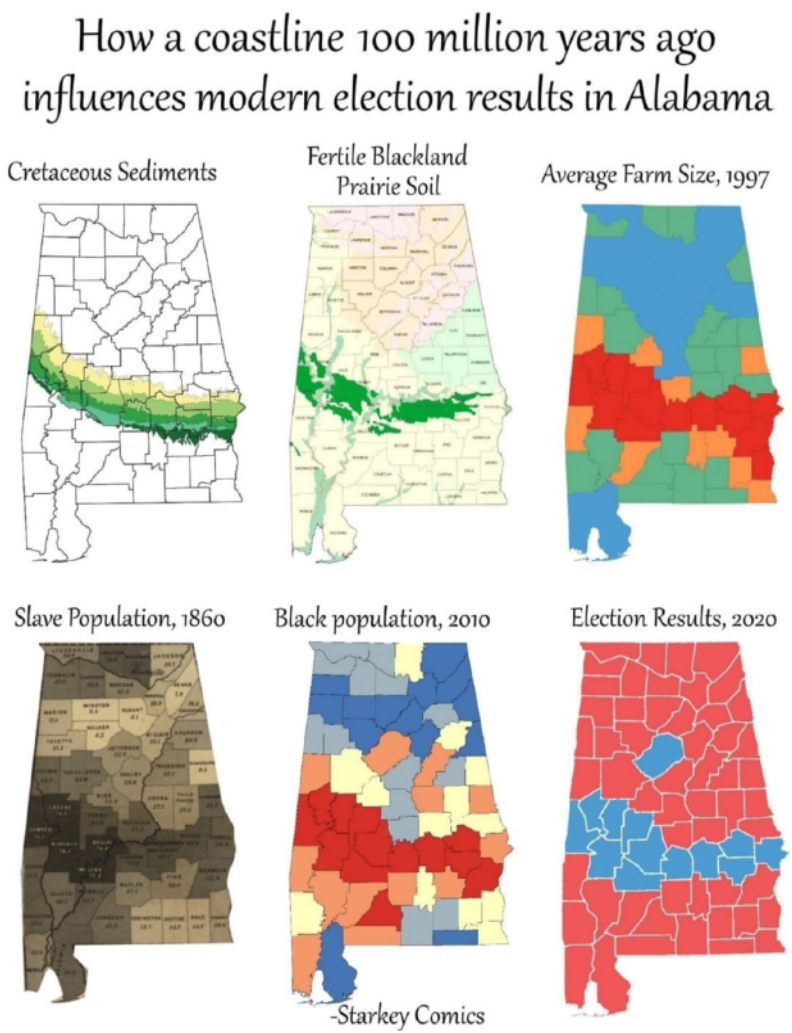
Geospatial analysis

Jason G. Fleischer, Ph.D.
Asst. Teaching Professor
Department of Cognitive Science, UC San Diego

jfleischer@ucsd.edu

@jasongfleischer

https://jgfleischer.com

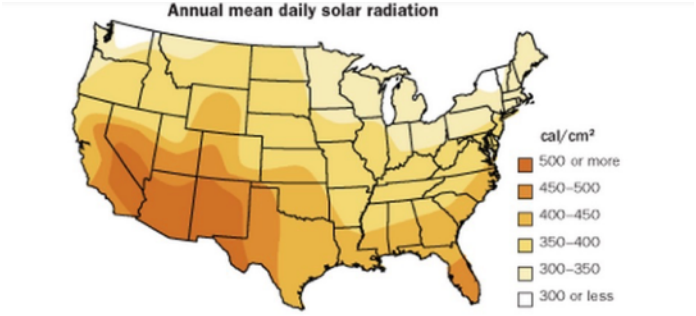


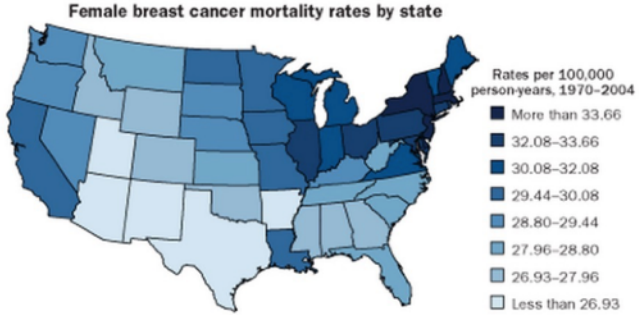
Why Geospatial Analysis?

'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

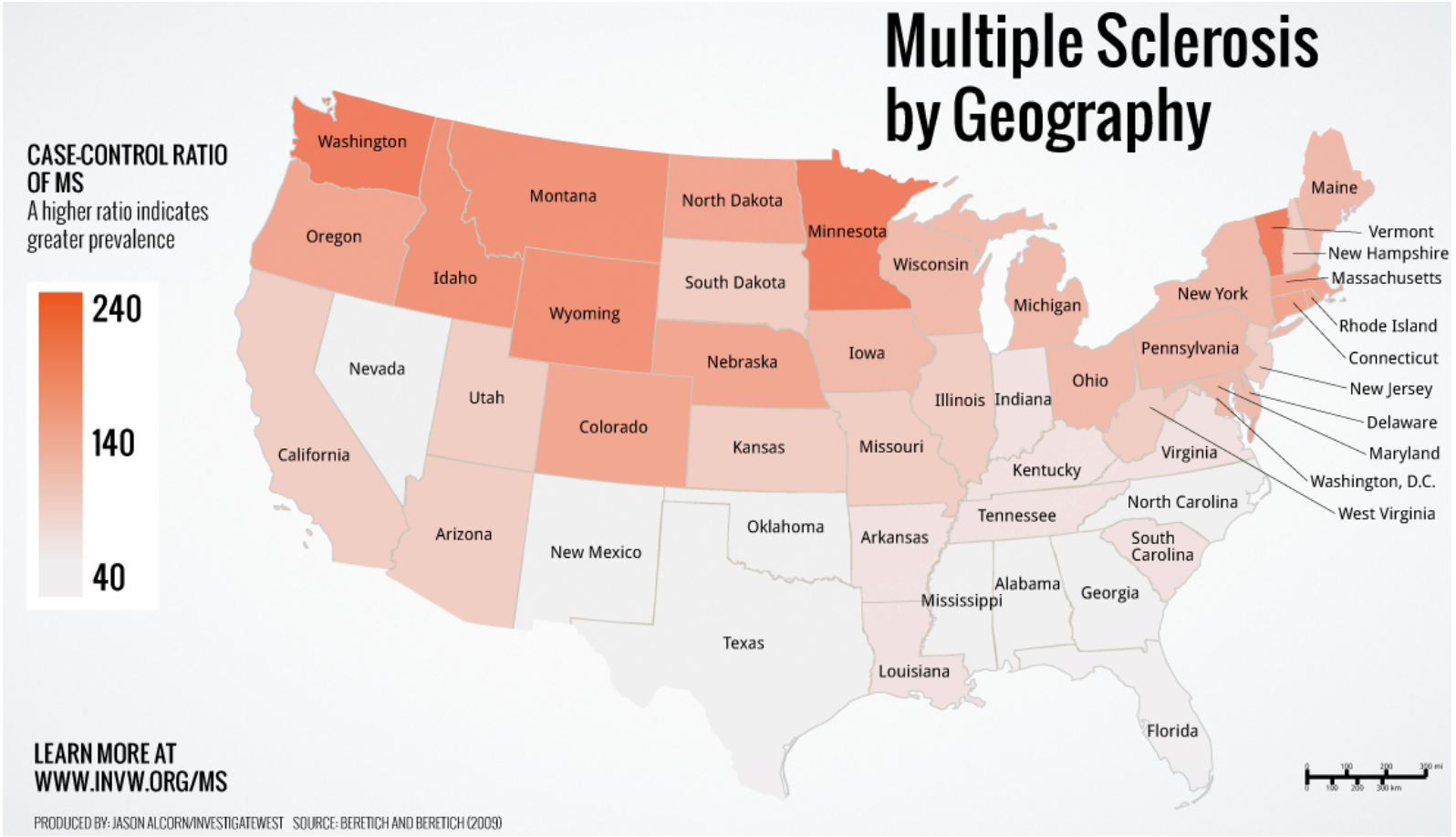




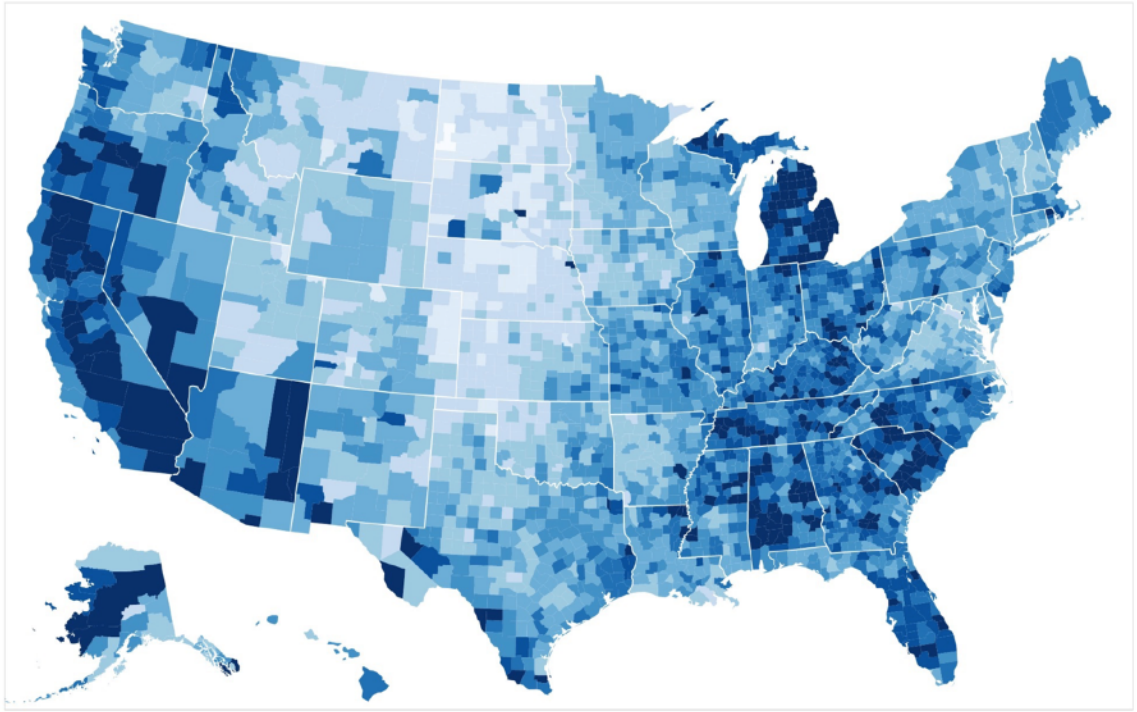
ON THE MAP Scientists who study vitamin D can't help but notice that a host of diseases seem to vary with latitude. Type 1 diabetes, multiple sclerosis and even some cancers appear to be more common in areas that get less sun -- meaning less opportunity for the body to produce vitamin D. The maps above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


Multiple Sclerosis by Geography
CASE-CONTROL RATIO OF MS
A higher ratio indicates greater prevalence
240
140
40
Washington
Oregon
California
Nevada
Idaho
Montana
Wyoming
Utah
Arizona
Colorado
New Mexico
North Dakota
South Dakota
Nebraska
Kansas
Oklahoma
Texas
Minnesota
Iowa
Missouri
Arkansas
Louisiana
Wisconsin
Illinois
Michigan
Indiana
Ohio
Kentucky
Tennessee
Mississippi
Alabama
Georgia
Florida
South Carolina
North Carolina
Virginia
West Virginia
Pennsylvania
New York
Maryland
Delaware
New Jersey
Connecticut
Rhode Island
Massachusetts
Vermont
New Hampshire
Maine
Washington, D.C.
LEARN MORE AT
WWW.INVW.ORG/MS
PRODUCED BY: JASON ALCORN/INVESTIGATEWEST SOURCE: BERETICH AND BERETICH (2009)

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.

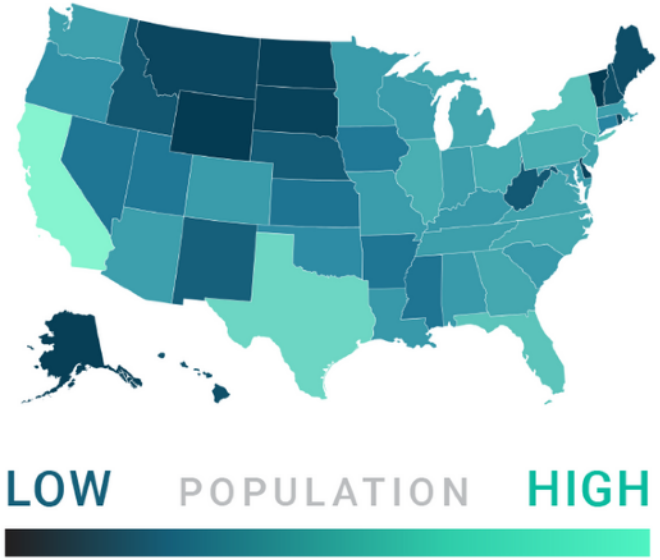
Choropleth maps are useful for visualizing *clear regional patterns* in the data

Source: http://bl.ocks.org/mbostock/4060606

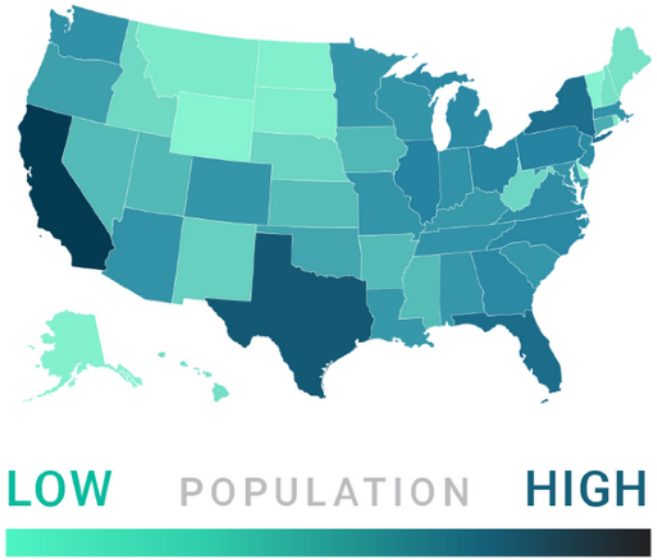
adapted from Brad Voytek

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

NOT IDEAL



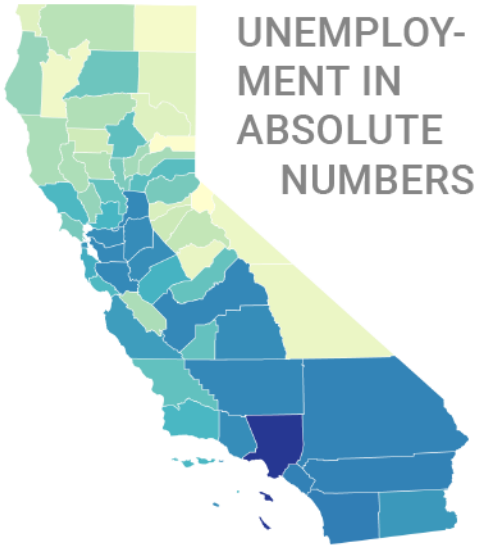
BETTER



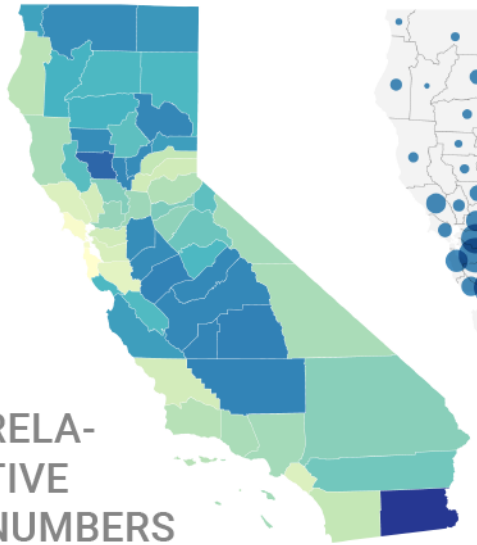
https://blog.datawrapper.de/colors/

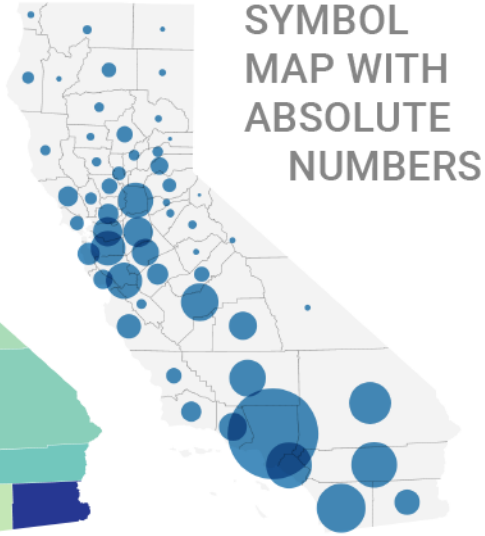
Choropleth should display relative differences, *not* absolute numbers

NOT IDEAL



BETTER

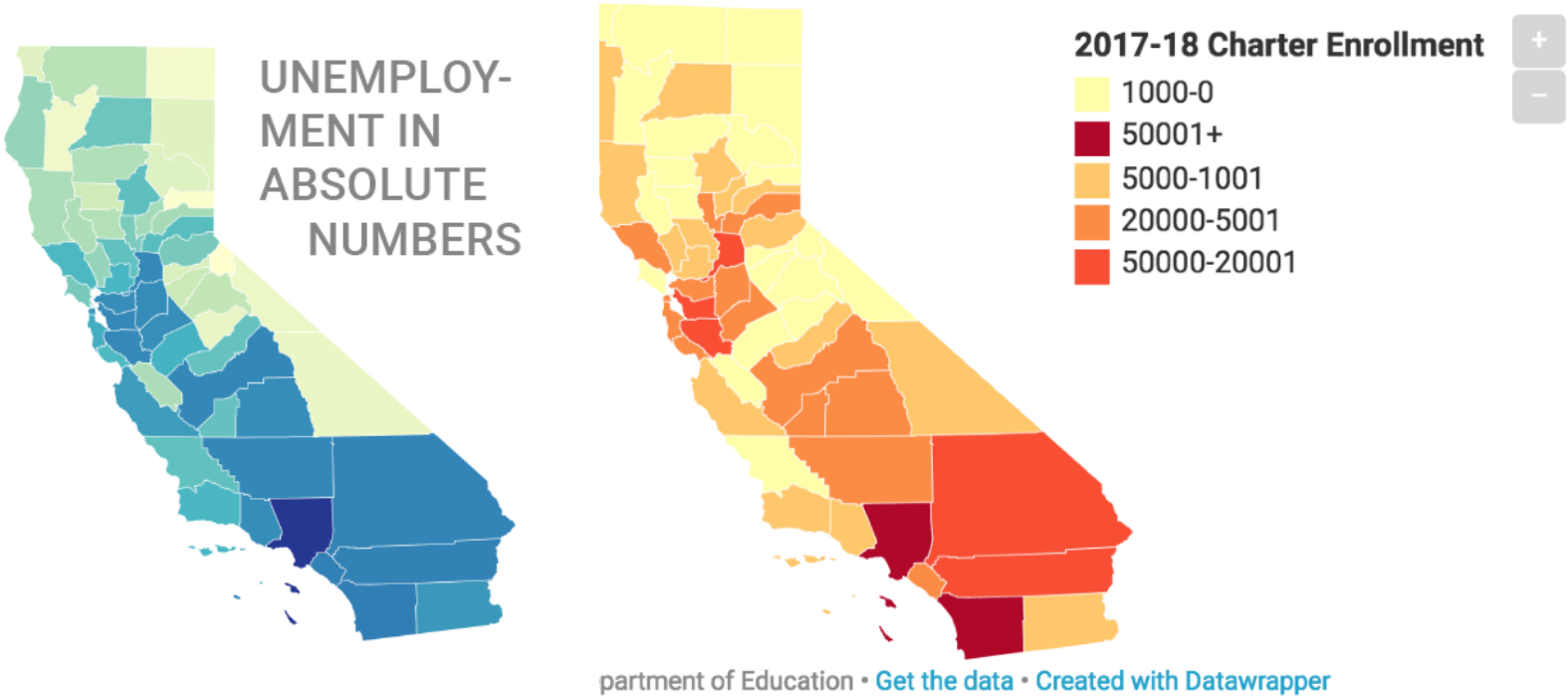




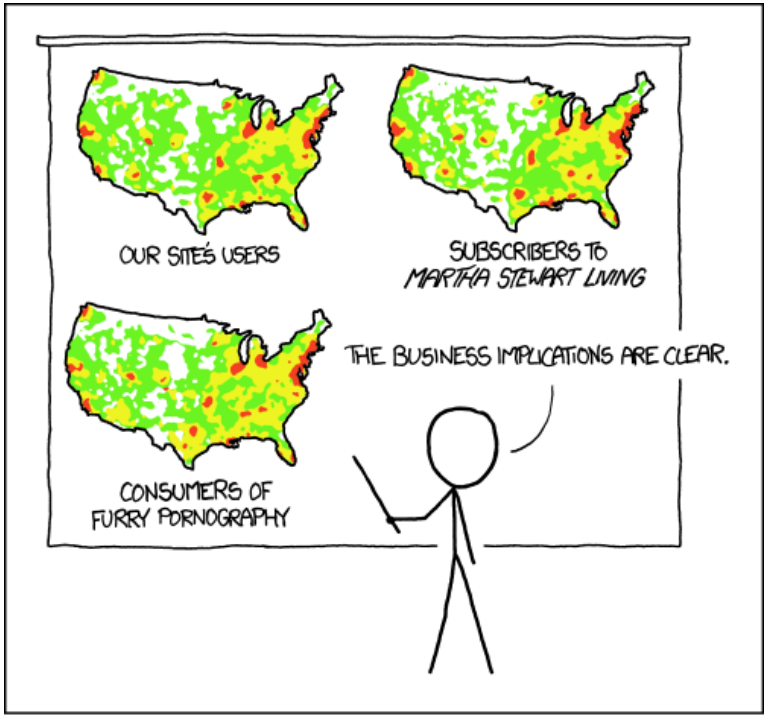
https://blog.datawrapper.de/choroplethmaps/

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



https://calmatters.org/articles/california-charter-school-data-enrollment-cost-teachers-unions/



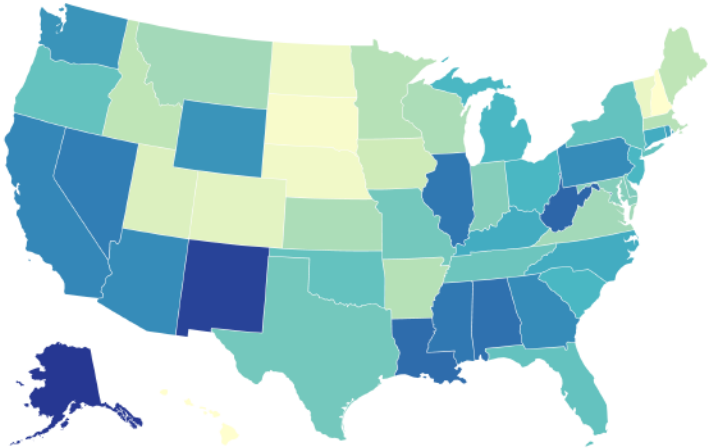
Choropleth maps can be misleading

https://xkcd.com/1138/

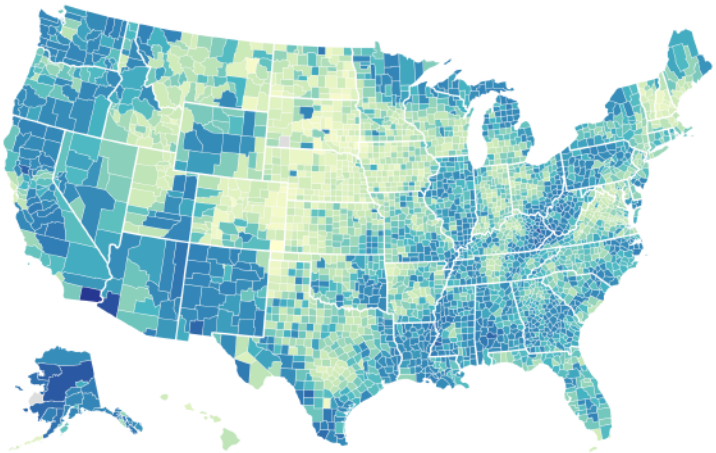
Consider using the smallest unit possible

(but there are exceptions!)

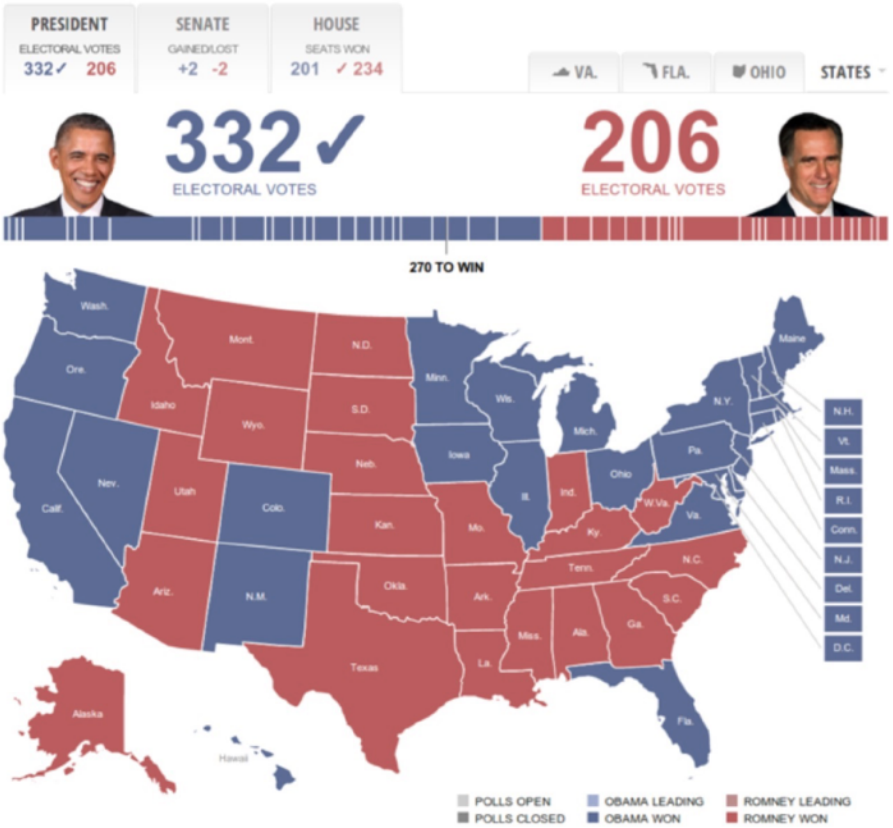
NOT IDEAL



BETTER



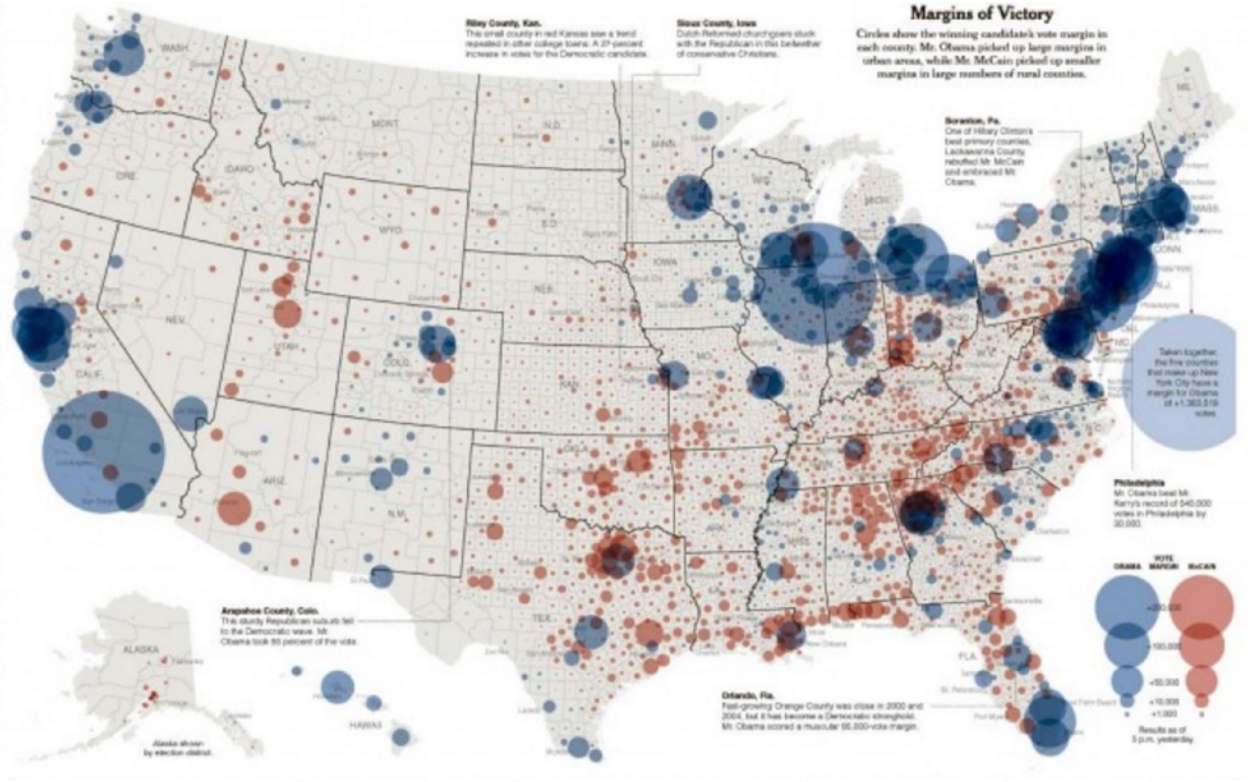
https://blog.datawrapper.de/choroplethmaps/



Sometimes summarizing at the state level is ok...

Source: *Huffington Post*, http://elections.huffingtonpost.com/2012/results

adapted from Brad Voytek



This **bubble graph** more accurately tells the full story, since the size of the bubbles is reflective of the population

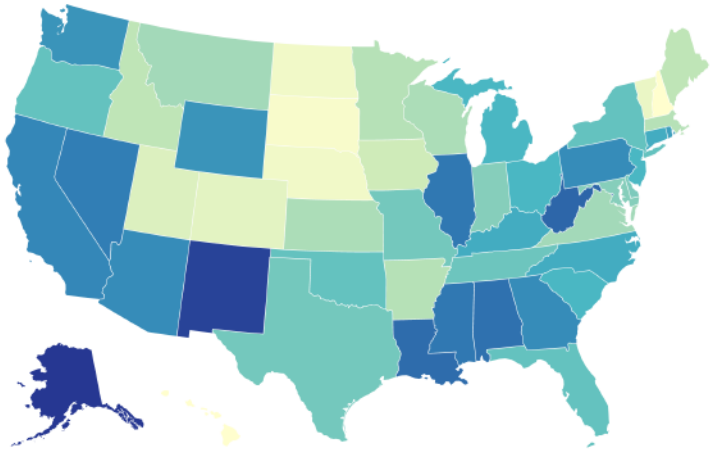
...but same data *can* be displayed more effectively and informatively.

Source: http://www.thefunctionalart.com/2012/11/when-choropleth-maps-deceive.html adapted from Brad Voytek

Visualization Choices

Cartograms should be considered when displaying how many people were affected

NOT IDEAL



Choropleths answer “How much area was affected?”

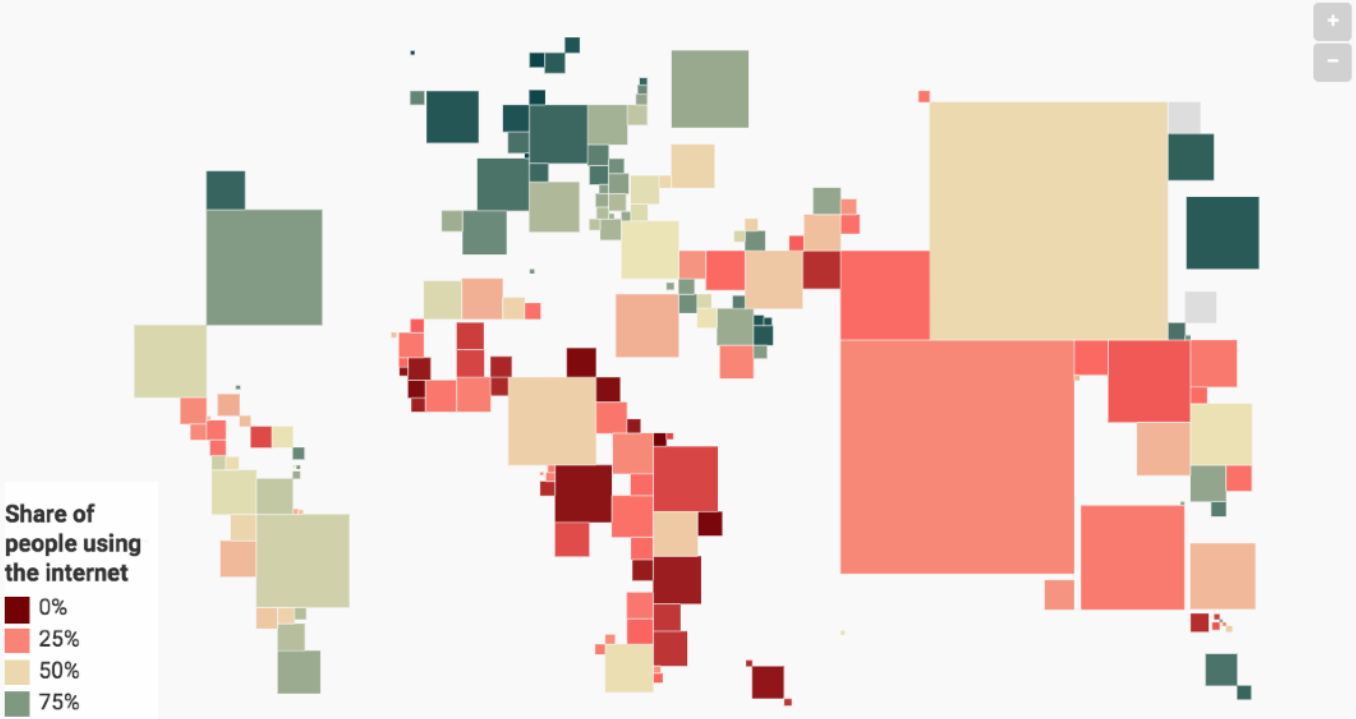
BETTER

Cartograms answer “How many people were affected?”

https://blog.datawrapper.de/choroplethmaps/

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.

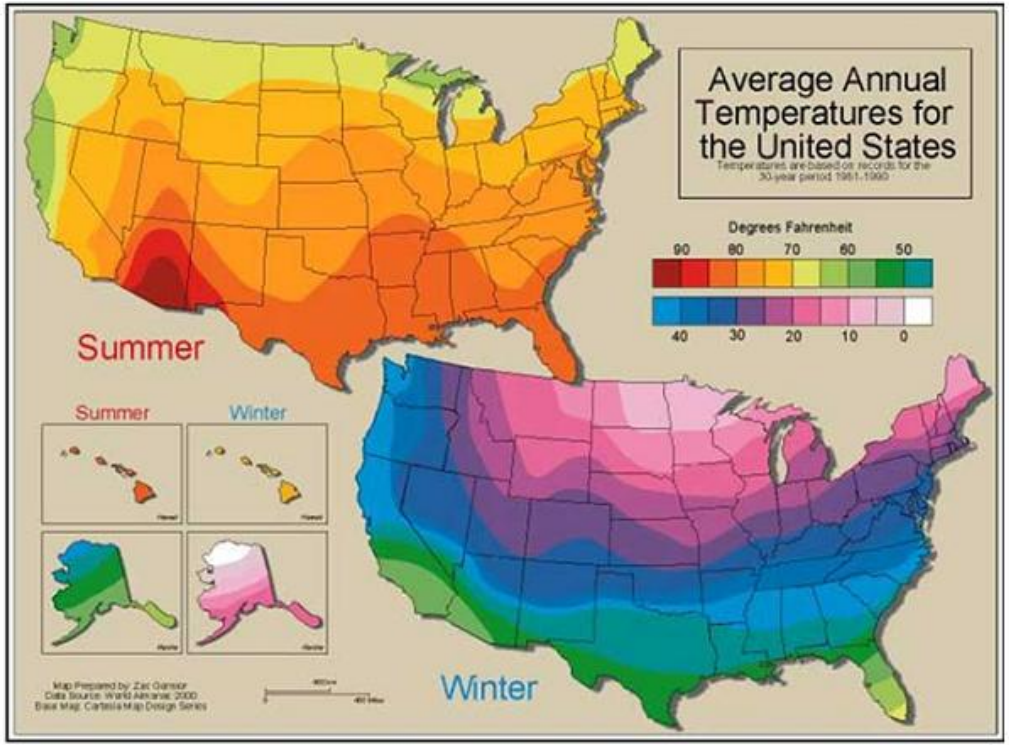


Source: Our World in Data · Get the data

https://blog.datawrapper.de/cartograms/

Isarithmic maps demonstrate smooth, continuous phenomena

(temperature, elevation, rainfall, etc.)

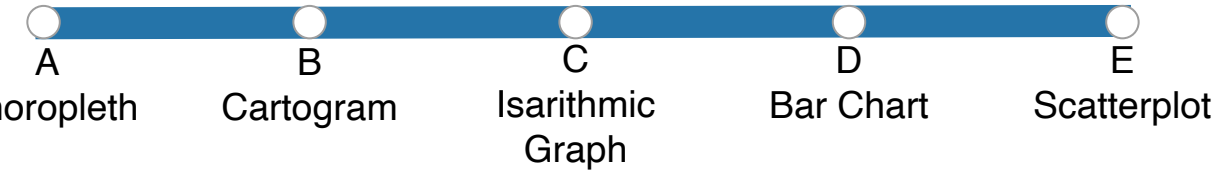


http://enb105-2012s-dem.blogspot.com/2012/03/isarithmic-map.html

Visualizing Geospatial Data

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:

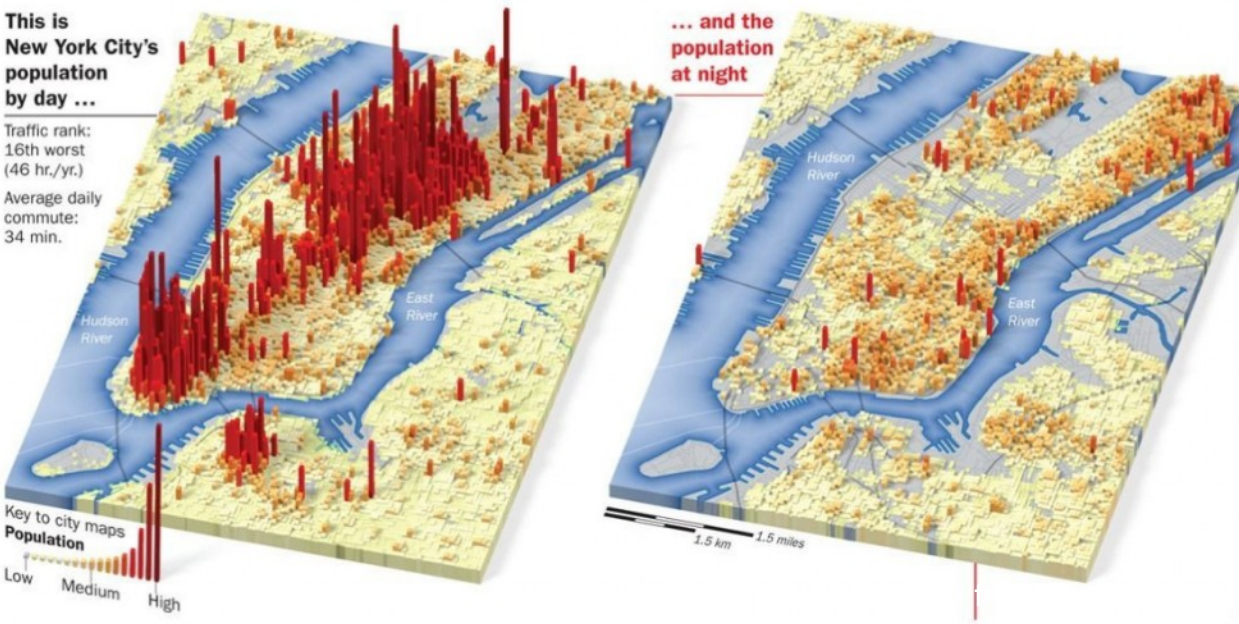
Violations of conventional statistics:

- 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



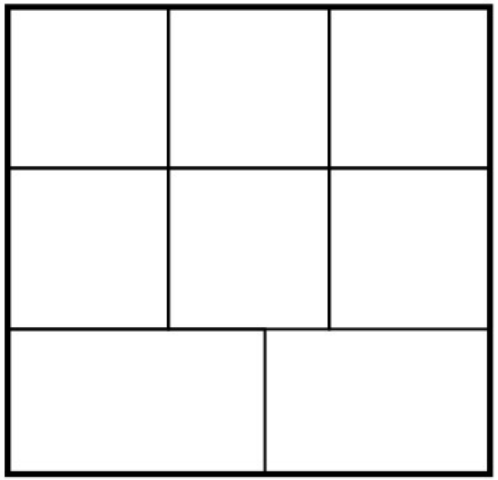
Modifiable Areal Unit Problem (MAUP)

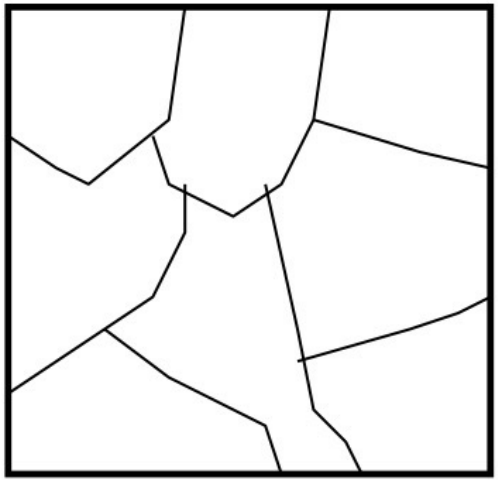
The aggregation units used are arbitrary with respect to the phenomena under investigation, yet the aggregation units used will affect statistics determined on the basis of data reported in this way.

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

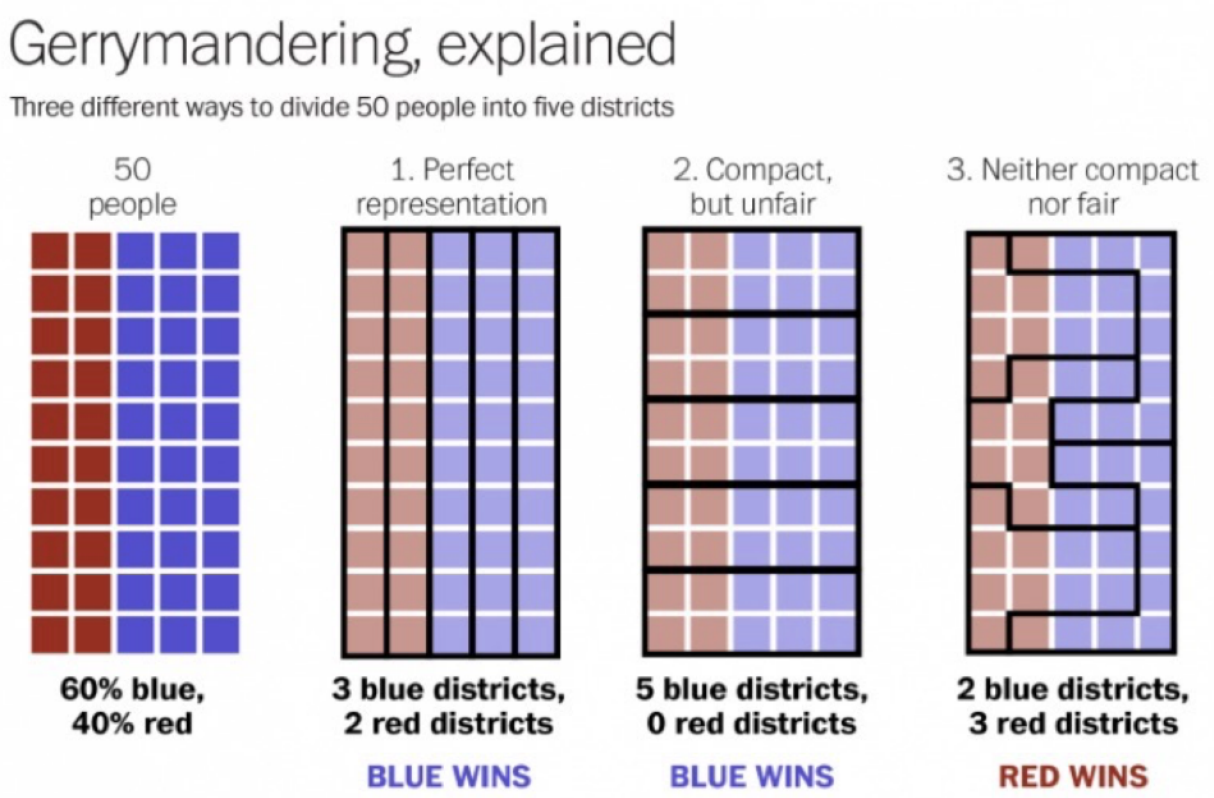
Modifiable Areal Unit Problem (MAUP)

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





For example...gerrymandering



https://www.washingtonpost.com/news/wonk/wp/2015/03/01/this-is-the-best-explanation-of-gerrymandering-you-will-ever-see/

adapted from Brad Voytek

For example...gerrymandering

North Carolina

DISTRICTS REDRAWN TO OPTIMIZE COMPACTNESS

SOURCE: U.S. Census Bureau (top), Brian Olson (bottom)
GRAPHIC: The Washington Post. Published June 3, 2014

Welcome to Hexapolis

Every 10 years, Hexapolis redraws its congressional district lines — just like the United States does. But Hexapolis is a simpler place.

Lawmakers in either the **Purple Party** or **Yellow Party** control redistricting. To increase their advantage in upcoming elections, they have been known to gerrymander egregiously — even if it means leaving some voters disenfranchised.

Hexapolis has nine districts. Even though a majority of voters favor the Purple Party, that does not mean that the Yellow Party can't shift the state's partisan tilt.

https://www.nytimes.com/interactive/2022/01/27/us/politics/congressional-gerrymandering-redistricting-game-2022.html

Modifiable Areal Unit Problem (MAUP)

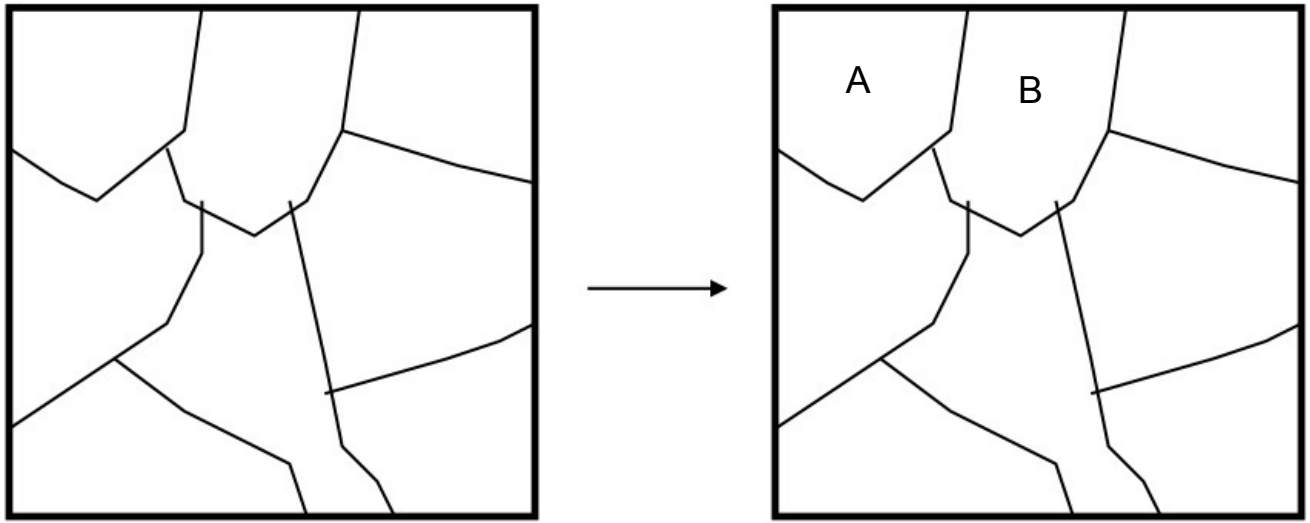
Potential problems in almost every field that utilizes spatial data.

In the 2016 U.S. presidential election, Hillary Clinton, with more of the population vote than Donald Trump, but failed to become president. (also true in Gore/Bush 2000)

A different aggregation of U.S. counties into states could have produced a different outcome.

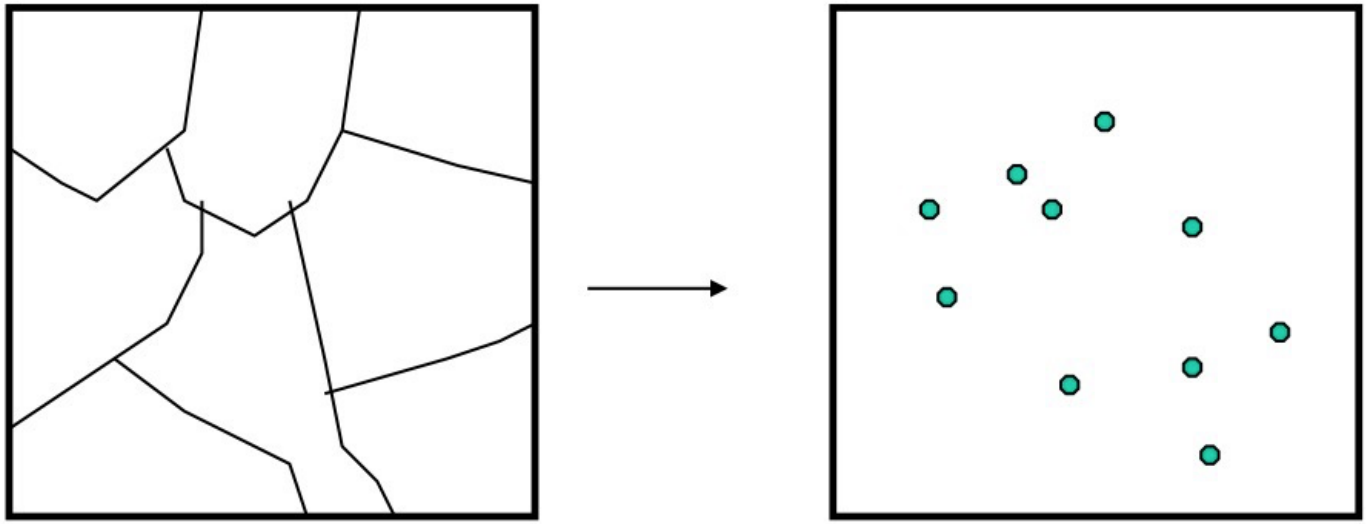
Edge Effects (The Boundary Problem)

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.

Conclusions we might draw:

- 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 The only valid conclusion!

Ecological Fallacy

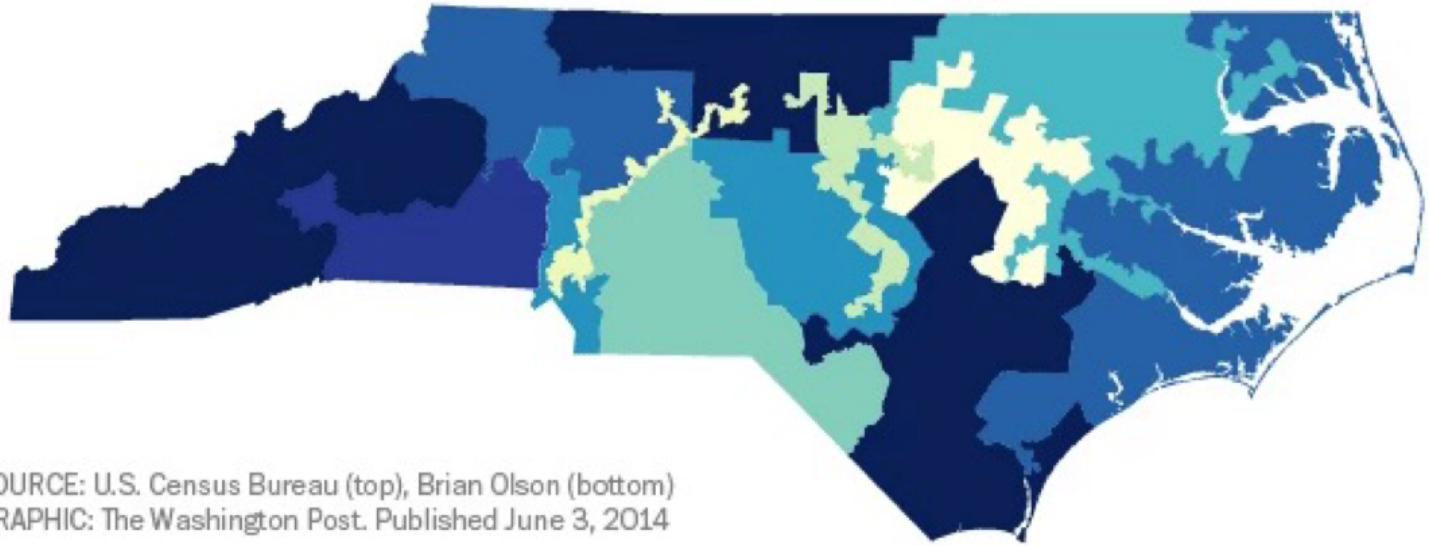
Issues:
Inferences drawn about associations between the characteristics of an aggregate population and the characteristics of sub-units within the population are wrong. That is: *results from aggregated data (e.g. counties) cannot be applied to individual people*

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

For example...gerrymandering

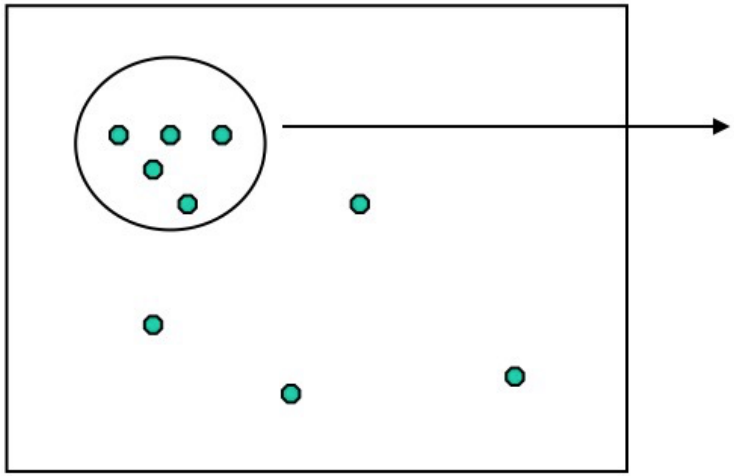
North Carolina

CURRENT CONGRESSIONAL DISTRICTS



SOURCE: U.S. Census Bureau (top), Brian Olson (bottom)
GRAPHIC: The Washington Post. Published June 3, 2014

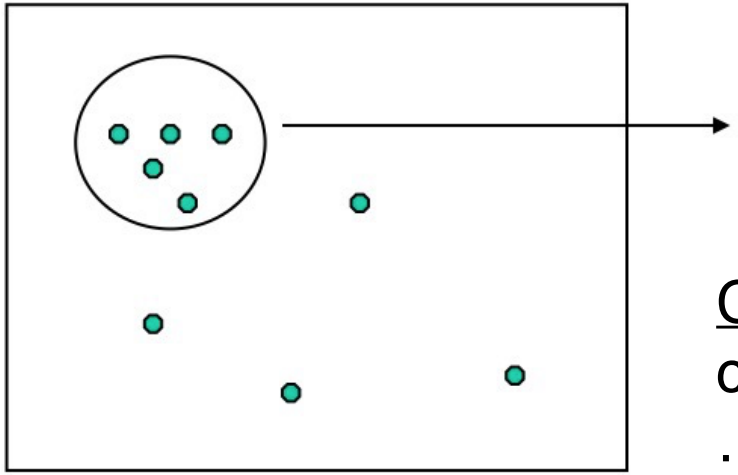
Nonuniformity



Area with high crime rates?

Crime locations

Nonuniformity

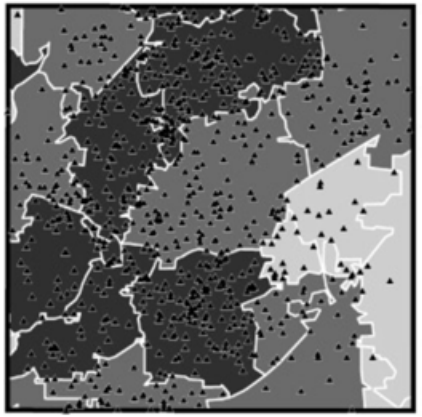


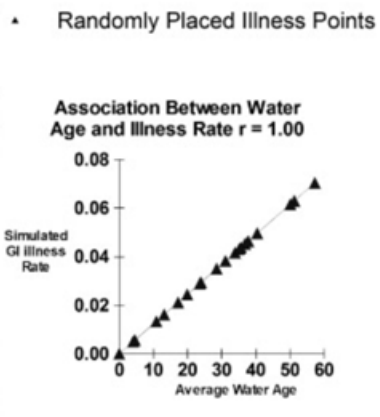
Crime locations

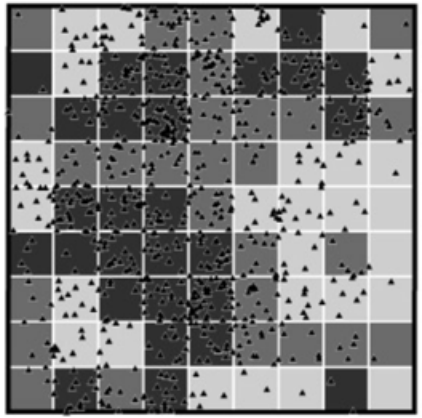
Area with high crime rates?

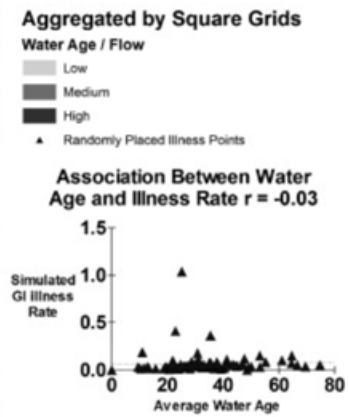
Conclusion: Bank robberies are clustered
….but only because banks are clustered!

Spatial Statistics









What explains what's going on here?

A
Spatial Autocorrelation

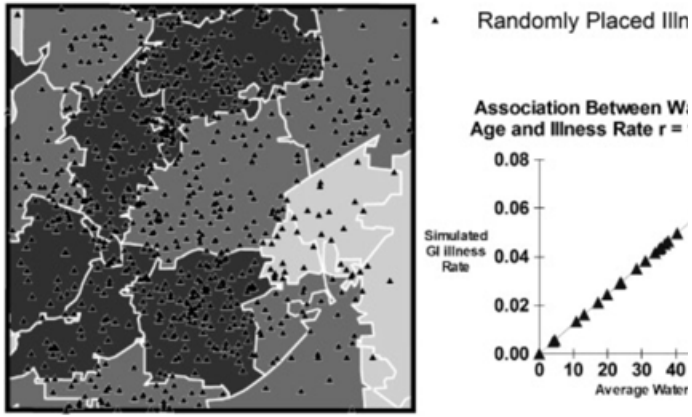
B
MAUP

C
Edge Effects

D
Ecological Fallacy

E
Nonuniformity

Spatial Statistics



What explains v

A
Spatial Autocorrelation

B
MAUP