Intro to Spatial Statistics

January 18, 2022

Pratt

Goals

- Provide motivation for **statistics**
  - Why do we care, why is it worth learning, how is it useful, what can we do with it?
  - Statistics and the scientific method
- Provide motivation for ***spatial* statistics**
  - Why do we need statistics in a spatial context? (Going beyond the map)
  - Why do we need to consider the effects of space in a statistical context? (Spatial is special)

First, an example

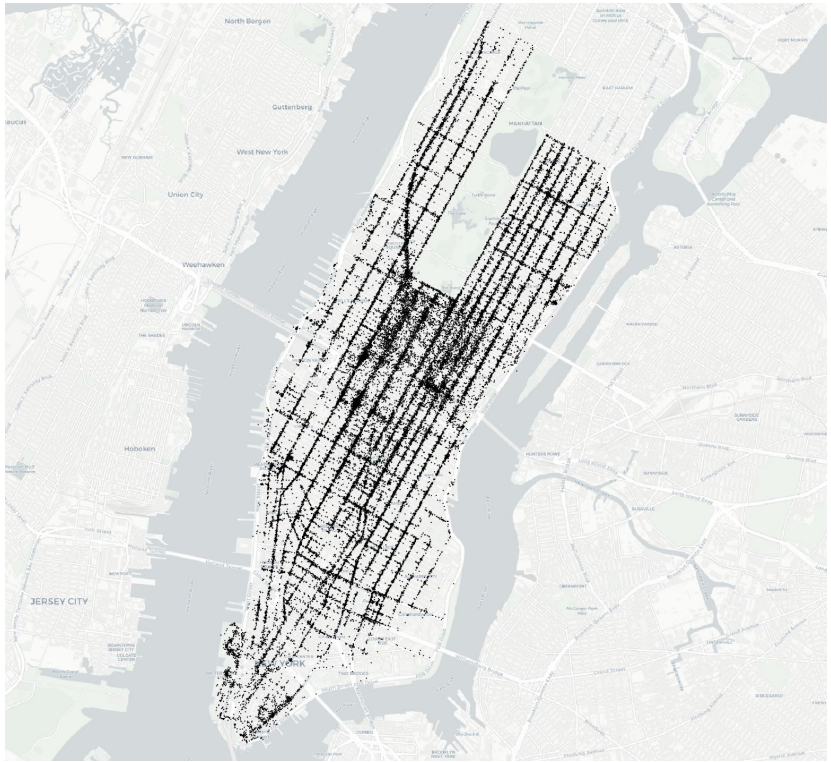
Going beyond the map

Analyzing Yellow Taxi Pickups in NYC

- Let's say we are interested in analyzing the spatial distribution of Yellow Taxi pickups in New York City
- Some initial questions:
  - Are there any clusters of high demand (i.e. "**hot spots**")?
  - **Where** are these hot spots located?
  - **Why** are hot spots located where they are?
  - Are there any **interesting outliers**?

Map the Taxi Pickup Points

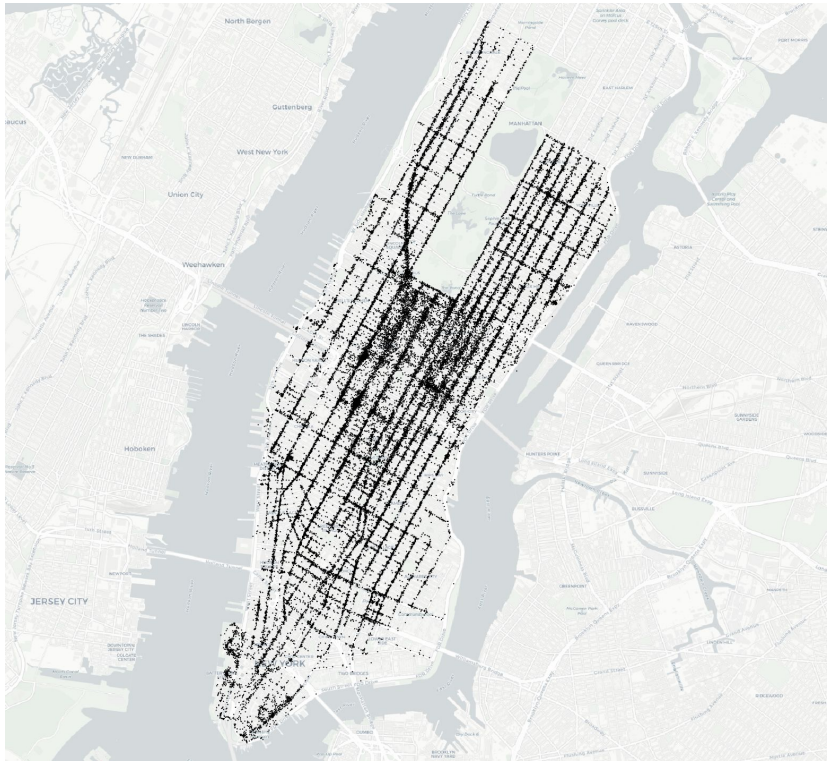
- The first thing we'll do is simply map the points and see what it looks like
- To the right is a sample of 50,000 Yellow Taxi pickup locations from the month of January 2016



Sample of 50,000 Yellow Taxi Pickups, January 2016

Map the Taxi Pickup Points

- This points map is somewhat useful
- We can see some patterns emerging
  - Individual streets stand out
  - Midtown looks like a hotspot
- Some drawbacks
  - No clear takeaways
  - Fails to capture “ambience”
  - Overlapping points might obscure each other
  - Relies entirely on human eye to detect patterns

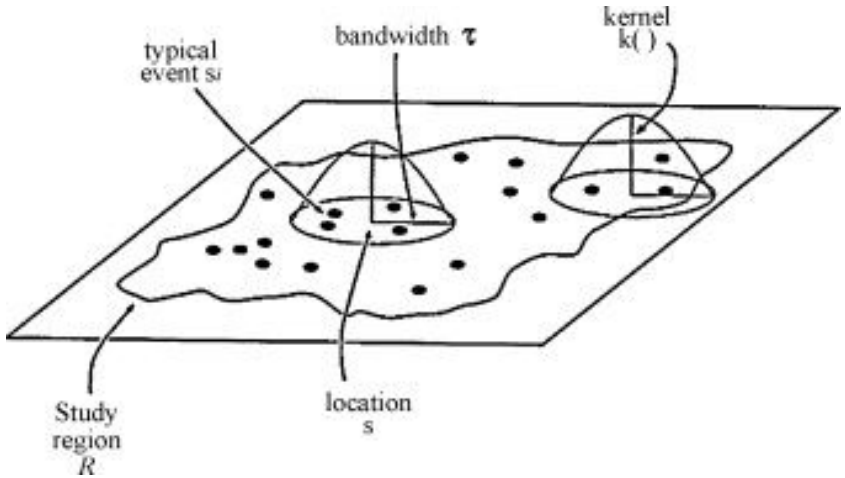


Sample of 50,000 Yellow Taxi Pickups, January 2016

Create a heatmap using Kernel Density Estimation (KDE)

- KDE calculates the density of features in a neighborhood around the features
- Conceptually, a smoothly curved surface is fitted over each point
- KDE requires us to choose:
    - Kernel (smoothing function)
    - Bandwidth (radius of kernel)

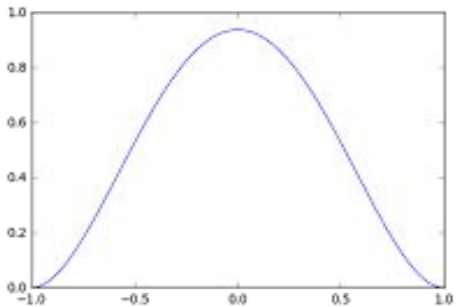
Kernel Density Estimation



Make a smoothed heatmap from the points

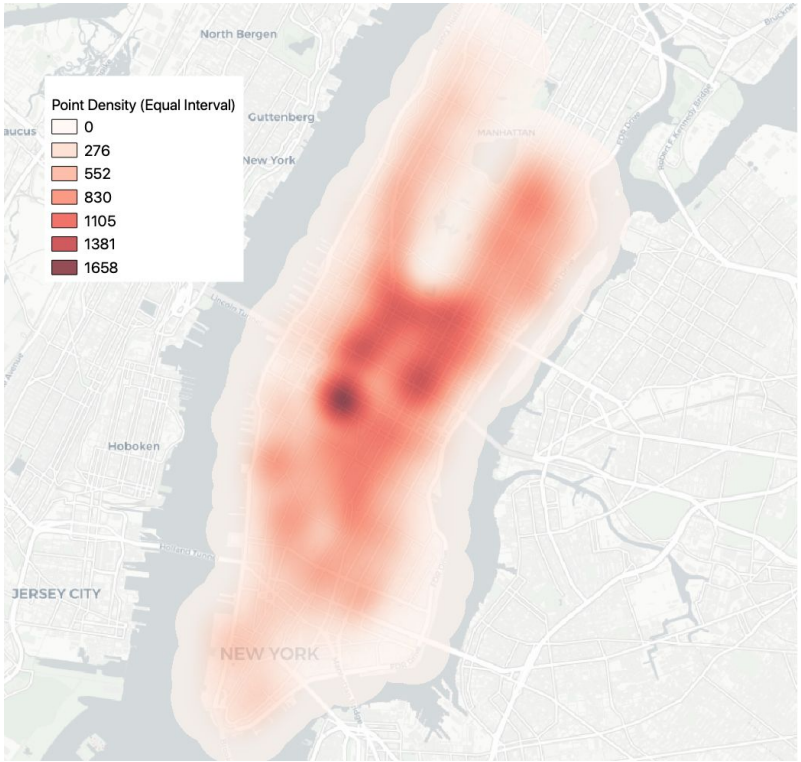
- Heatmaps are a great visualization method to display the density of point events
- The map to the right uses a Quartic kernel (typically the default option) and a bandwidth of 2,000 feet

Quartic kernel



Bandwidth

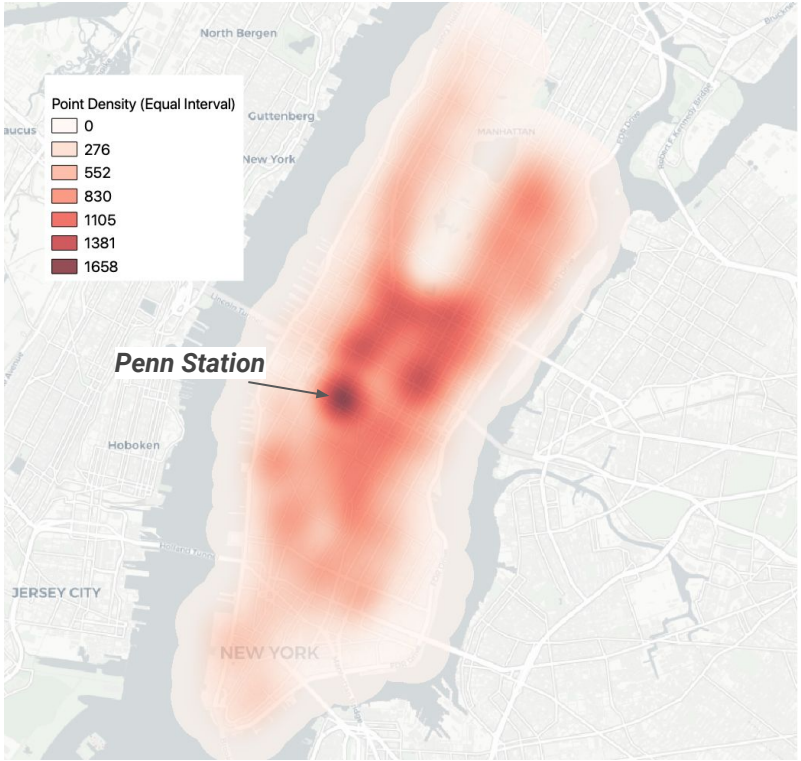
2,000 Feet



Heatmap with quartic kernel and bandwidth (radius) = 2,000 feet

Make a smoothed heatmap from the points

- This heatmap is pretty useful
- Gives a good general sense that hotspots do in fact exist
- Visually suggests where the hotspots might be located
- Penn Station stands out as a primary hot spot



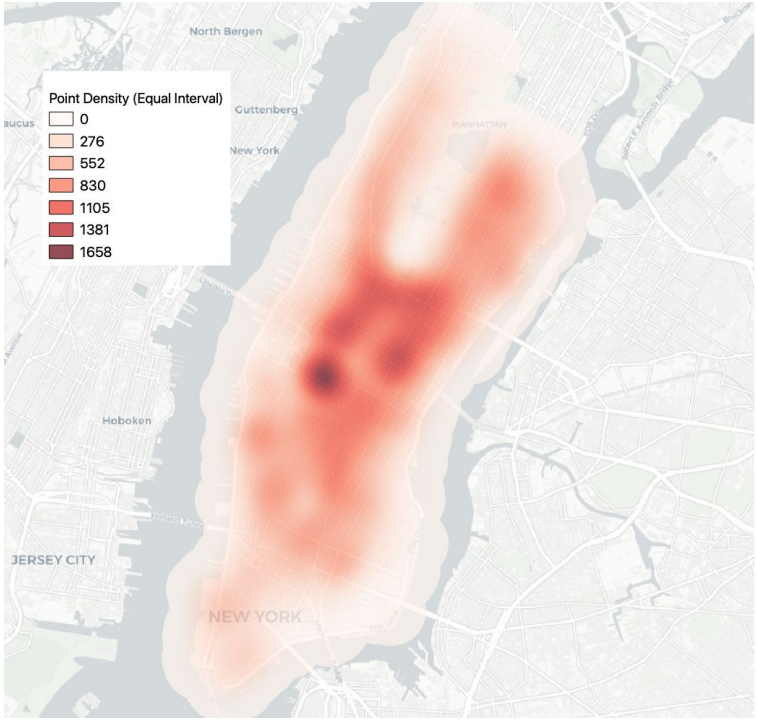
Heatmap with quartic kernel and bandwidth (radius) = 2,000 feet

Problems with Heatmaps

- Heatmaps rely solely on the **human eye** for pattern detection
    - The human eye is very good at detecting patterns, even when they are not there
    - Allows room for subjective judgement and non-reproducibility
- Heatmaps require several **subjective choices** by the map-maker which can significantly impact the overall impression of the map
    - How to choose the **color scheme**?
    - How to choose the **bandwidth radius**?
    - How to choose the **kernel function**?

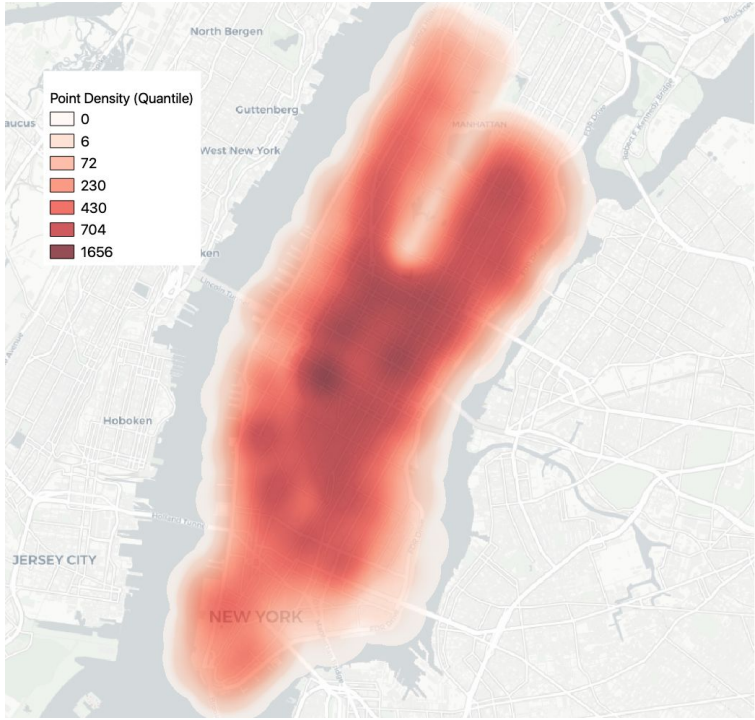
Different Color Schemes Yield Different Patterns

Equal Interval Classification



*Equal interval classification divides a set of values into equal sized groups. The number of observations **may differ** in each class.*

Quantile Classification

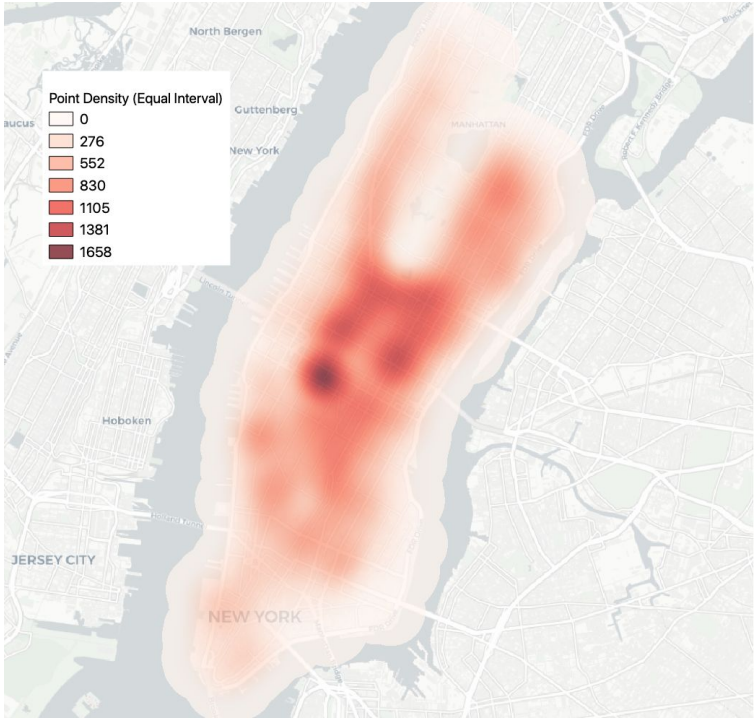


*Quantile classification distributes a set of values into groups that contain an **equal number** of observations.*

Different Color Schemes Yield Different Patterns

- This equal interval classification makes it look like there is significant variability in the point density
- Hot spots stand out very starkly
- Cold spots also stand out

Equal Interval Classification

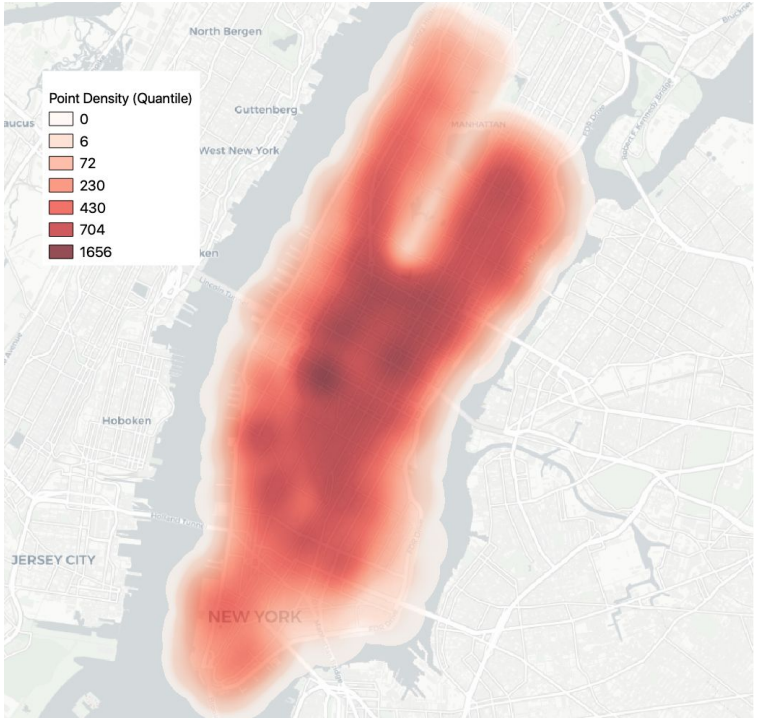


*Equal interval classification divides a set of values into equal sized groups. The number of observations **may differ** in each class.*

Different Color Schemes Yield Different Patterns

- This quantile classification makes it look like there is less variability in the point density
- The whole map seems to have a more uniform distribution
- Hot spots do not stand out as well

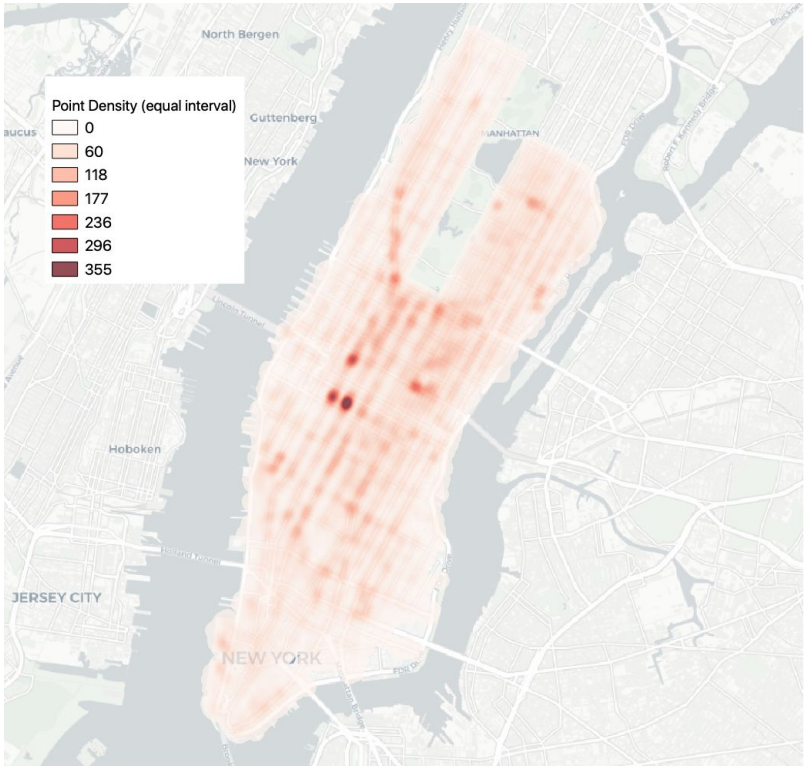
Quantile Classification



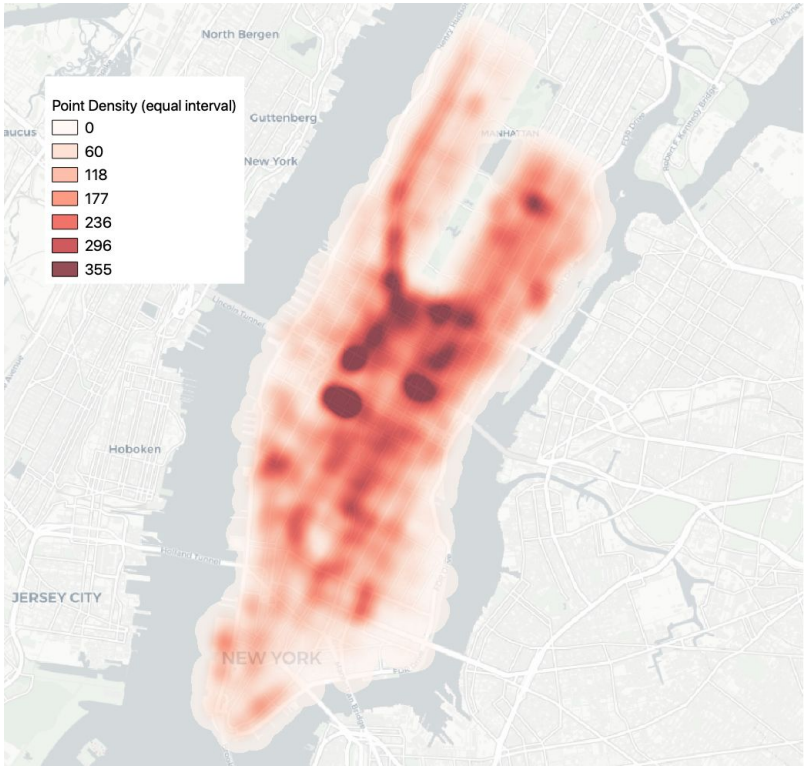
*Quantile classification distributes a set of values into groups that contain an **equal number** of observations.*

Different Bandwidths Yield Different Patterns

Bandwidth = 500 feet

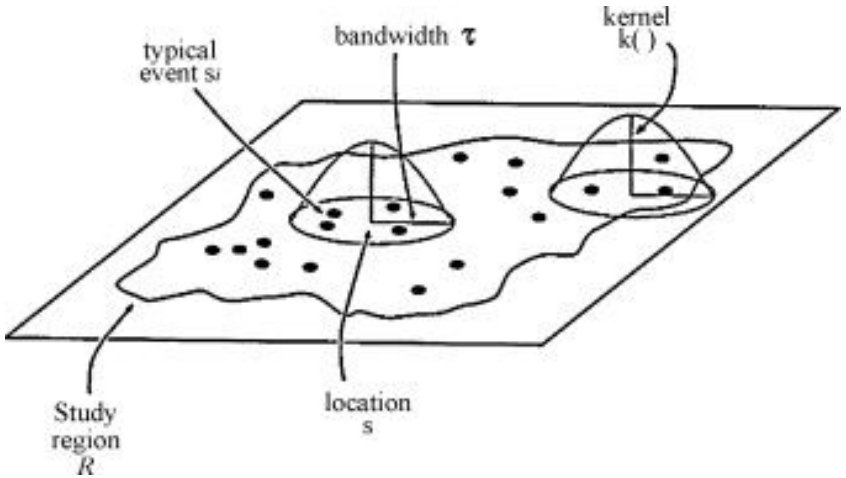


Bandwidth = 1,000 feet

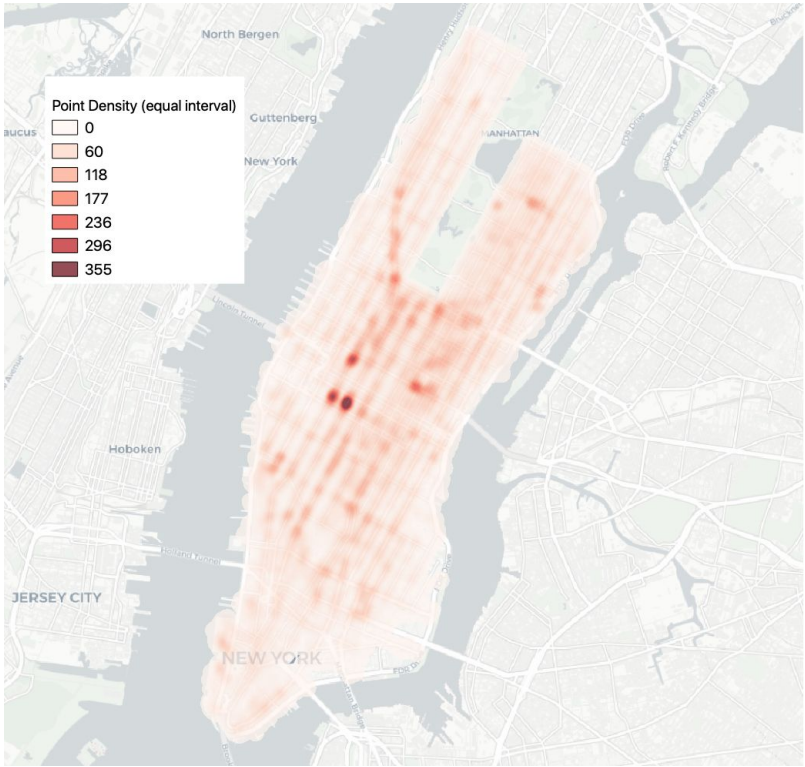


Different Bandwidths Yield Different Patterns

- Bandwidth is the radius of the smoothing function

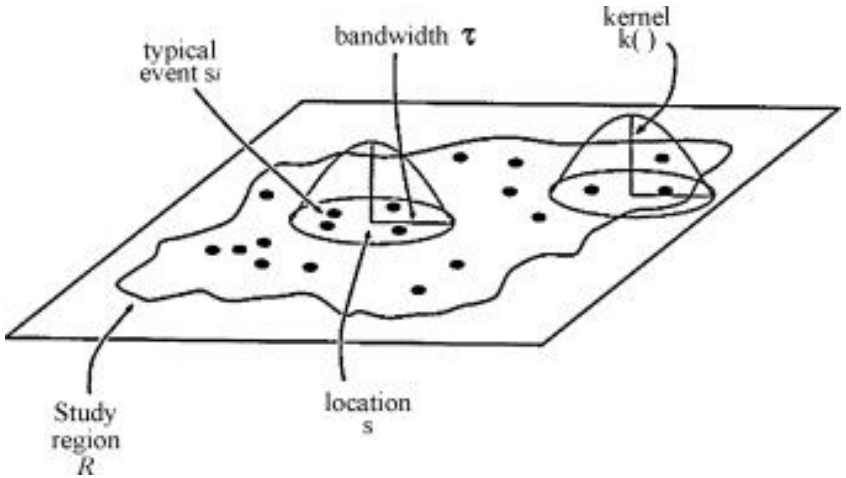


Bandwidth = 500 feet

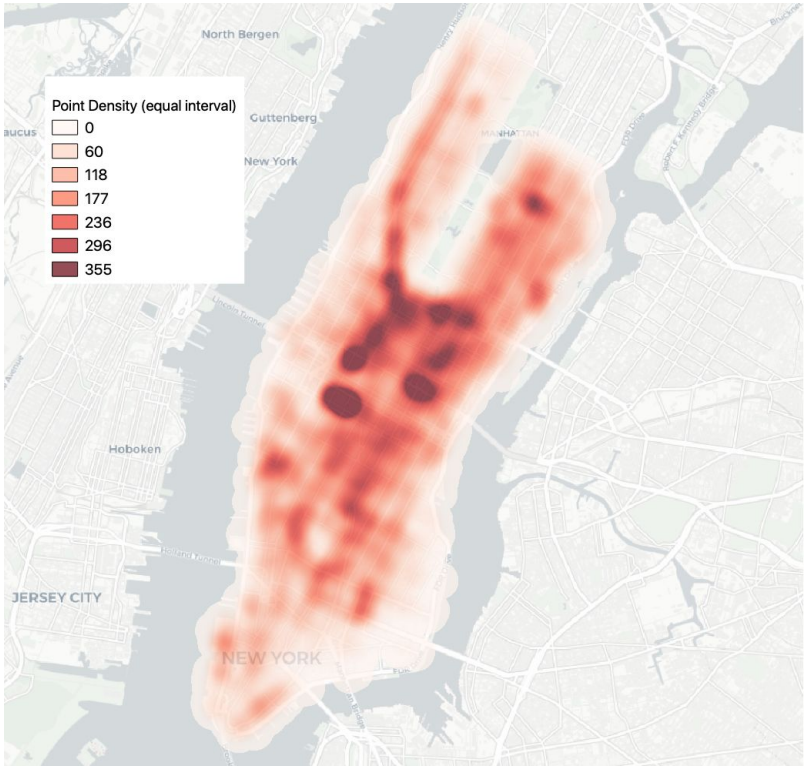


Different Bandwidths Yield Different Patterns

- Bandwidth is the radius of the smoothing function

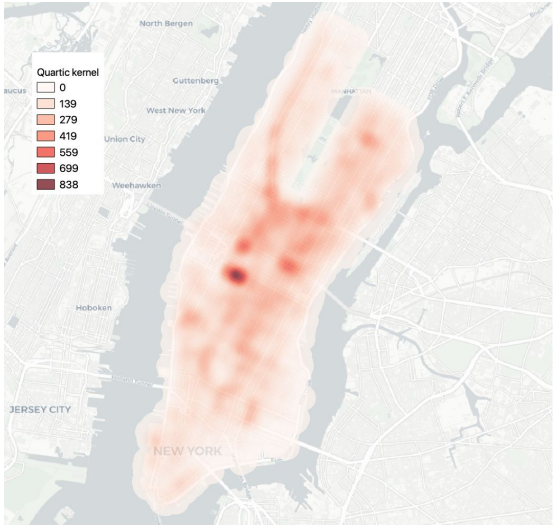


Bandwidth = 1,000 feet



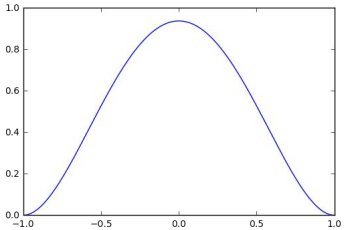
Different Kernel Functions Yield Different Patterns

Quartic Kernel

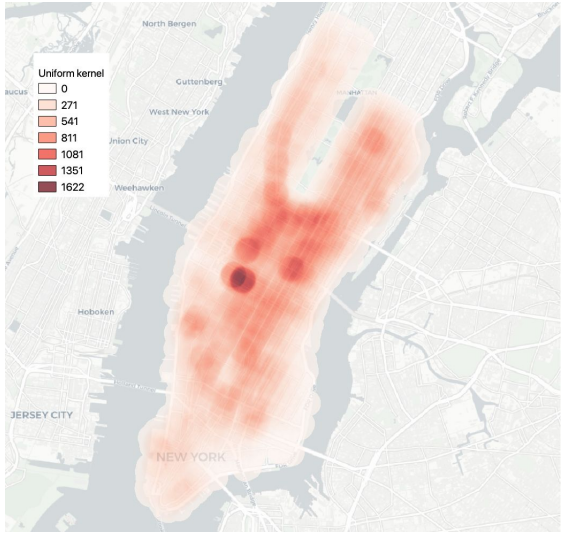


Quartic

$$P(x) = \frac{15}{16}(1 - d^2)^2$$

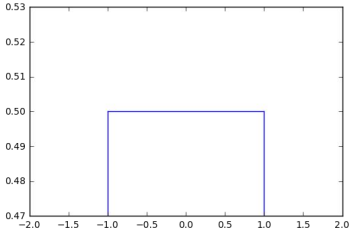


Uniform Kernel

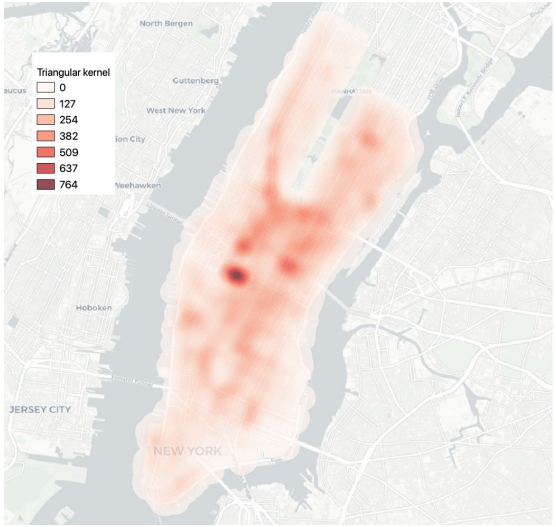


Uniform

$$P(x) = \frac{1}{2}$$

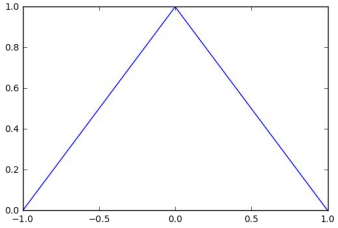


Triangular Kernel



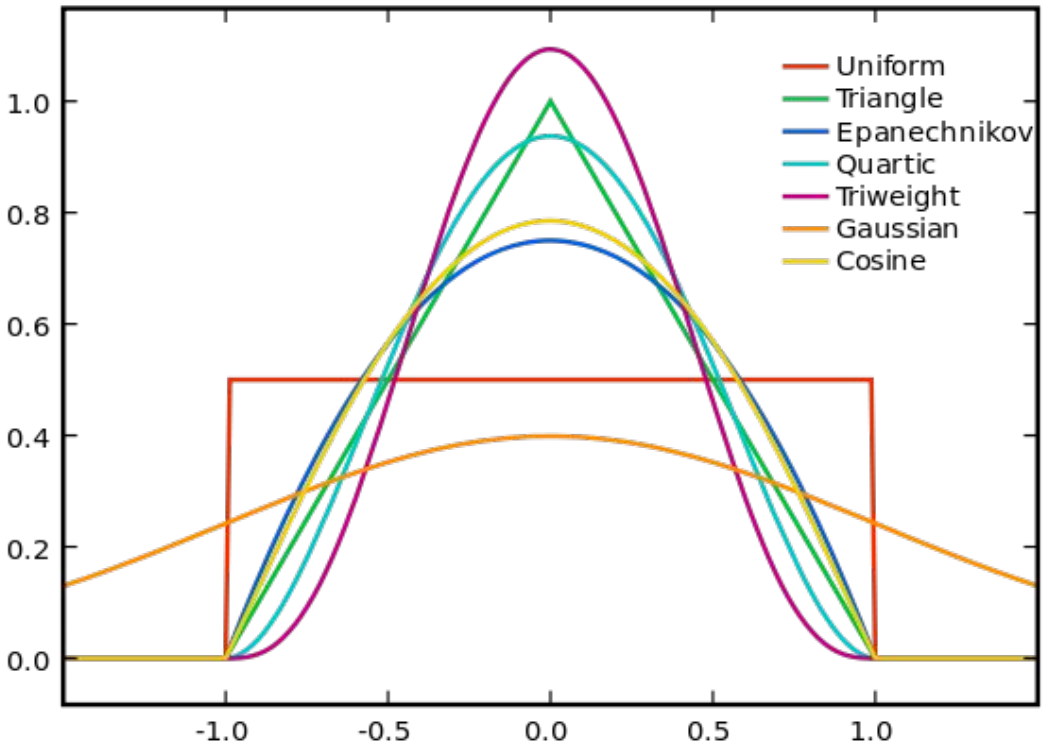
Triangular

$$P(x) = (1 - |d|)$$



Kernel Shapes

- There are many kernel functions to choose from



Recap

- We made a KDE heatmap with the goal of visually analyzing patterns in the dataset
- We were required to make several subjective decisions along the way, each of which could have a significant impact on the map and resulting takeaways
- We can use our eye to look for patterns in the map, but this process is subjective, prone to errors, and not reproducible (different eyes can see different patterns)
- How to assess if patterns are actually there, beyond a reasonable doubt?

Adding value with spatial statistics

- Spatial statistics can help!
- We can formulate hypotheses and test them
- We can make use of decades of theory and methods
- We can come to more definitive and defensible conclusions
- We can ensure our analysis is reproducible by others
- Subjectivity is never completely eliminated, but we can minimize the potential for subjectivity (on the part of both the map-maker and the viewer) to bias the results

Spatial Autocorrelation

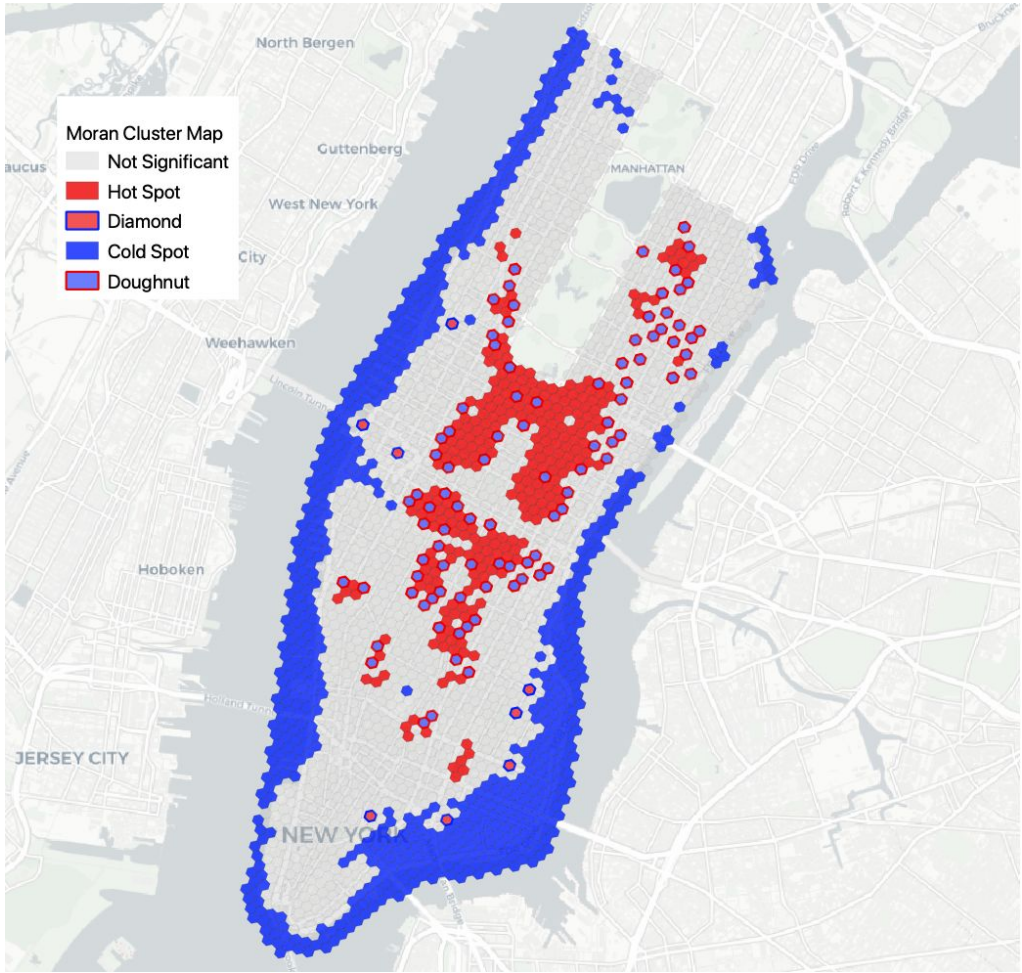
- Spatial autocorrelation refers to the degree to which an object is similar to or different from other nearby objects (i.e. “spatial lag”)
    - Auto = “self”
    - Correlation = “mutually related”
    - Autocorrelation = “mutually related to itself”
- We can perform formal mathematical tests for the presence of autocorrelation
- We will study spatial autocorrelation in depth later in the semester

Hot Spot Analysis

- In a formal hot spot analysis, we test for the presence of spatial autocorrelation to identify and characterize clusters and outliers
    - Spatial **clusters** (nearby values are similar)
        - Hot spots = clusters of high values
        - Cold spots = clusters of low values
    - Spatial **outliers** (nearby values are different)
        - Diamonds = high values surrounded by low values
        - Doughnuts = low values surrounded by high values

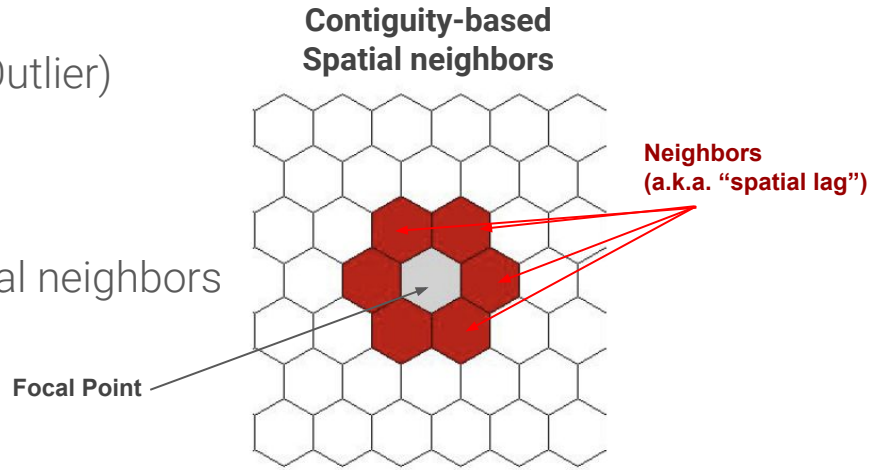
Moran Cluster Map

- Uses a statistic called **Moran's I** to categorize objects as either:
  - Clusters (hot vs. cold)
  - Outliers (diamond vs. doughnuts)
  - Neither (not statistically significant)
- The clusters and outliers that result from this process are ***statistically significant***, meaning that it is highly unlikely that they came about by random chance



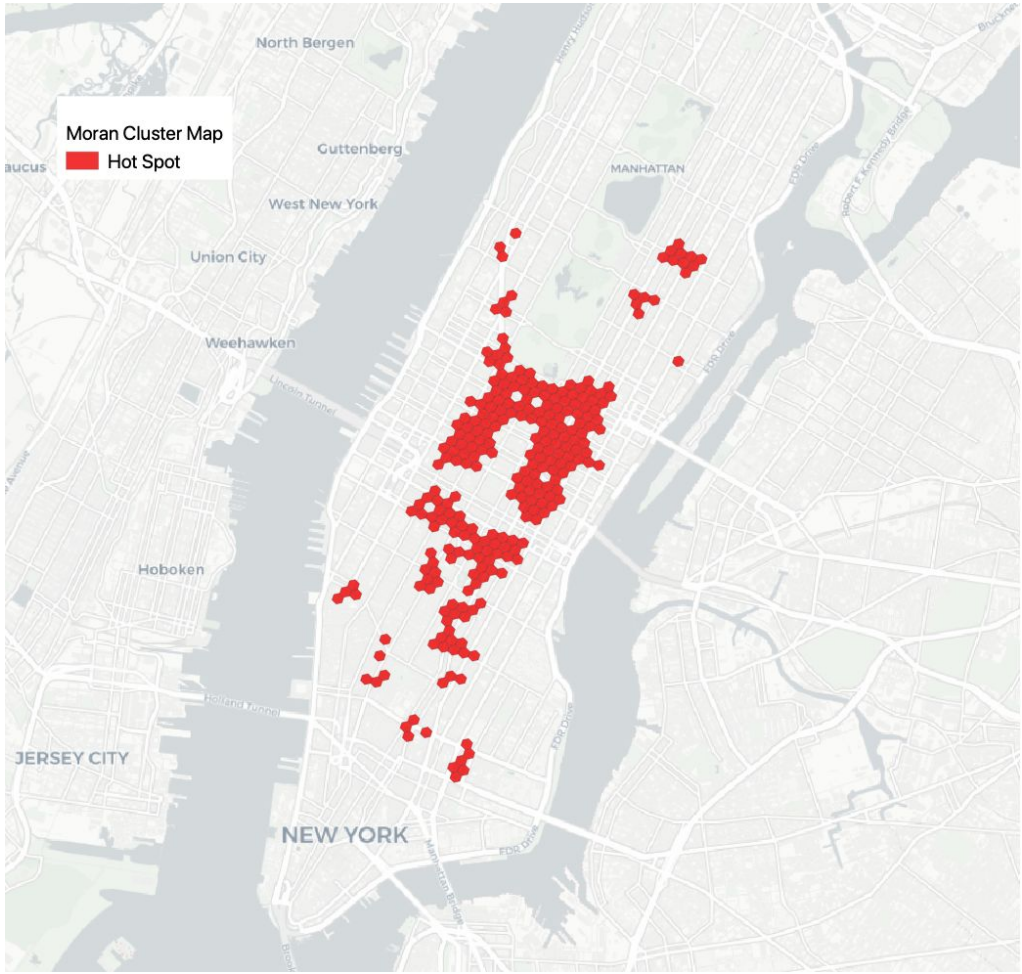
Moran's I

- We will go into much more detail on Moran's I later in the semester
- The basic idea is to compare a value to the average among its spatial neighbors
    - Is it very similar to its neighbors? (Cluster)
    - Is it very different from its neighbors (Outlier)
- There are many ways to define neighbors
- In this example, I use contiguity based spatial neighbors



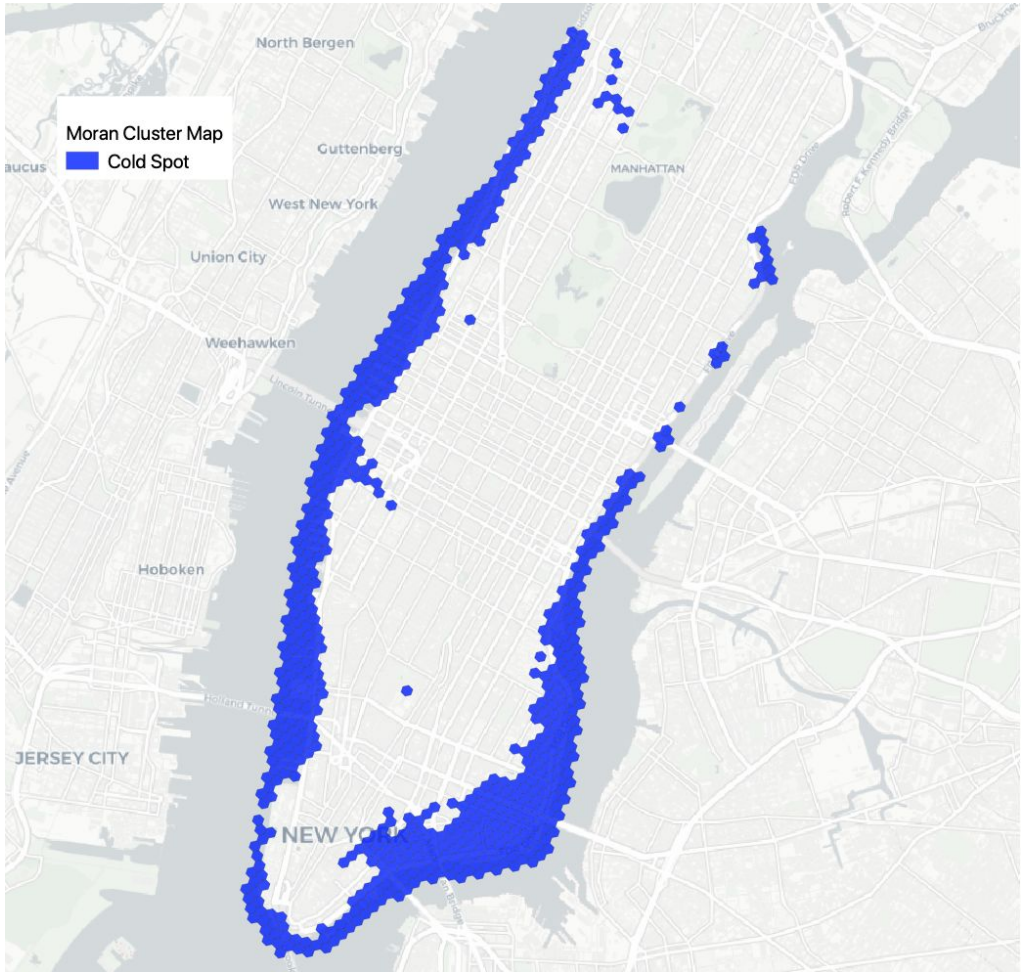
Hot Spots

- **High** values near other **High** values (HH)
- Statistically significant clusters



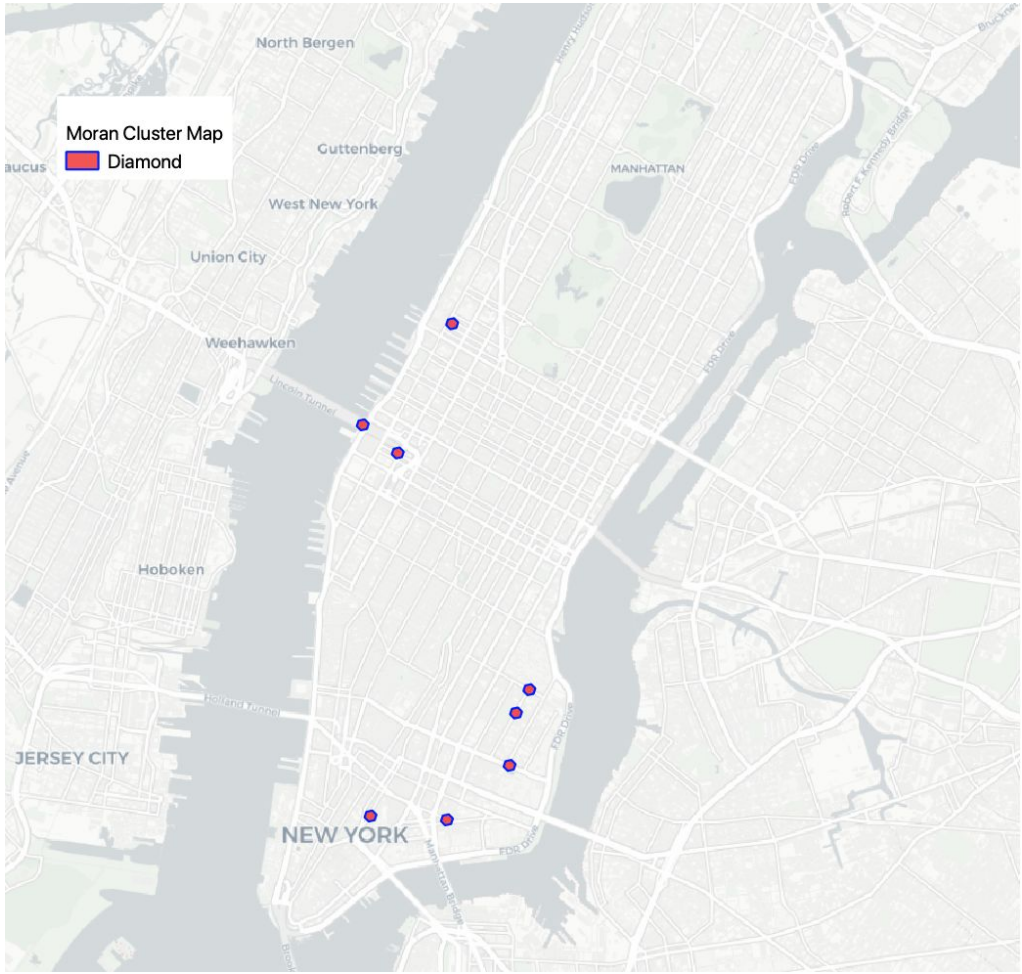
Cold Spots

- **Low** values near other **Low** values (LL)
- Statistically significant clusters



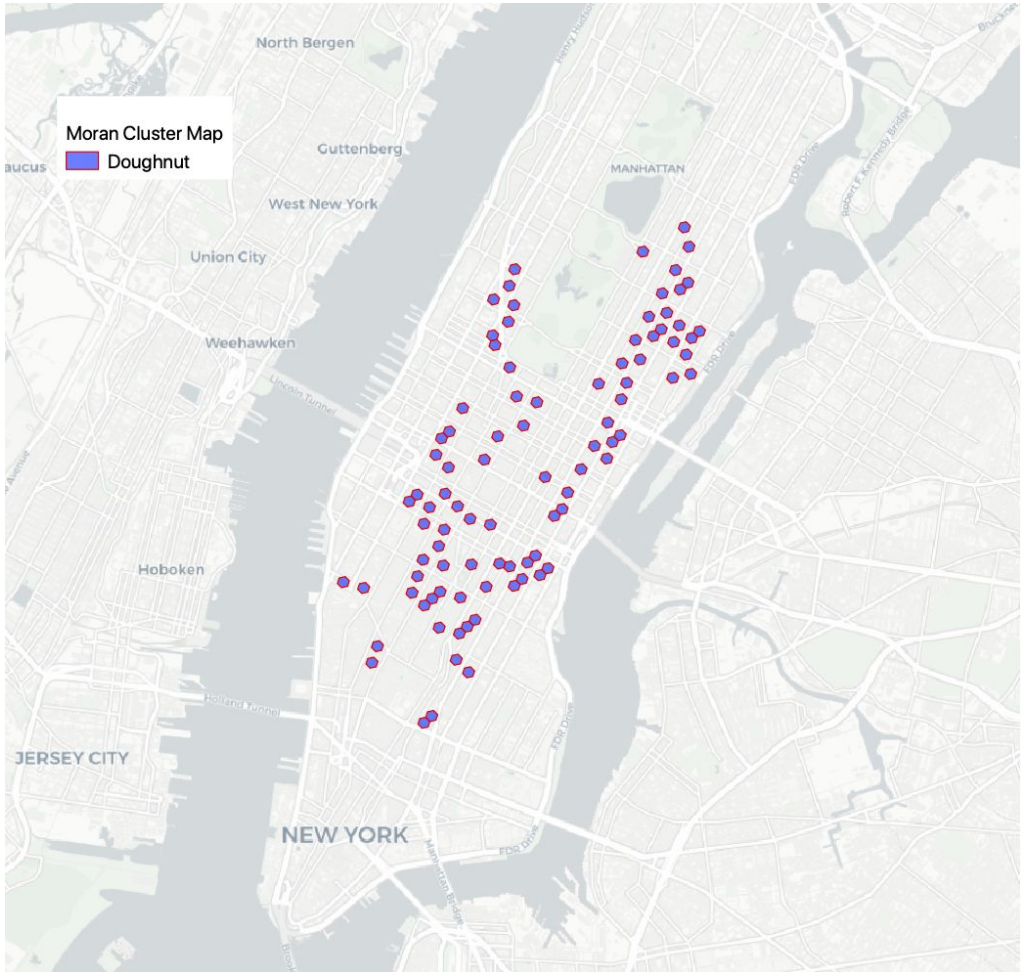
Diamonds

- **High** values near **Low** values (HL)
- Statistically significant outliers



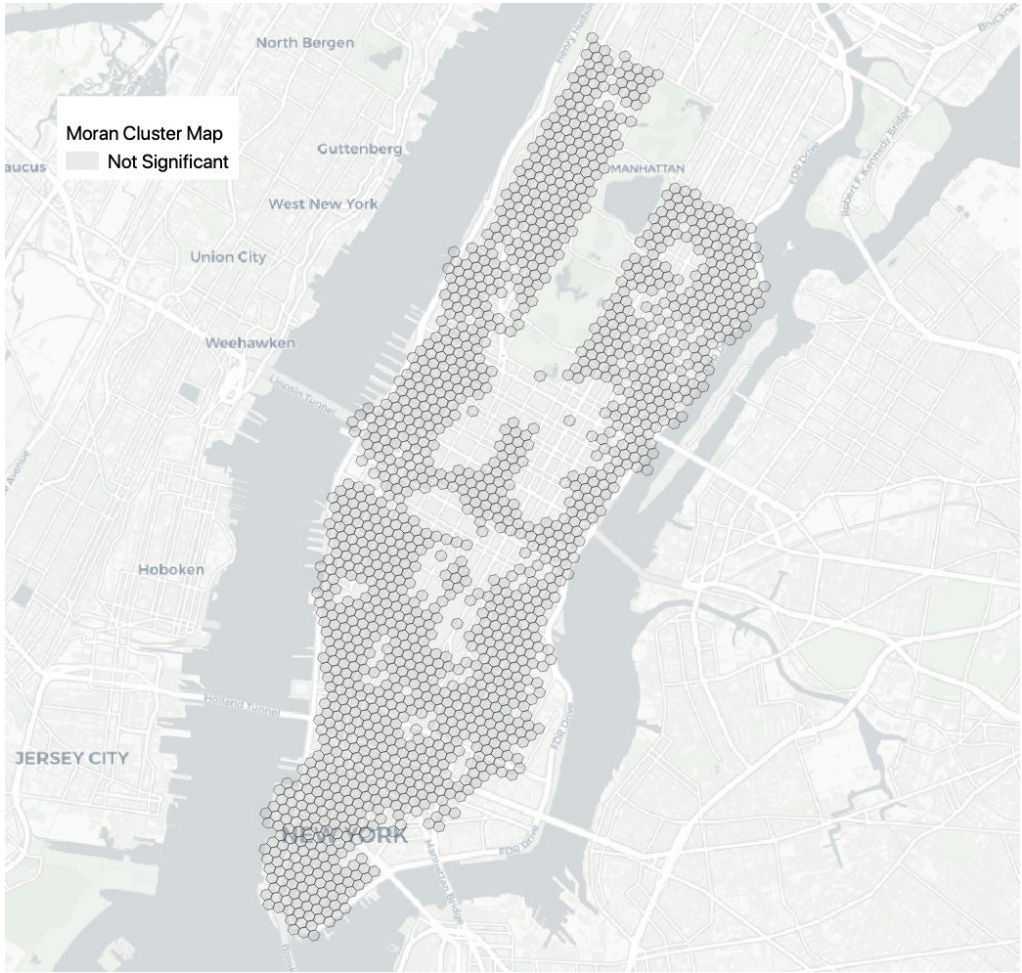
Doughnuts

- **Low** values near **High** values (LH)
- Statistically significant outliers



Not Significant

- Not statistically significant
- The relationship between each feature and its neighbors could theoretically have come about through random chance alone

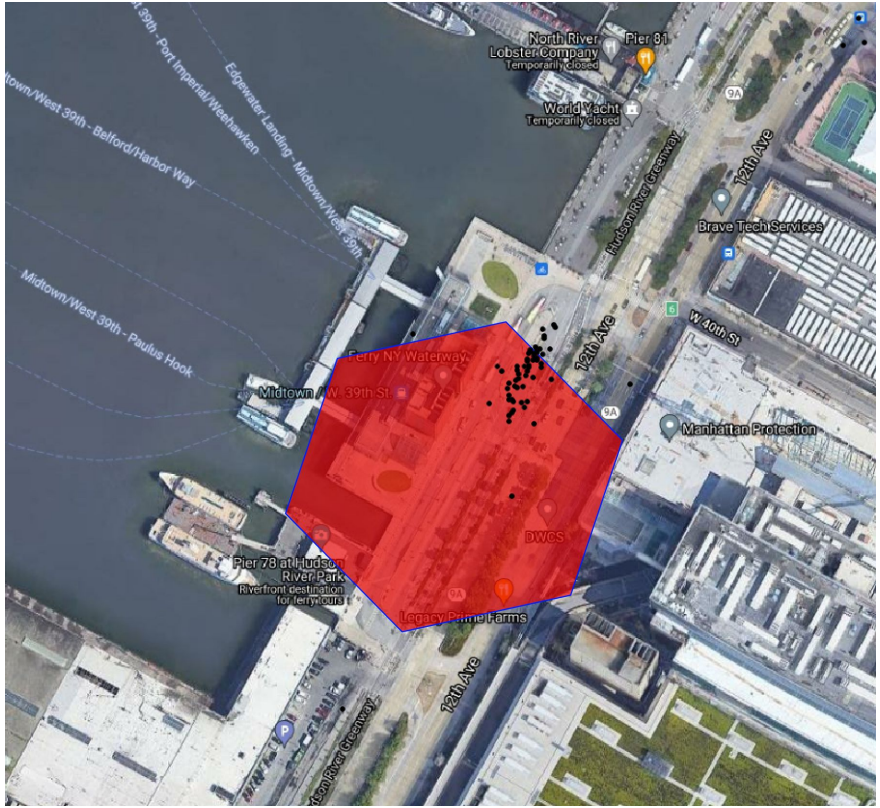


Diamond hunting

West 40th St and 12th Ave

Nearby attractions:

- Ferry Terminal
- Hudson River Park

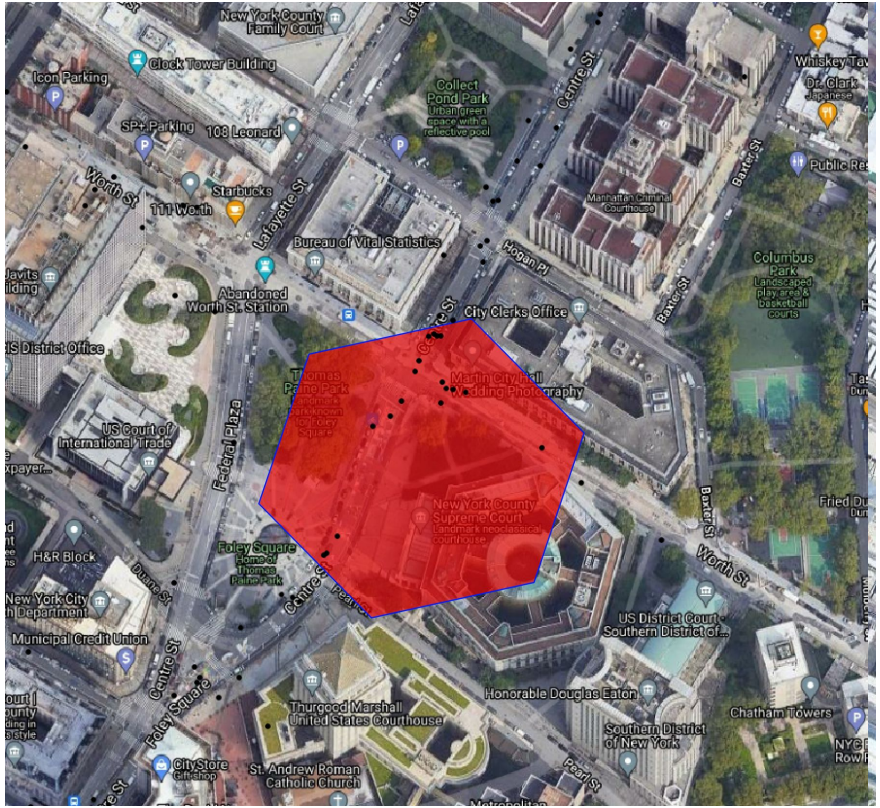


Diamond hunting

Civic Center (Financial District)

Nearby attractions:

- New York County Supreme Court
- Government buildings

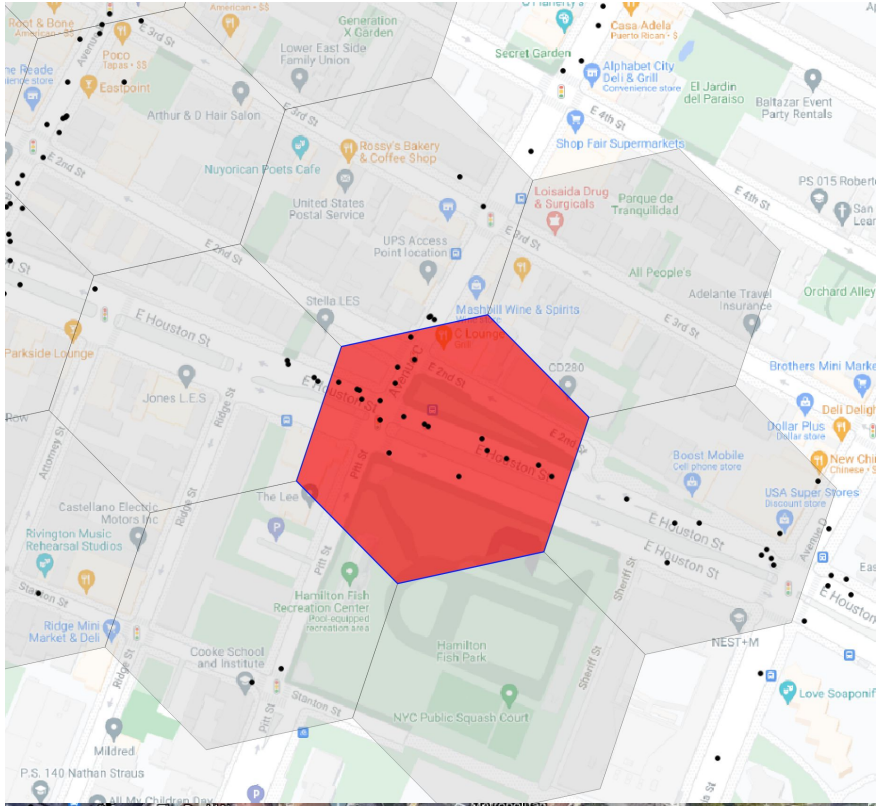


Diamond hunting

Houston St and Avenue C

Nearby attractions:

- Bus stop
- Park and recreation center



Concluding the example

- Maps are great, but visualizing spatial data requires many subjective decisions on behalf of the mapmaker, which can impact conclusions drawn from the map
- The human eye is a strong yet imperfect pattern detector
- Spatial analysis often requires more than simply looking at data on a map
- Spatial statistics can add value, robustness and reproducibility to our analysis

What is Statistics?

Why should we care?

What is Statistics?

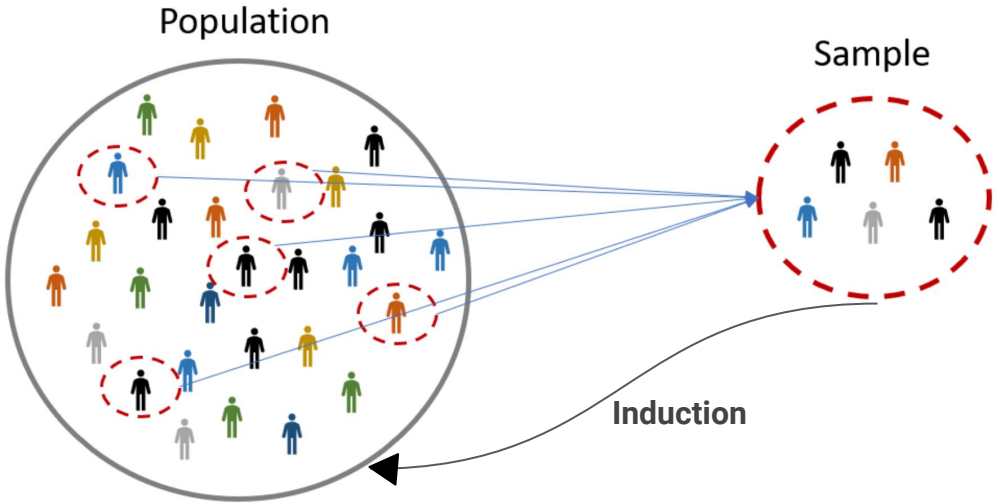
- The science (and art) of learning from data
- Concerned with the collection, analysis, interpretation and communication of empirical data
- Collection of theory and techniques with which the **uncertainty** of *inductive inferences* may be evaluated

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

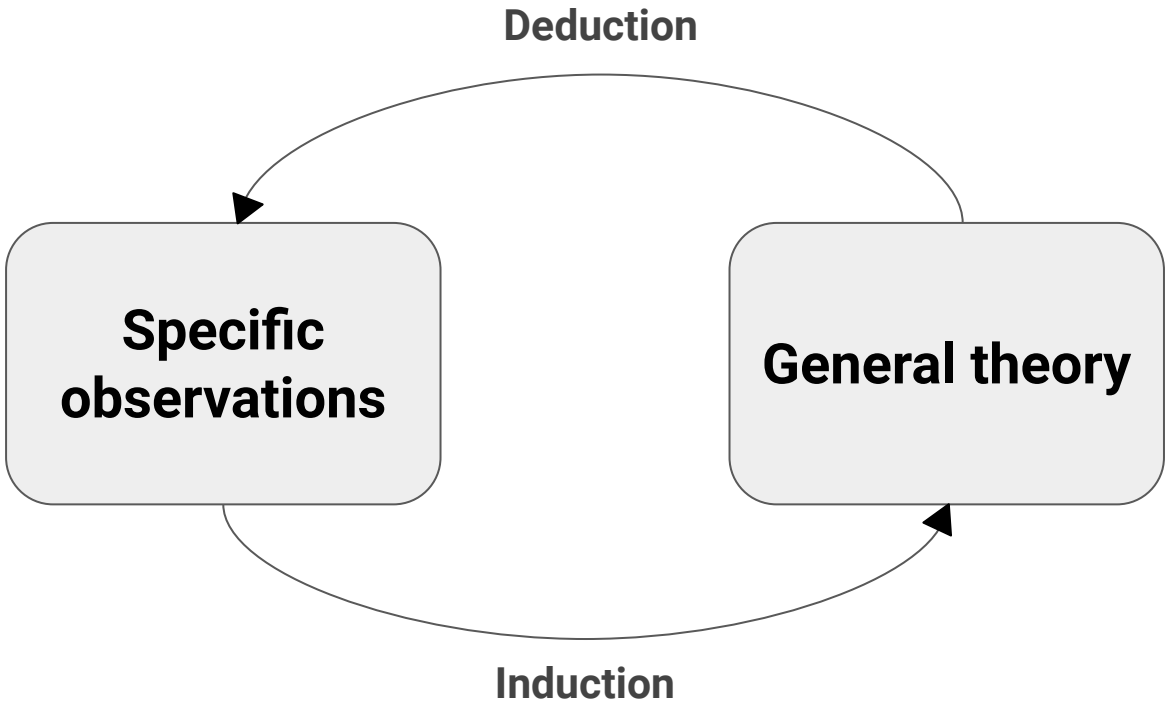
- Example
  - Every raven in a random sample of 1,000 ravens is black
  - This supports the conclusion that all ravens are black
  - Thus, if something is a raven, then it is (probably) black
  - Thus, if something is not black, then it is not a raven

Inductive Inference

≠

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

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

Statistics and Hypothesis Testing

- Hypothesis testing is an essential part of the scientific method
- Statistics are used to determine whether a hypothesis is supported or not supported
    - Hypothesis testing is a type of statistics that determines the probability of a hypothesis being true or false
- The basic idea is: *could these results be due to random chance?*
    - We use statistics to define what randomness means and to compare our definition of randomness to results
- “Statistically significant” results means that we can be reasonably confident that results obtained are not due to random chance

The Scientific Method

Scientific Method

- First outlined by Francis Bacon in *Novum Organum* (1620; “New Instrument”)
    - Bacon thought that prevailing systems of thought relied too much on guesswork and the mere citing of authorities to establish truths of science
- The goal is to offer an objective methodology for scientific experimentation that results in unbiased interpretations of the world and refines knowledge

Theory vs. Hypothesis

- Two key concepts in the scientific approach are theory and hypothesis
- A **theory** is used to make predictions about future observations
- A **hypothesis** is a testable prediction that is arrived at logically from a theory

Basic Principles of the Scientific Method

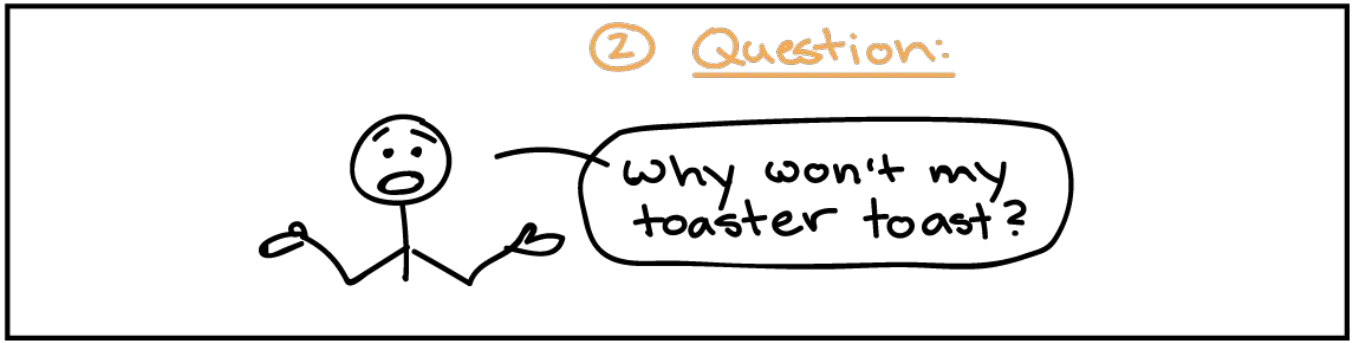
- Reproducibility
  - Results obtained by an experiment should be achieved again when the experiment is replicated by others
- Predictability
  - A theory should enable us to make predictions about future or unknown events
- Falsifiability
  - It must be logically possible that a hypothesis could be shown to be false by observation or experiment
- Fairness
  - All data must be considered when evaluating a hypothesis – the researcher cannot pick and choose what data to keep and what to discard or focus specifically on data that support or do not support a particular hypothesis

Basic Steps of the Scientific Method

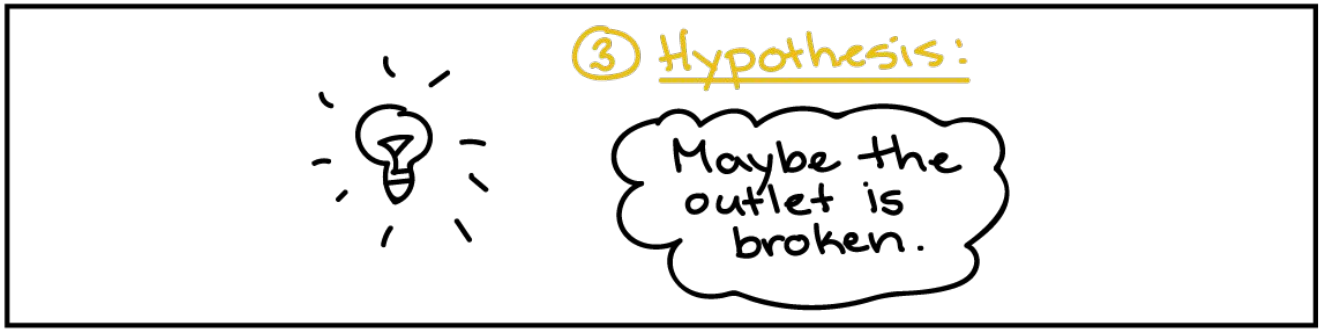
1. Make an observation
2. Ask a question
3. Form a hypothesis
4. Make a prediction based on that hypothesis
5. Test the prediction
6. (Optional) Iterate: use the results to make new hypotheses or predictions

Make an observation

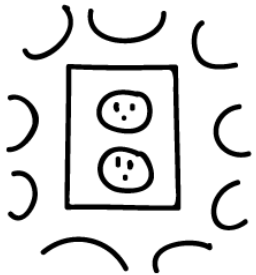
Ask a question



Propose a hypothesis



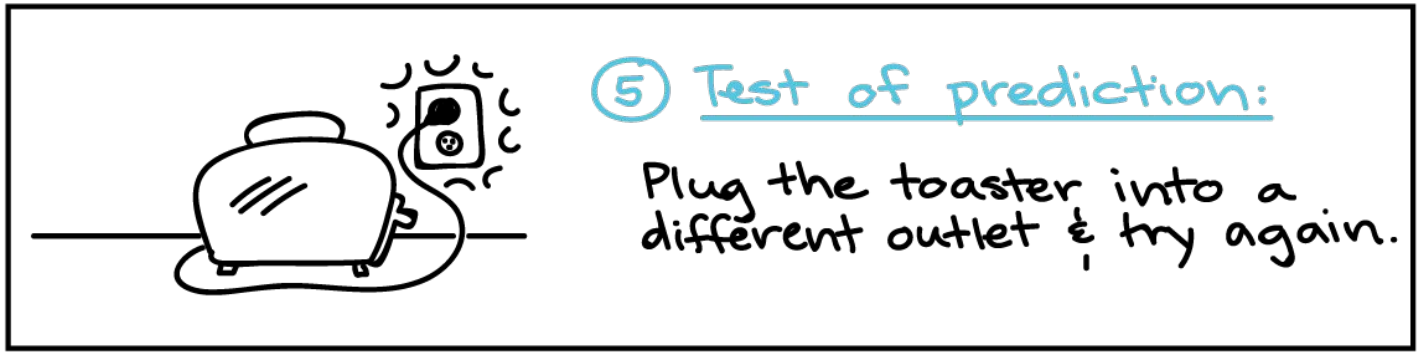
Make a prediction



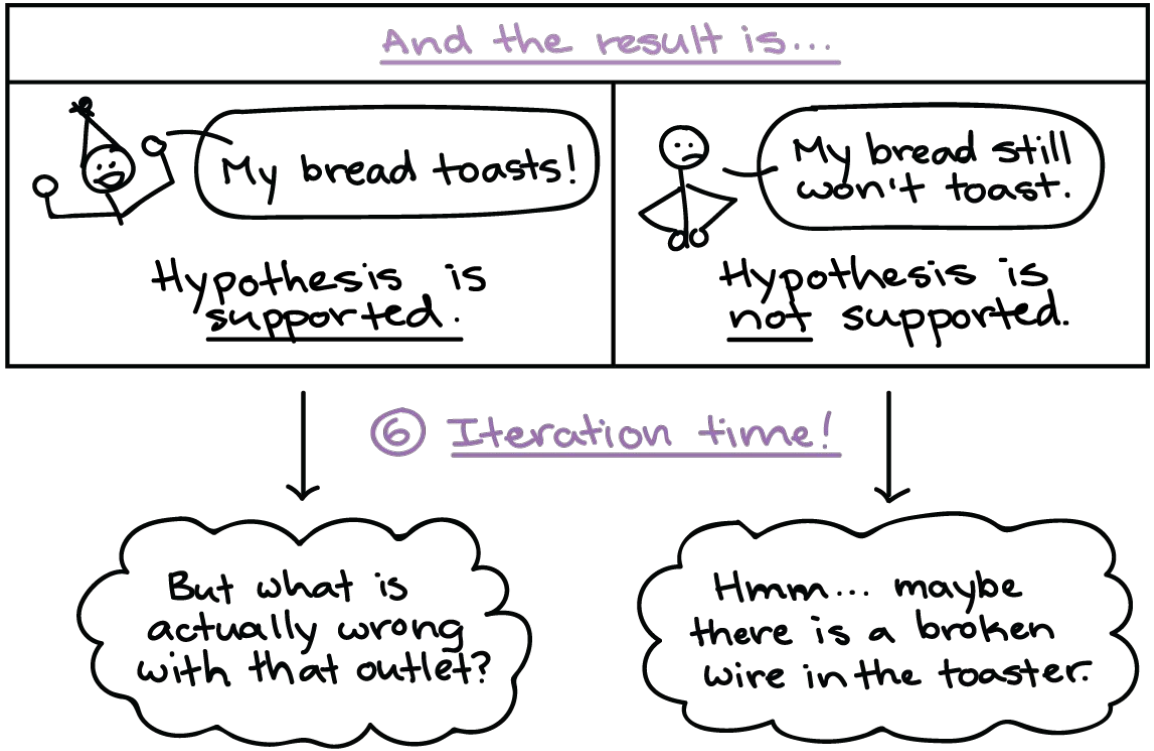
④ Prediction:

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

Test the prediction

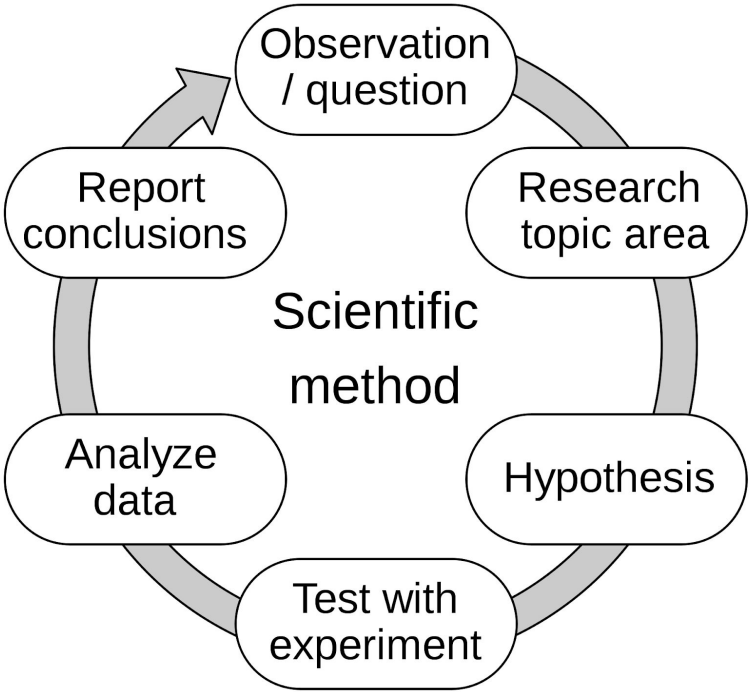


Iterate


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? (interaction)
- How does the context affect what happens? (environment)
- Where should things be located? (optimization)

When is data spatial?

- Non-spatial data = values
- Spatial data = values at **locations**

When is analysis spatial?

- Non-spatial analysis: location does not matter (spatial stationarity)
- Spatial analysis: when the location changes, the information content of the data changes (spatial nonstationarity)

Why is Spatial Special?

Why Spatial is Special

- Spatial data can be seen as both a nuisance and a feature
- As a nuisance, spatial data often complicates statistical tests that assume independent and identically distributed variables
- As a feature, spatial data can enable us to incorporate the effects of space directly into our analysis (with a proper grasp of spatial statistics, that is!)

Why Spatial is Special

- Statistical models are simplified representations of reality
- Things get more complex very quickly when you are forced to grapple with the fact that everything happens somewhere and sometime, and that the effects of space (and time) are not random

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)

Spatial data as bunches of grapes 🍇 🍇 🍇

Data of geographic units are tied together, like bunches of grapes, not separate, like balls in a urn. Of course, mere contiguity in time and space does not of itself indicate independence between units in a relevant variable or attribute, but in dealing with social data, we know that by very virtue of their social character, persons, groups, and their characteristics are interrelated and not independent. Sampling error formulas may yet be developed which are applicable to these data, but until then, the older formulas must be used with great caution. Likewise, other statistical measures must be carefully scrutinized when applied to these data.

Stephan, 1934

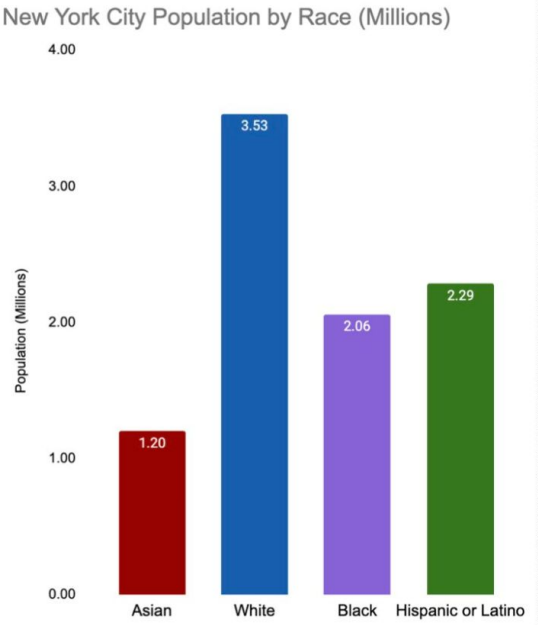
Tobler's First Law of Geography

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

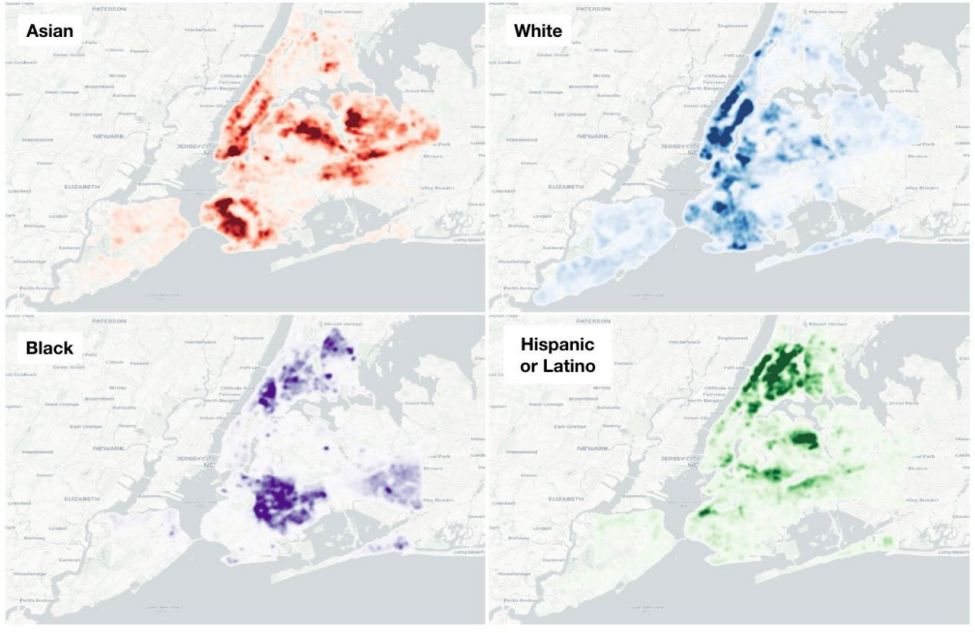
- Waldo Tobler, 1970

Autocorrelation can be hidden when we ignore space

Non-Spatial View



Spatial View



Statistical issues with spatial data

- Spatial data often violates the i.i.d. assumption
  - Spatial autocorrelation
- 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 very 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 statistics vs. Global statistics
  - Local modeling may capture relationships that global models fail to capture

Modifiable Areal Unit Problem

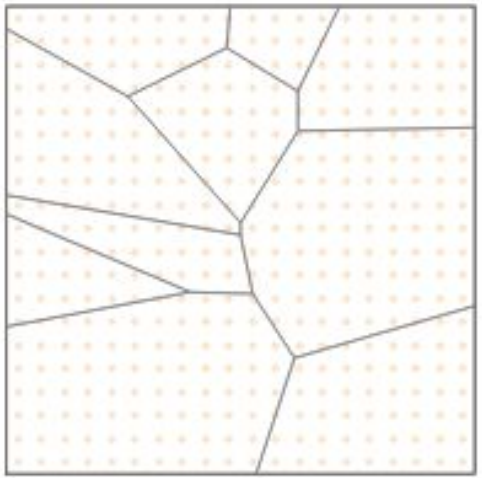
Modifiable Areal Unit Problem

- 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

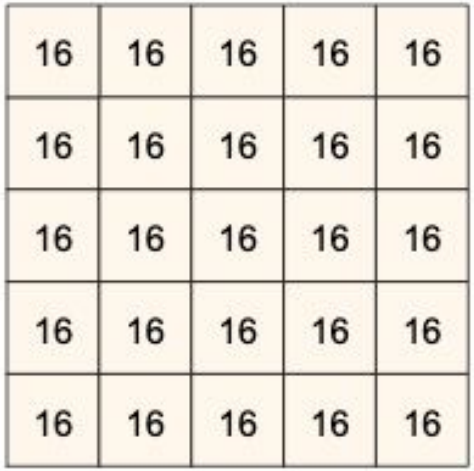
Uniform grid



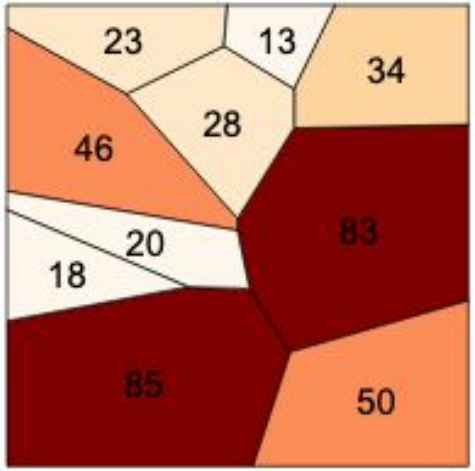
Irregular polygons

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



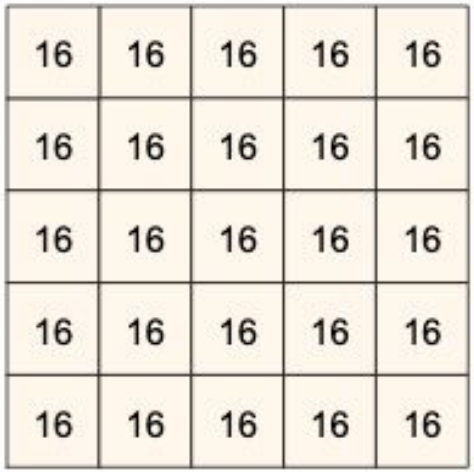
Uniform grid



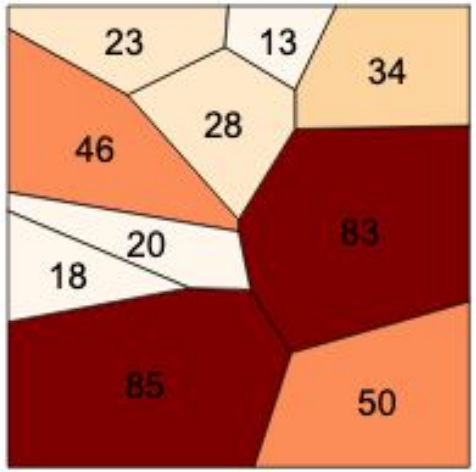
Irregular polygons

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

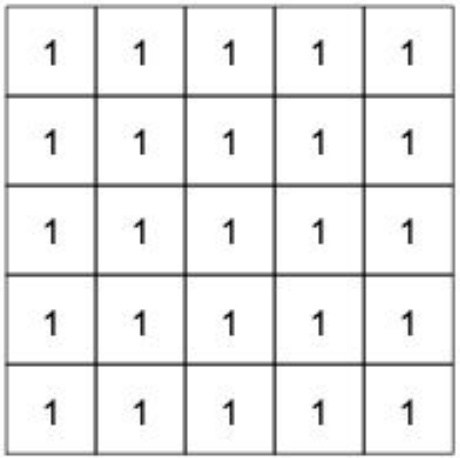
False pattern: artifact of the aggregation scheme

⚠️ WARNING ⚠️

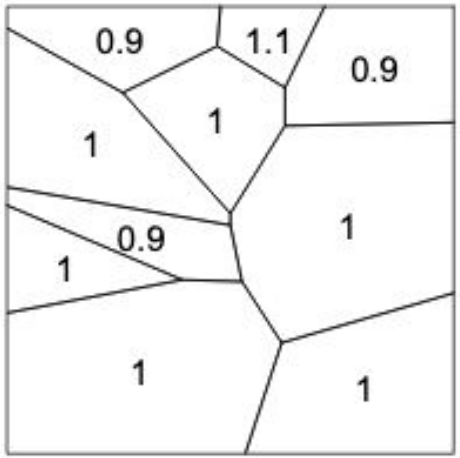
Never map counts aggregated across polygons of irregular size and shape!

Modifiable Areal Unit Problem

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



Uniform grid

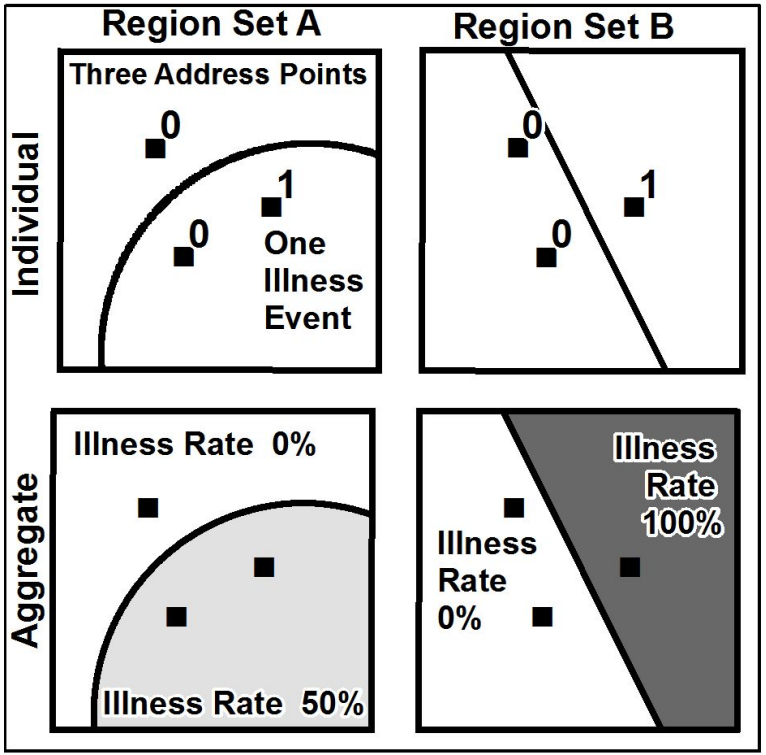


Irregular polygons

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

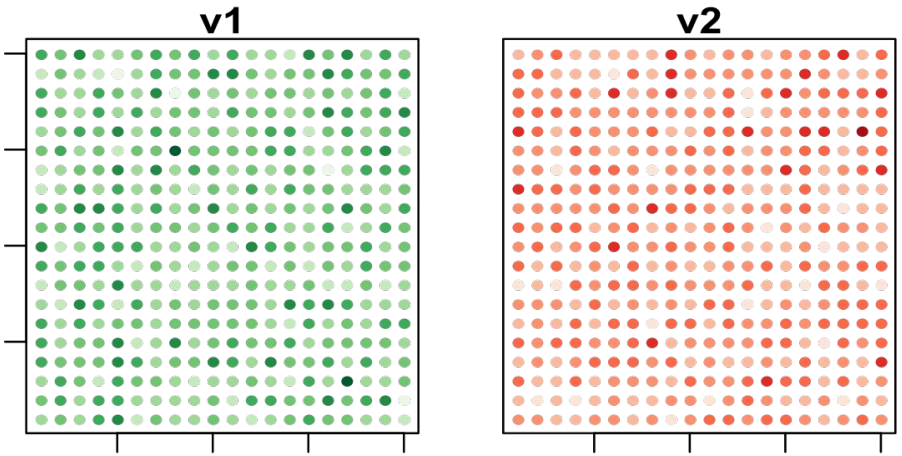


Modifiable Areal Unit Problem

- 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

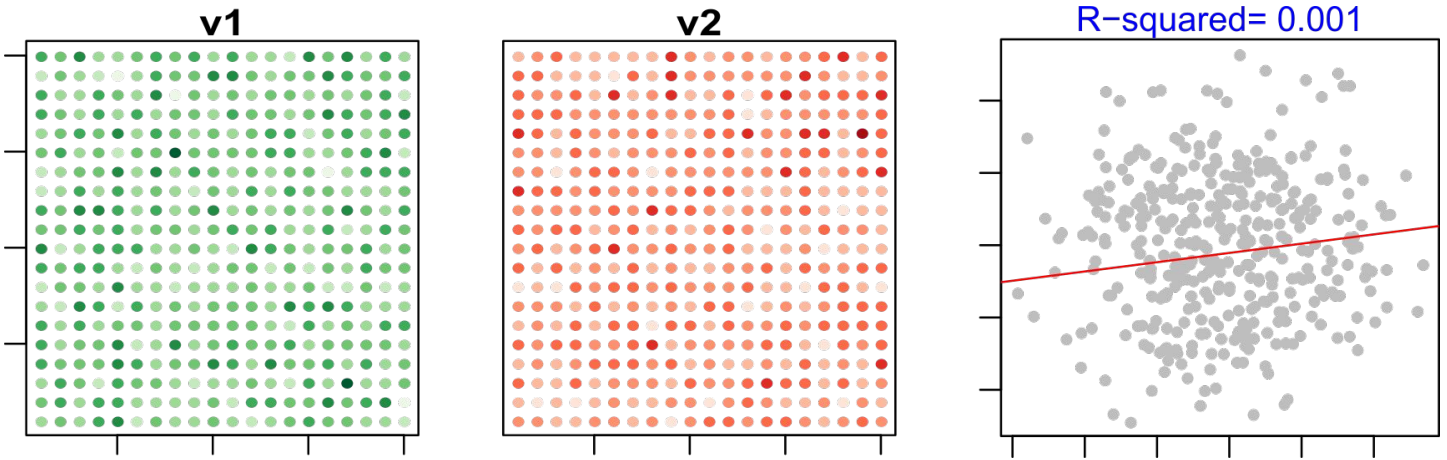
Modifiable Areal Unit Problem

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



Modifiable Areal Unit Problem

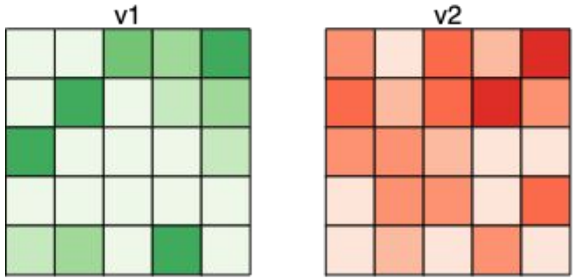
- We might want to know if there is correlation between v1 and v2



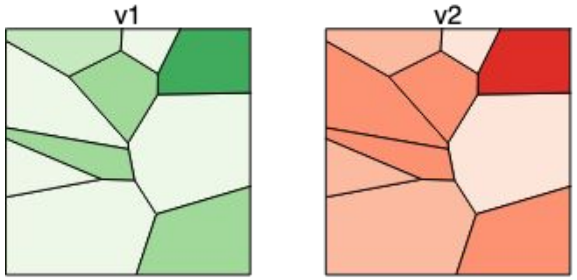
Modifiable Areal Unit Problem

- However, data often comes aggregated by polygons

Data summarized using a **uniform** aggregation scheme



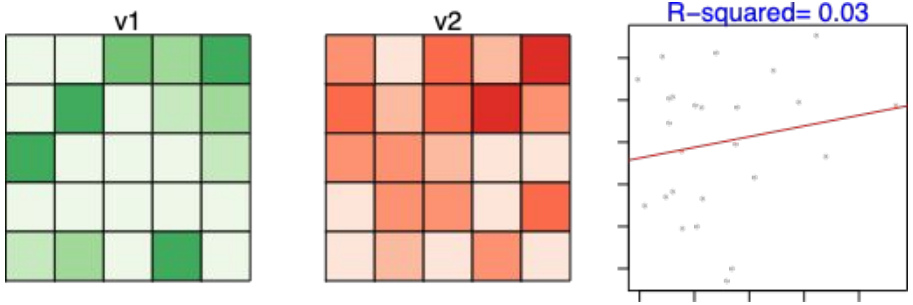
Data summarized using a **non-uniform** aggregation



Modifiable Areal Unit Problem

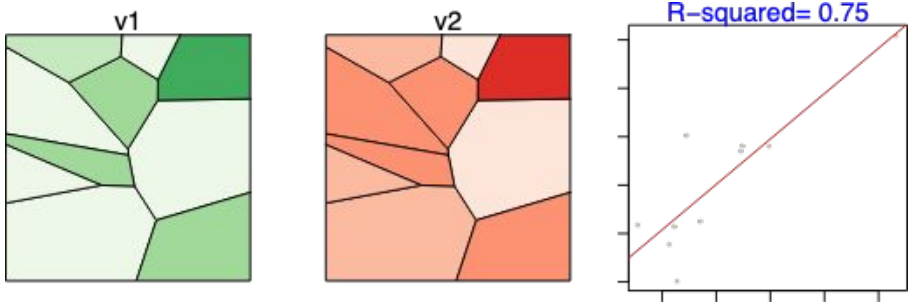
- If we are not careful, this can fool us into thinking a pattern exists when it doesn't

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!