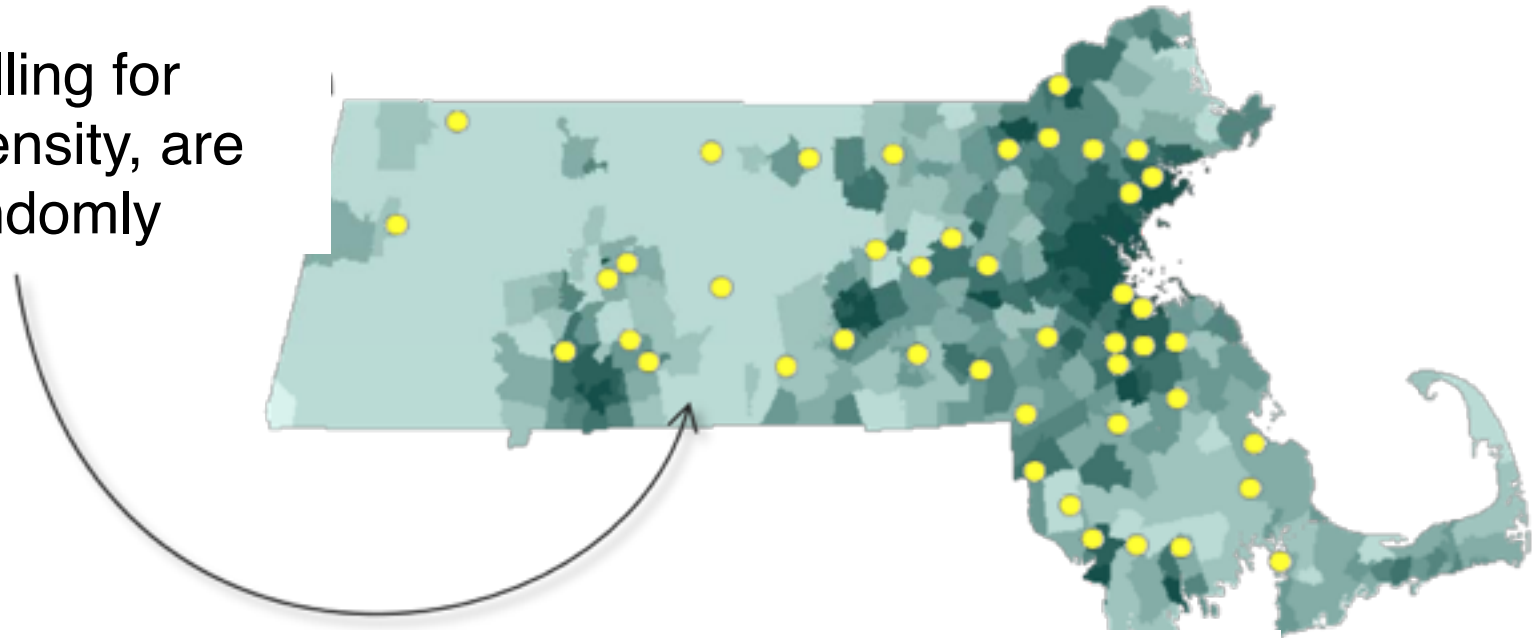




Fail to reject the null

Suggests that our results come from a CSR

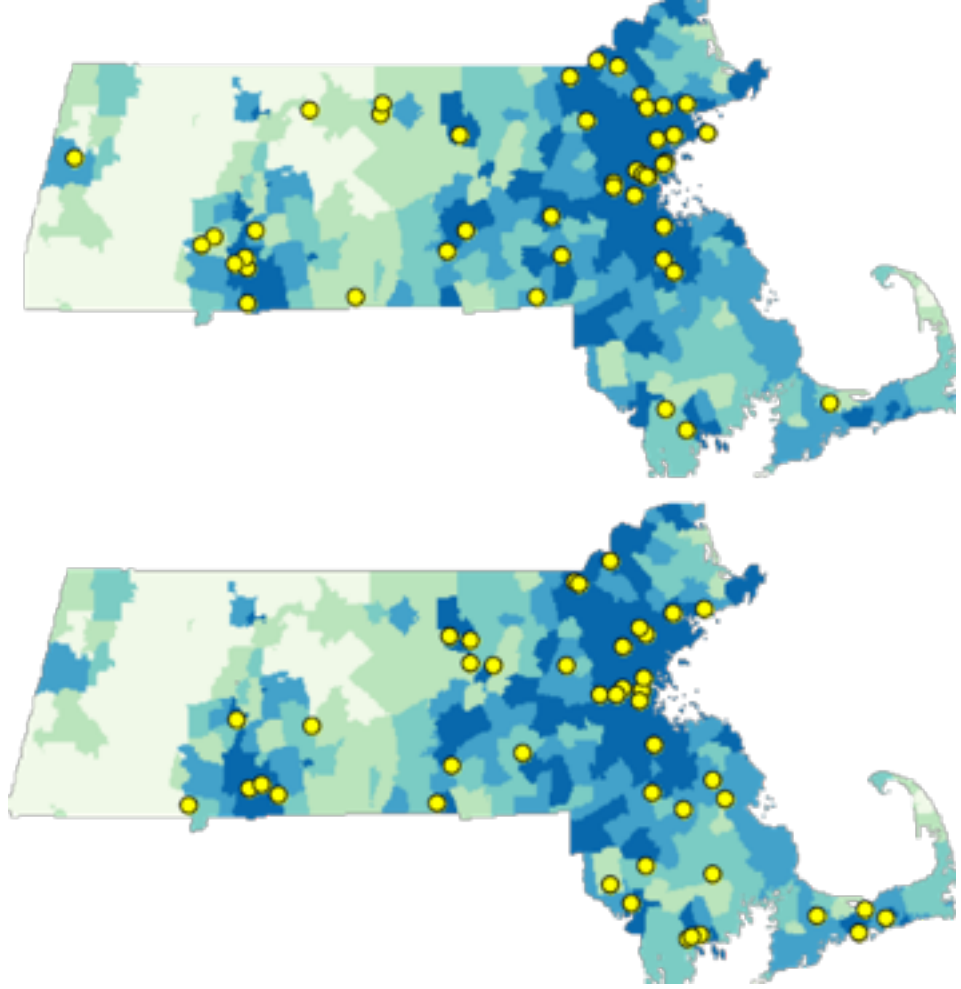
When controlling for population density, are Walmarts randomly distributed?

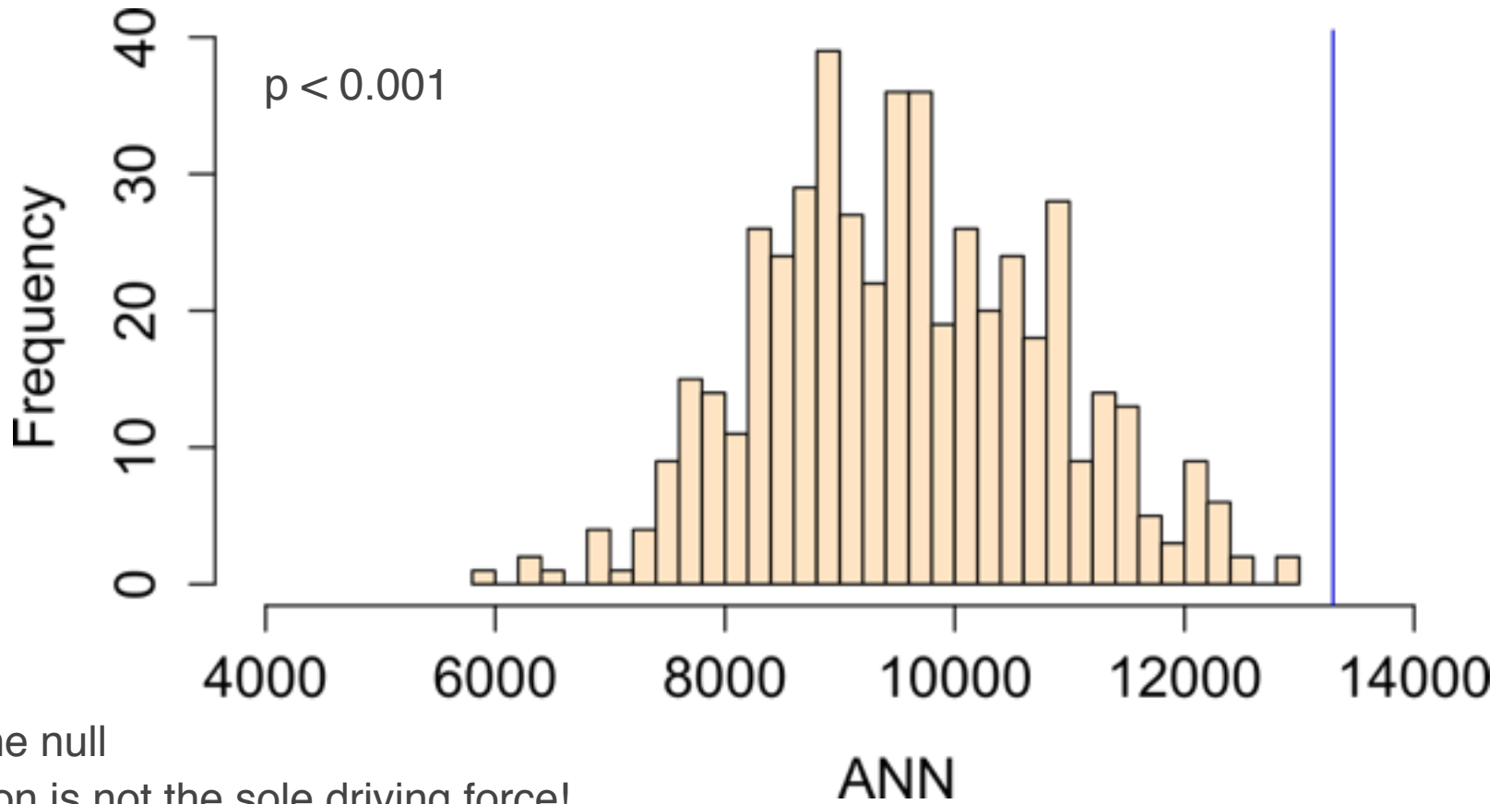


H_0 : Walmarts are distributed according to population density alone

H_a : Walmarts are *not* distributed based on population density alone

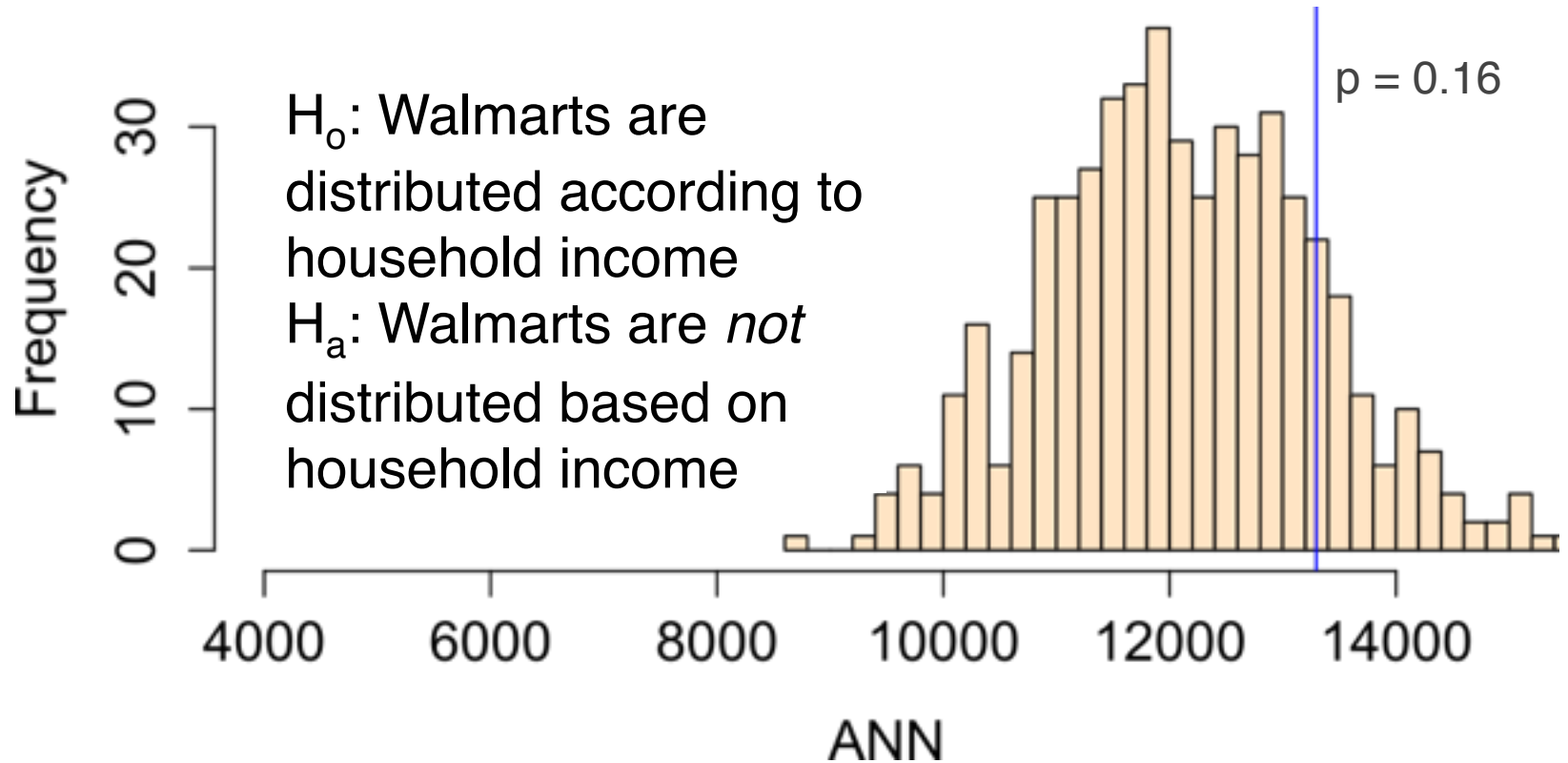
two randomly generated
point patterns using
population density as
the underlying process





Reject the null
Population is not the sole driving force!

Maybe median household income is the driving force...?



...Is it CSR or median household income?

hints at plausible scenarios, but doesn't tell us which one it is definitively.

Basic Geospatial Analysis: Summary

1. Considerations when visualizing spatial data important to conclusions drawn
 - a. values to plot?
 - b. map type?
 - c. color scale?
2. Traditional statistics fail with geospatial data:
 - a. Spatial autocorrelation
 - b. MAUP
 - c. Edge effects
 - d. Ecological fallacy
 - e. Nonuniformity of space
3. Analysis still possible
 - a. Global Point Density, Quadrat Density, Kernel Density
 - b. Poisson Point Process
 - c. K-Nearest Neighbor (KNN)
 - d. Comparison to a CRP (using simulation)