Intro to Spatial Statistics

January 18, 2022

Pratt

What is Statistics?

What is Statistics?

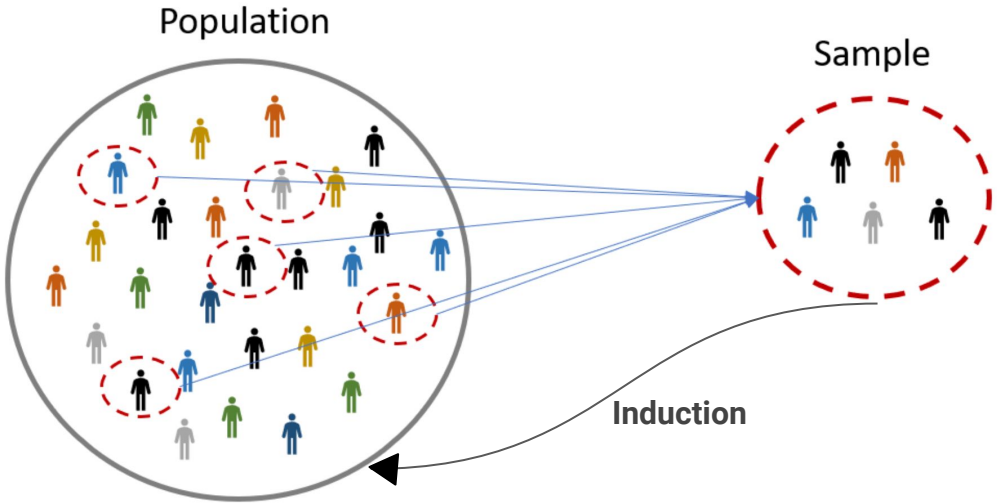
- The science (and art) of learning from data
- Concerned with the collection, analysis, interpretation and communication of empirical data
- Collection of theory and techniques with which the uncertainty of ***inductive inferences*** may be evaluated

Inductive Inference

- A method of reasoning in which a general principle is synthesized from a finite set of observations
- Extends the observed pattern or relation to other future or unknown instances
- Relies on using evidence to make a generalized claim

Inductive Inference

- Make claims about an unknown population from a known sample, while acknowledging and evaluating the uncertainty inherent in doing so



Inductive Inference

- Example
  - Every raven in a random sample of 1,000 ravens is black
  - This supports the conclusion that **all ravens are black**
  - Thus, if something is a raven, then it is (probably) black
  - Thus, if something is not black, then it is not a raven

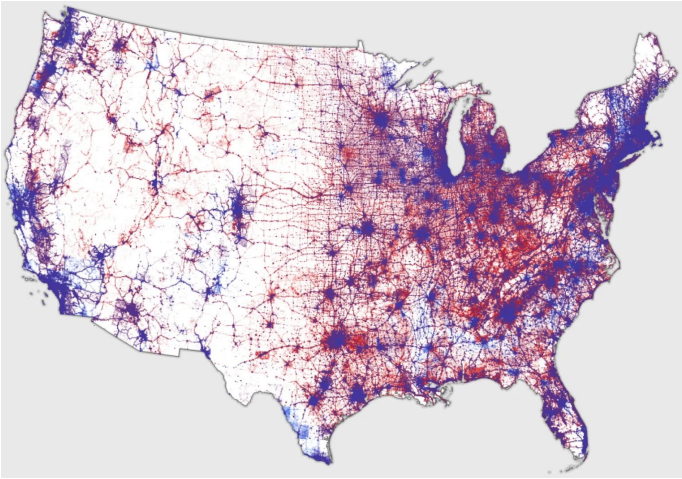
Inductive Inference

≠

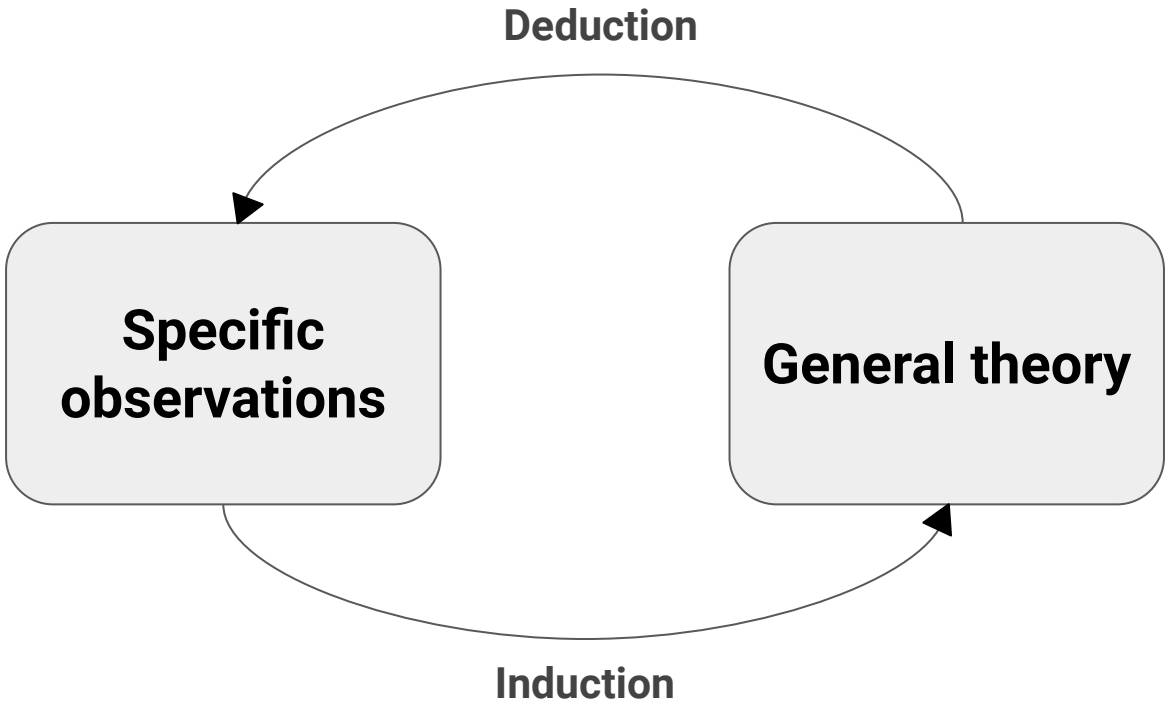
All ravens are black. This apple is green. Therefore, this apple is not a raven.

Inductive Inference

- Example
  - 62% of voters in a random sample of 400 registered voters said that they favor John Kerry over George W. Bush for President in the 2004 Presidential election
  - This supports with a probability of at least 95% the following conclusion: Between 57% and 67% of all registered voters favor Kerry over Bush for President



Induction vs. Deduction



Why do we need Statistics?

- In general, the goal of the scientific method is to objectively evaluate a hypothesis on the basis of experimental results
- Objective evaluation of a hypothesis is often difficult because:
  - It is not possible to observe all conceivable events (population vs. sample)
  - Inherent variability exists when exact laws of cause and effect are unknown

Why do we need Statistics?

- Scientists must often reason from particular cases to wider generalizations
    - This is a process of uncertain inductive inference
- Statistics is a collection of theory and tools that enables the researcher to make inductive inferences and evaluate their uncertainty, thereby drawing meaningful conclusions from data

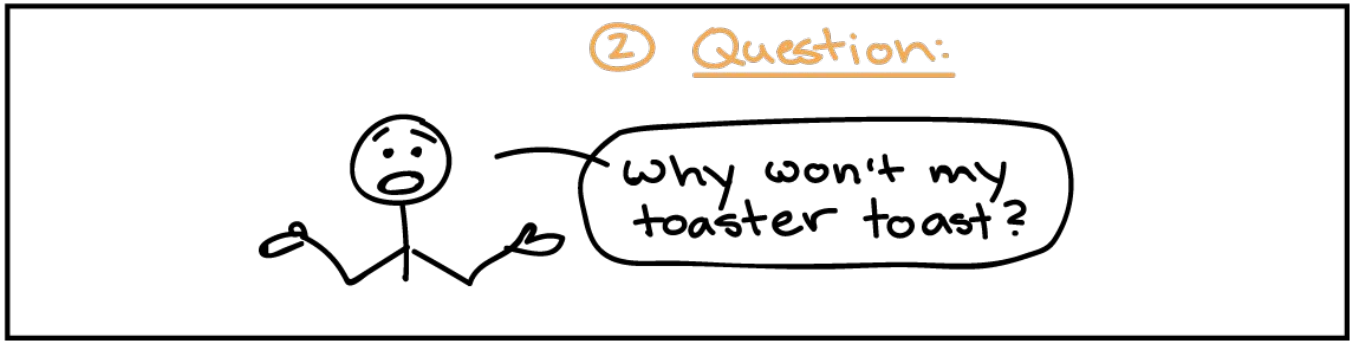
The Scientific Method

Scientific Method

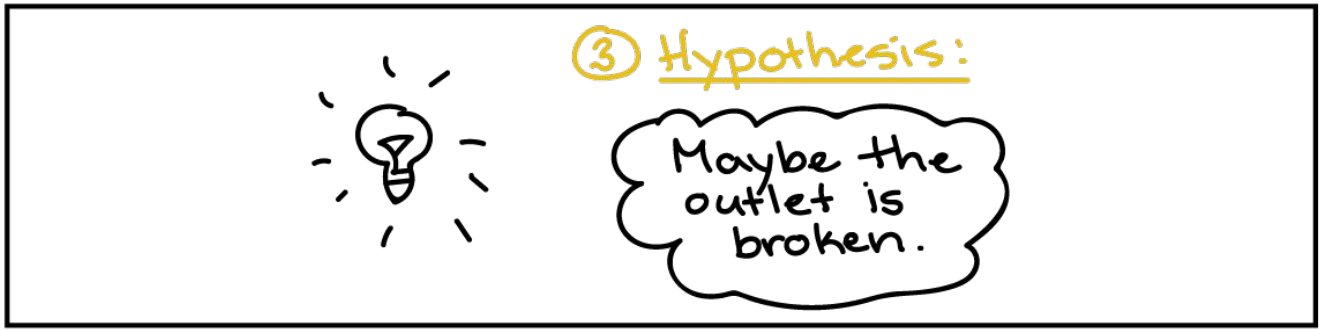
1. Make an observation
2. Ask a question
3. Form a hypothesis
4. Make a prediction based on that hypothesis
5. Test the prediction
6. Iterate: use the results to make new hypotheses or predictions

Make an observation

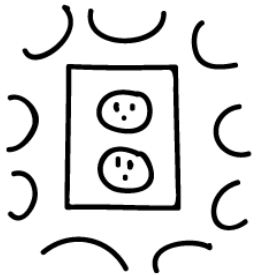
Ask a question



Propose a hypothesis



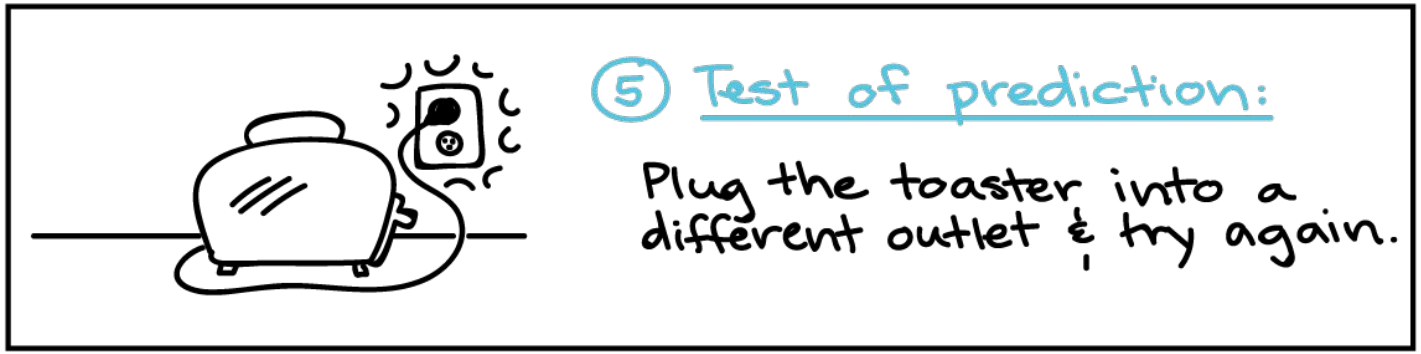
Make a prediction



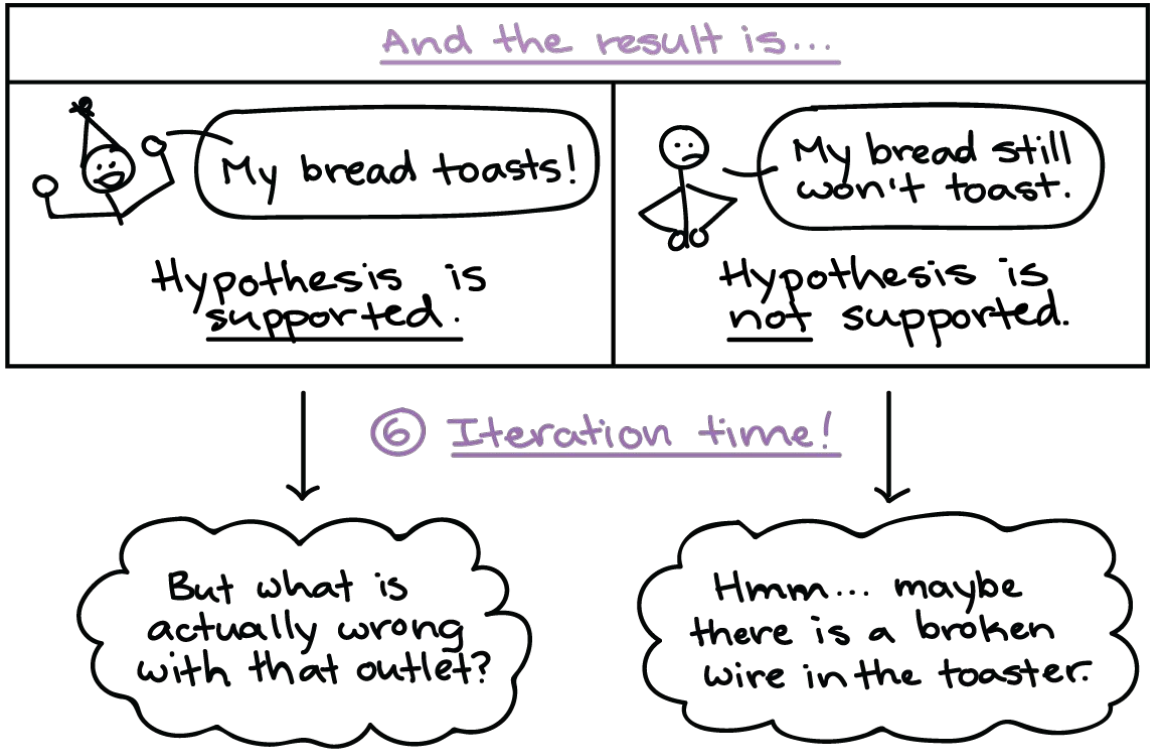
④ Prediction:

If I plug the toaster into a different outlet, then it will toast the bread.

Test the prediction

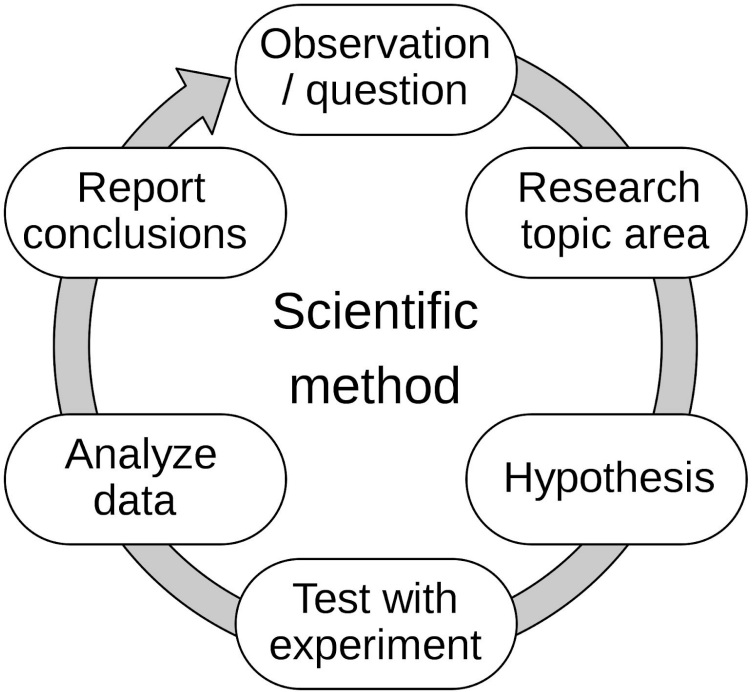


Iterate


And the result is...
My bread toasts!
Hypothesis is supported.
My bread still won't toast.
Hypothesis is not supported.
6 Iteration time!
But what is actually wrong with that outlet?
Hmm... maybe there is a broken wire in the toaster.

Statistics and the scientific method

- Statistics is relevant to all facets of the scientific method, from the initial planning of an experiment, to collecting and analyzing data, summarizing the results and evaluating the uncertainty of any inference drawn from the results



What is Spatial Statistics?

What is spatial statistics?

- Methods for analyzing patterns, processes and relationships in spatial data
- Developed specifically for use with geographic data
- Incorporates space directly into the mathematics
    - Proximity
    - Area
    - Connectivity
    - And/or other spatial relationships

Going beyond the map

- We use maps to display and interpret spatial information
- Spatial statistics enables us to go “beyond the map”
- Adds value and robustness to inductive inferences made from spatial data
- (Geospatial) knowledge discovery from data (KDD)

Spatial statistics questions

- **Where** do things happen? (patterns, clusters, hot spots, outliers)
- **Why** do they happen where they happen? (spatial processes)
- **How** does where things happen affect other things? (context, environment)
- **How** does the context affect what happens? (interaction)
- Where **should** things be located? (optimization)

When is analysis spatial?

- Spatial data = value + **location**
- Non-spatial analysis: location does **not** matter (location invariance)
- Spatial analysis: when the location changes, the information content of the data changes

The i.i.d. assumption

- Classical, non-spatial statistical methods often assume that variables are "independent and identically distributed" **(i.i.d.)**
- For example, consider tossing a coin several times
  - Each coin toss does not affect the others (independence)
  - The chance of heads coming up in each toss is always 0.5 (identically distributed)

Statistical issues with spatial data

- Spatial data often violates the i.i.d. Assumption
  - Units of geographic data are often like “bunches of grapes” (Stephan, 1934)
- Artificially imposed aggregation units can introduce bias
  - Modifiable Areal Unit Problem (MAUP)
- Global models might not explain local phenomenon
  - Spatial nonstationarity

Statistical opportunities with spatial data

- Analyzing clustering, hot spots & cold spots, outliers can be quite useful
  - Spatial autocorrelation is a measure of the tendency for neighbors to have similar values
- Techniques to work with spatially aggregated data
  - Avoiding common pitfalls due to issues such as MAUP
- Local Indicators of Spatial Association (LISA)
  - A class of local statistics that measure spatial association in sub-regions of the study area

Tobler’s First Law of Geography

Everything is related to everything else, but near things are more related than distant things.

- Waldo Tobler, 1970

Spatial Non-Stationarity

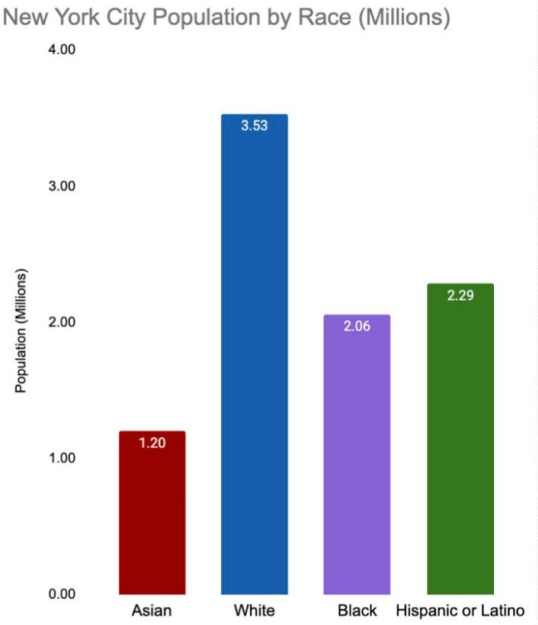
- As the location changes, the information changes
- Non-stationary
    - Variables behave differently based on their location/regional variation

Spatial Autocorrelation

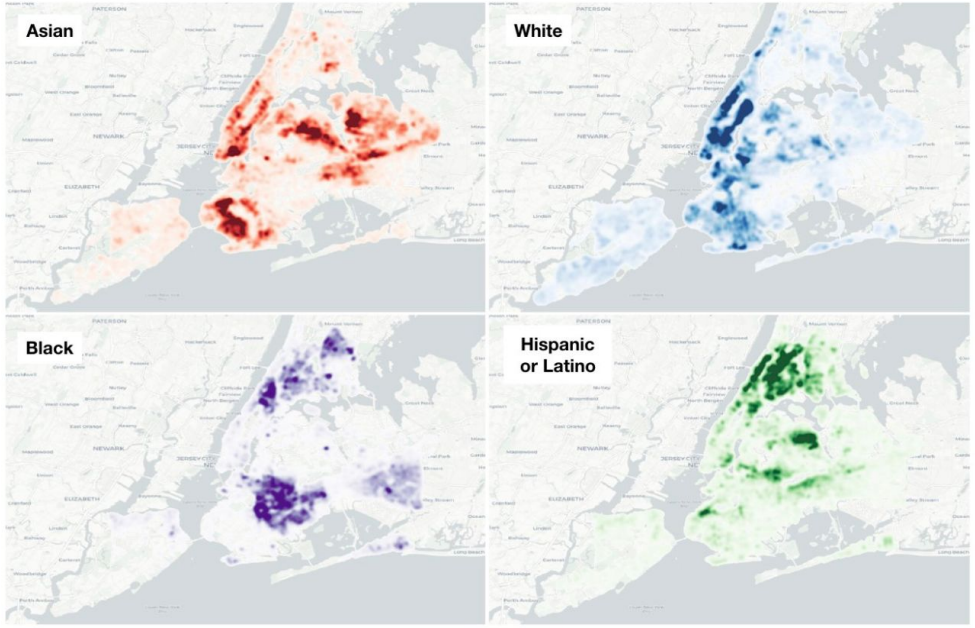
- Indicates if there is clustering or dispersion in a mapped pattern
    - Clustering = features near each other are more similar than those further away

Ignoring space can hide important information

Non-Spatial View



Spatial View



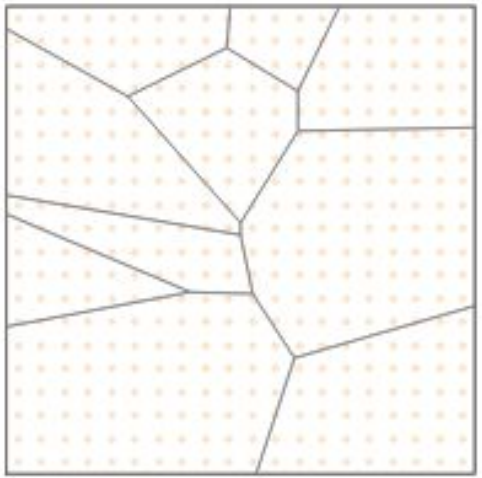
Modifiable Areal Unit Problem

- When point-based measures of spatial phenomena are aggregated into districts, the resulting summary values are influenced by both the shape and scale of the aggregation unit

Modifiable Areal Unit Problem

- Consider a bunch of points distributed uniformly across space

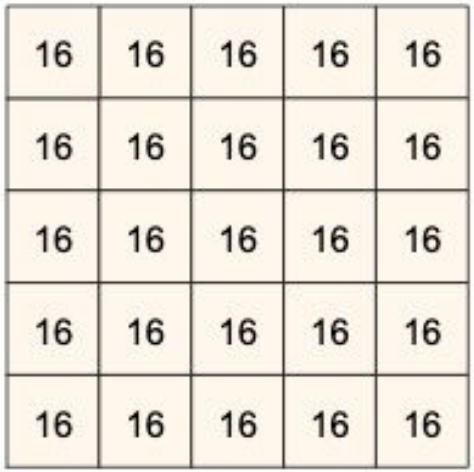
Uniform grid



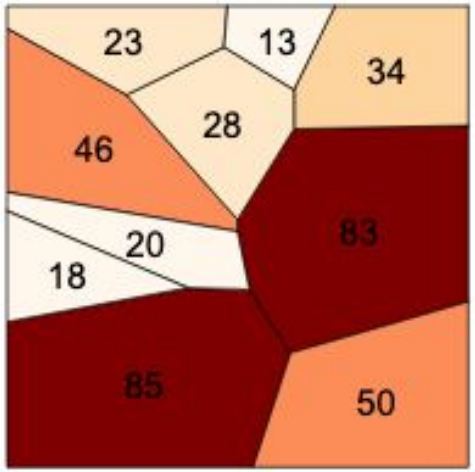
Irregular polygons

Modifiable Areal Unit Problem

- If we sum the number of individuals in each polygon, we get two maps that appear to be giving us two completely different population distribution patterns



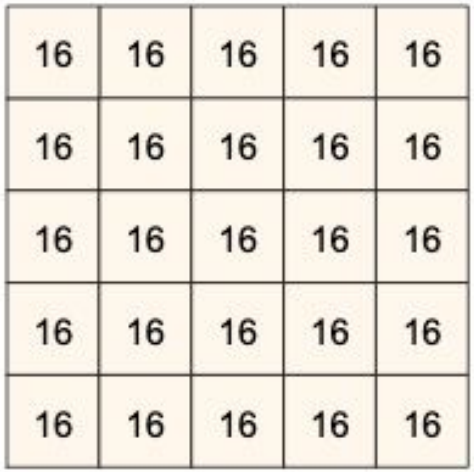
Uniform grid



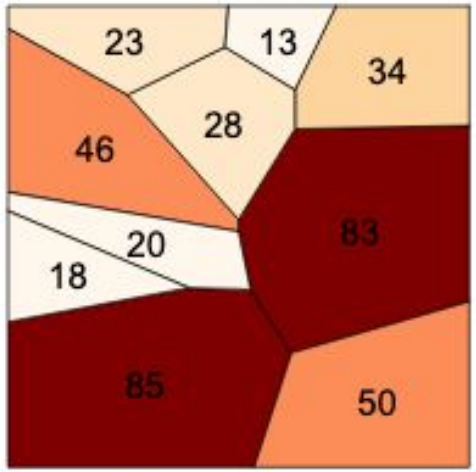
Irregular polygons

Modifiable Areal Unit Problem

- The maps highlight how count data aggregated by non-uniform areal units can fool us into thinking a pattern exists when in fact this is just an *artifact of the aggregation scheme*



Uniform grid



Irregular polygons

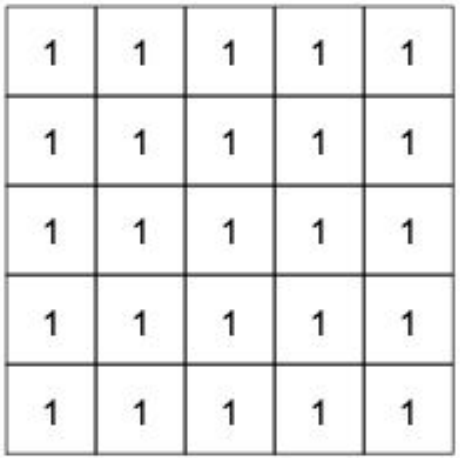
False pattern: artifact of the aggregation scheme

⚠️ WARNING ⚠️

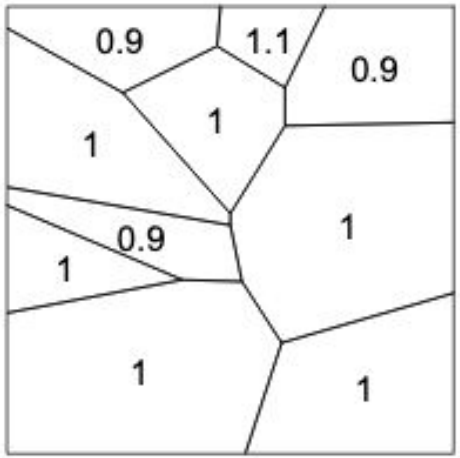
Never map counts aggregated across polygons of irregular size and shape!

Modifiable Areal Unit Problem

- A quasi-solution to this problem is to represent rates, rather than counts



Uniform grid

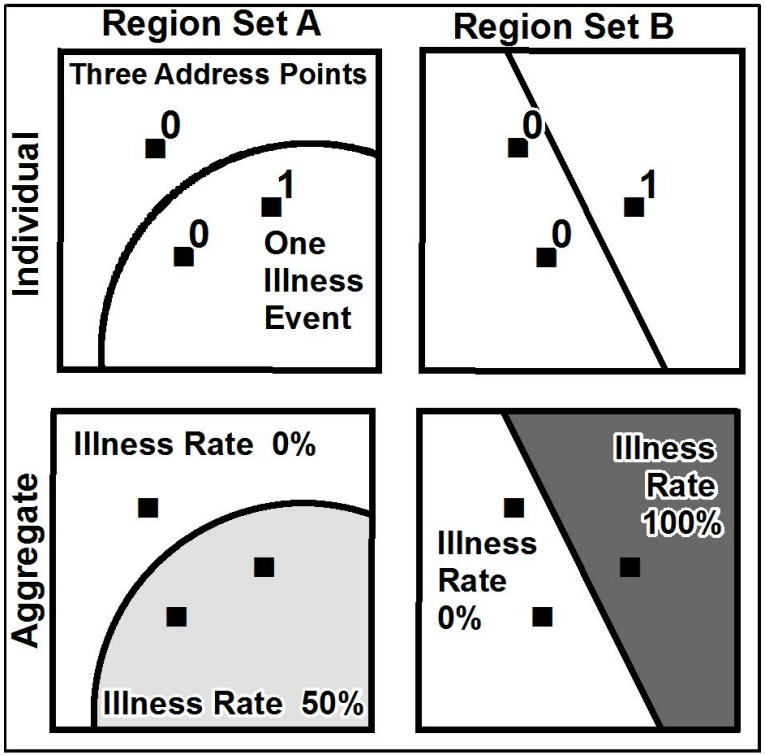


Irregular polygons

The slight discrepancy in values for the map on the right is to be expected given that the zonal boundaries do not split the distance between points exactly

Modifiable Areal Unit Problem

- But the problem is not fully solved by rates, especially when the are small numbers of points in each zonal unit

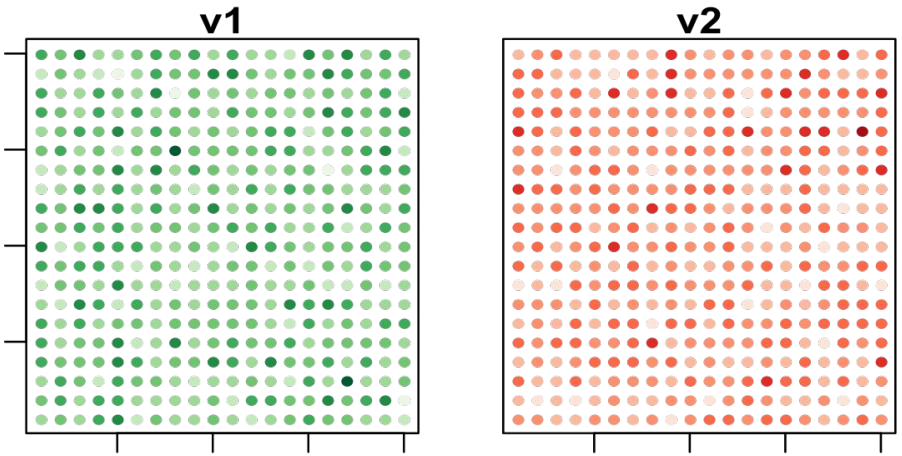


Modifiable Areal Unit Problem

- Different aggregation schemes can result in completely different outcomes
- This problem is often referred to as the modifiable areal unit problem (MAUP) and has some statistical implications as well as visual
- Unfortunately, this problem is often overlooked in many analyses that involve aggregated data

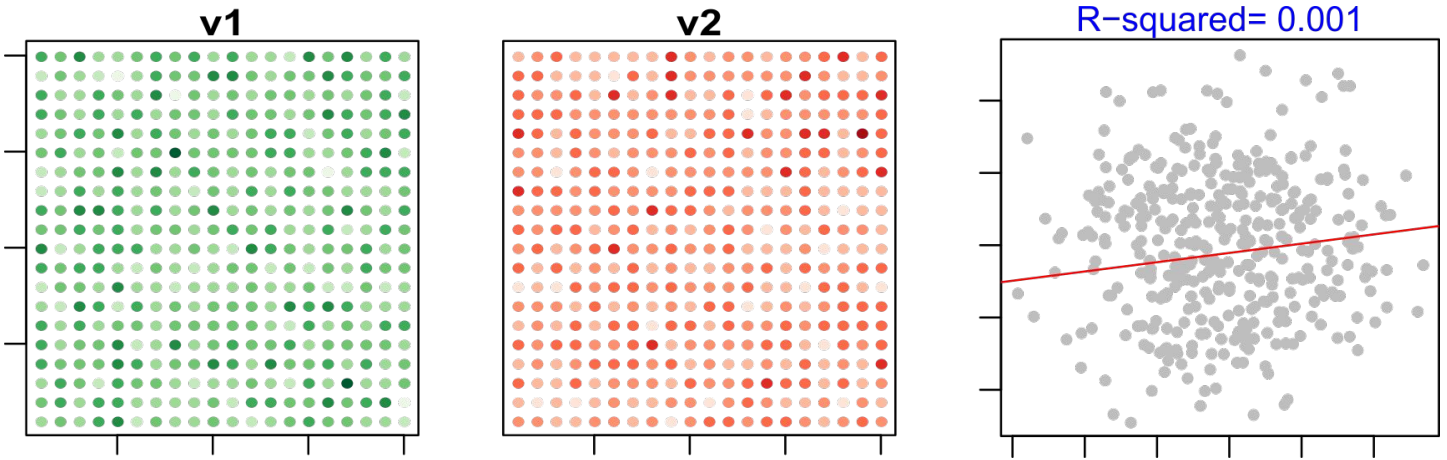
Modifiable Areal Unit Problem

- Let’s say two variables, v1 and v2, are recorded at each point



Modifiable Areal Unit Problem

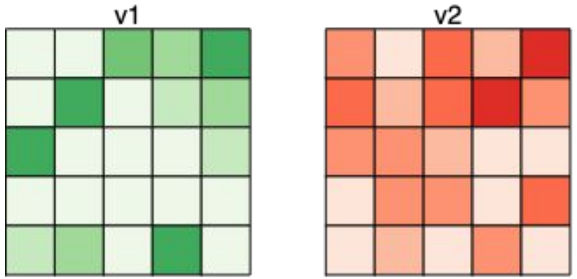
- We might want to know if there is correlation between v1 and v2



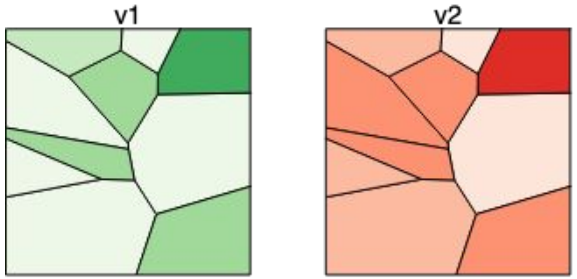
Modifiable Areal Unit Problem

- However, data often comes aggregated by polygons

Data summarized using a **uniform** aggregation scheme



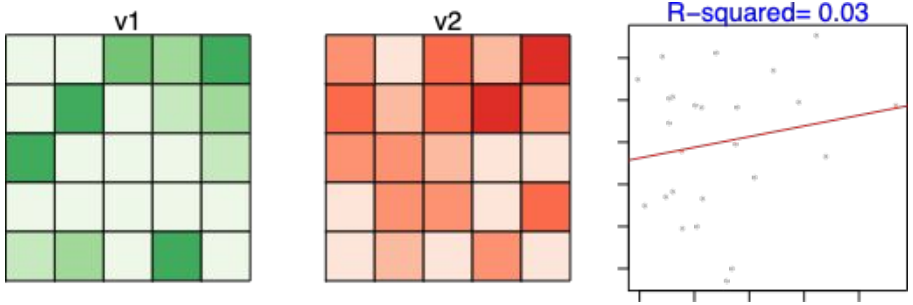
Data summarized using a **non-uniform** aggregation



Modifiable Areal Unit Problem

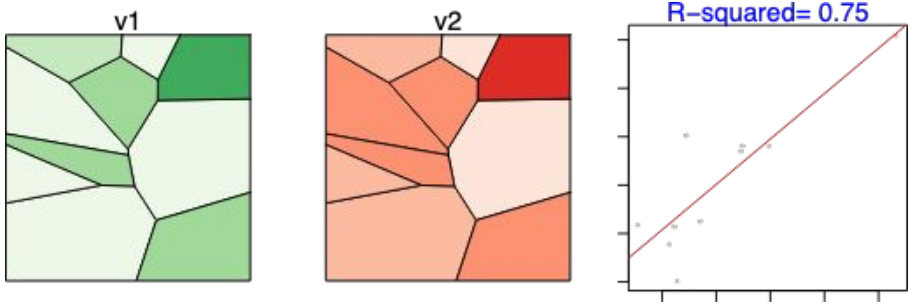
- If we are not careful, this can fool us into thinking a pattern exists when it doesn't

Data summarized using a **uniform** aggregation scheme



Slight increase in slope and R-squared

Data summarized using a **non-uniform** aggregation



Big increase in slope and R-squared!