Student project survey!

Class syllabi requested!

We're crowdsourcing data from PRIOR QUARTERS' (ideally mostly prior years') classes students have taken, on a per-class basis, using two methods: A. direct syllabus submission AND/OR B. a survey with questions corresponding to policies for that class.
We're doing this to research how course logistics policies (attendance and assignment lateness, as well as providing materials like class recordings or lecture notes) correlate to student performance and satisfaction (using CAPES data).
This is for our project for this class, and we'd appreciate if you'd submit for at least one class, or ideally multiple (and try to do ones from before this year, if you can).

Form A: Syllabus submission – .https://docs.google.com/forms/d/e/1FAIpQLSc7iF4juanYzbRetRmfeBBVrSSWkFADa9nU50KzYLVe2dbdkA/viewform

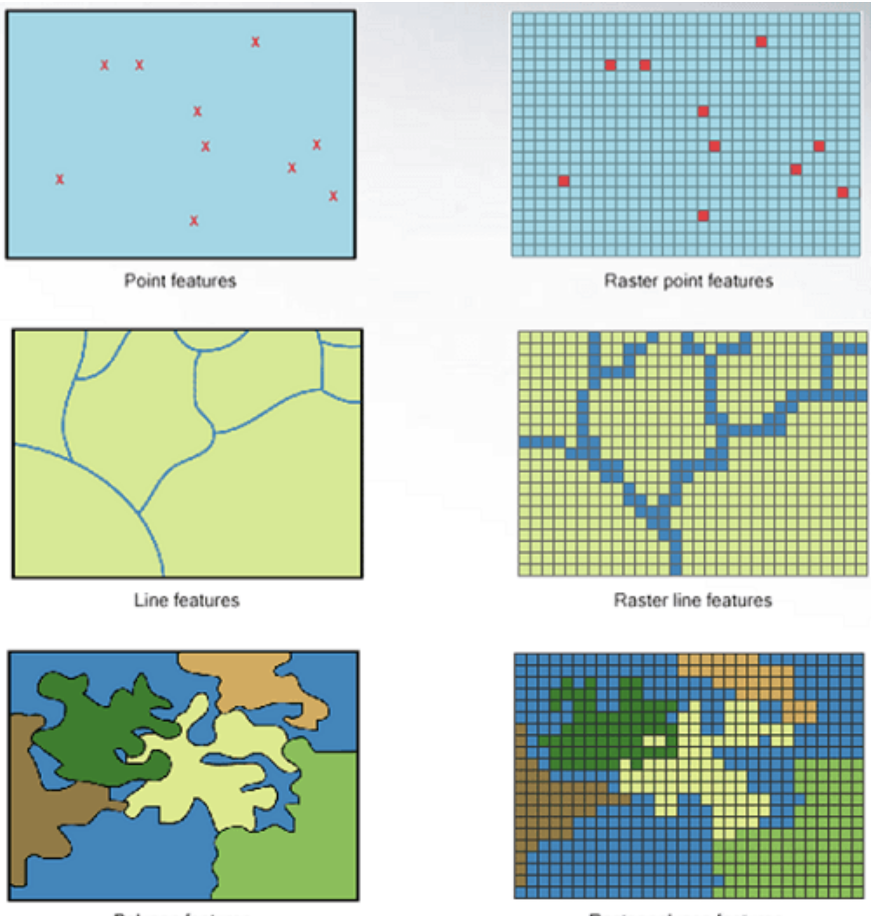
Form B: Direct info submission – https://docs.google.com/forms/d/e/1FAIpQLSfL5GjmCkemxm6t6XFKc6uF3xnU8sQ7U9Zw7aAzEfsKWBU-Xw/viewform

Representing spatial data and making maps like these

Representing shapes on maps

- Vector data
  - Points, lines, polygons
- Raster data
  - Encodes the world as a continuous surface represented by a grid, such as the pixels of an image.

Source: Humboldt State University



Point features

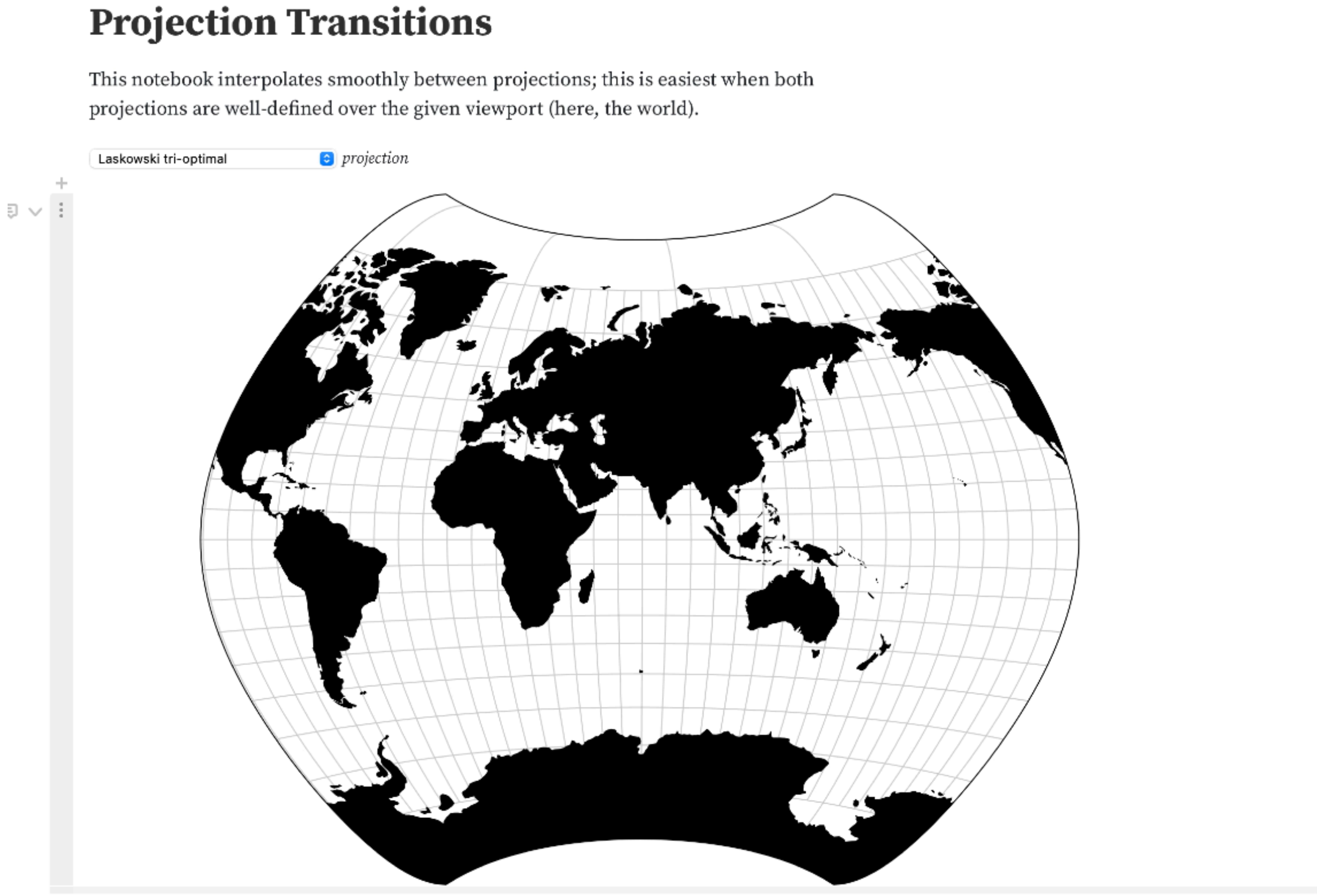
Raster point features

Line features

Raster line features

Polygon features

Raster polygon features



Projection Transitions

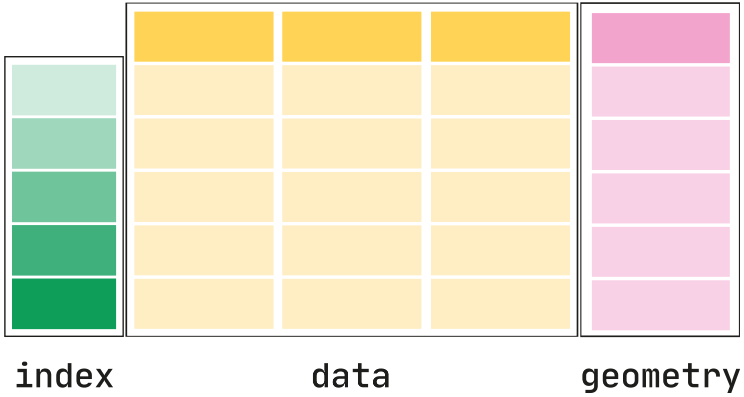
This notebook interpolates smoothly between projections; this is easiest when both projections are well-defined over the given viewport (here, the world).

Laskowski tri-optimal *projection*

Source: https://observablehq.com/@d3/projection-transitions

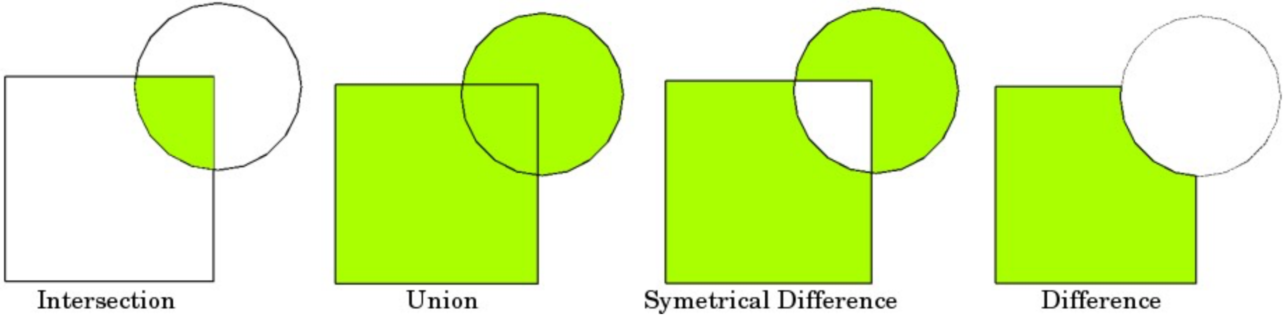
Geopandas

- Extends pandas, adds support for geospatial data
- Uses shapely library to represent vector geometry (e.g., POINT, LINE, POLYGON)
- Writes/reads shape files, GeoJSON and other formats
- Uses Coordinate Reference Systems (CRS) to represent how data aligns with the real world
- Different types of CRS: Geographic Coordinate Systems and Projected Coordinate Systems.



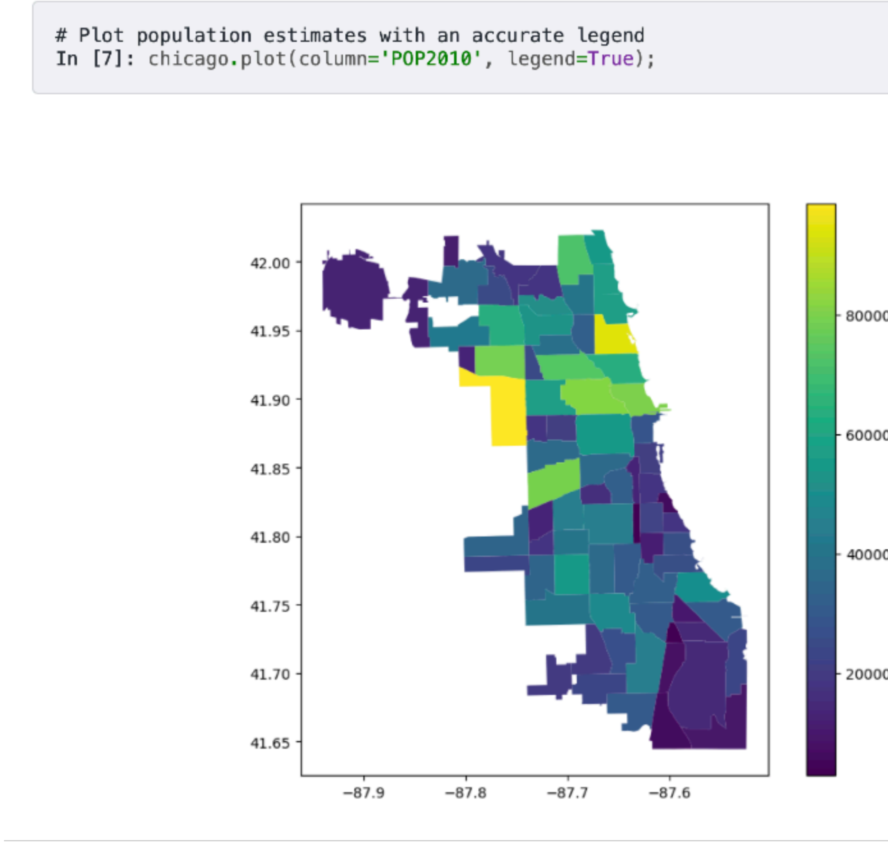
Geopandas skills

- Overlay spatial datasets: e.g. calculate what regions of Chicago are within 1km of a grocery store
- Join dataframes on shape information
- Aggregate across different subregions to get measurements at a more granular regional level



Geopandas tutorials

- https://geopandas.org/en/stable/getting_started/introduction.html
- https://geopandas.org/en/stable/docs/user_guide/mapping.html
- https://geopandas.org/en/stable/docs/user_guide/set_operations.html
- https://geopandas.org/en/stable/gallery/index.html



```
# Plot population estimates with an accurate legend
In [7]: chicago.plot(column='POP2010', legend=True);
```

Spatial Statistics : The Basics

Are countries with a high conflict index score geographically clustered?

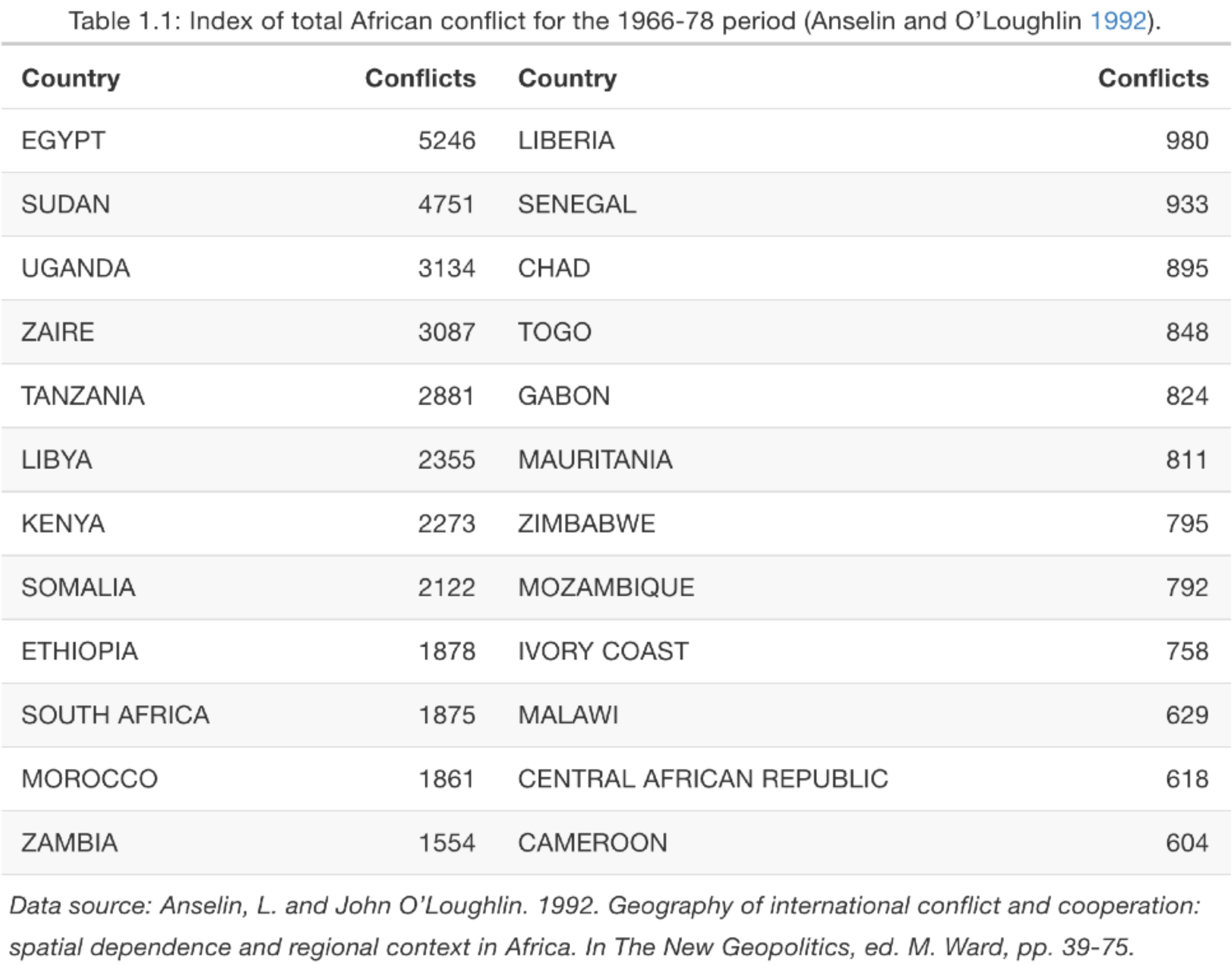


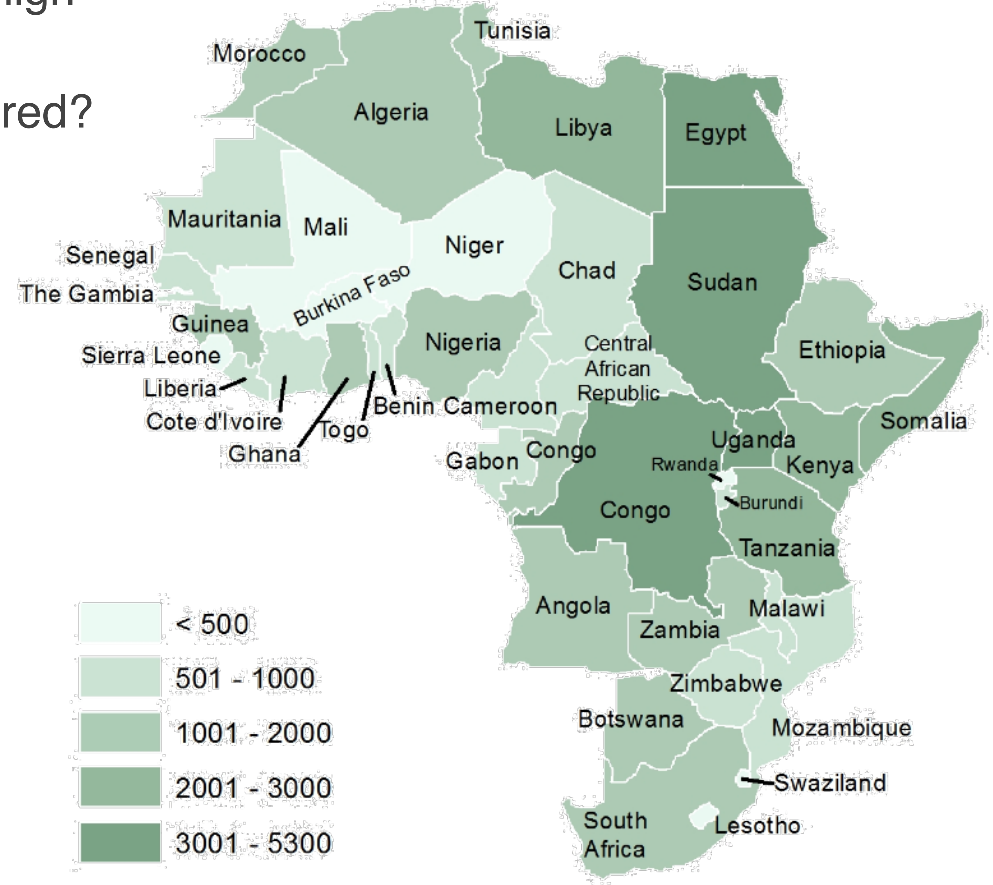
Table 1.1: Index of total African conflict for the 1966-78 period (Anselin and O'Loughlin 1992).

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.

https://mgimond.github.io/Spatial/introGIS.html

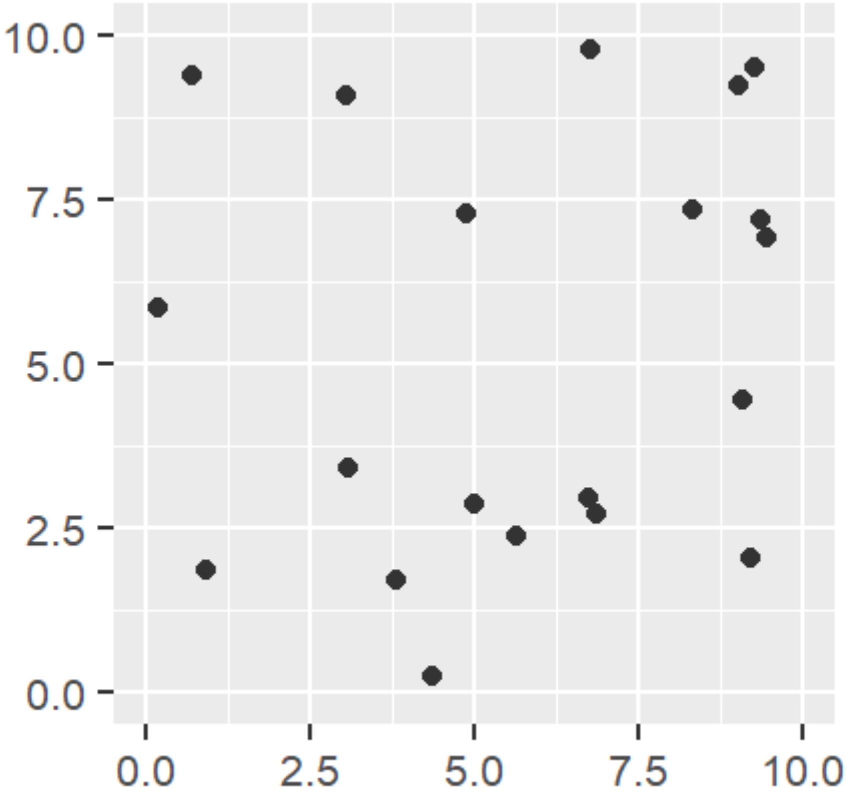
Are countries with a high conflict index score geographically clustered?



https://mgimond.github.io/Spatial/introGIS.html

Global Point Density

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

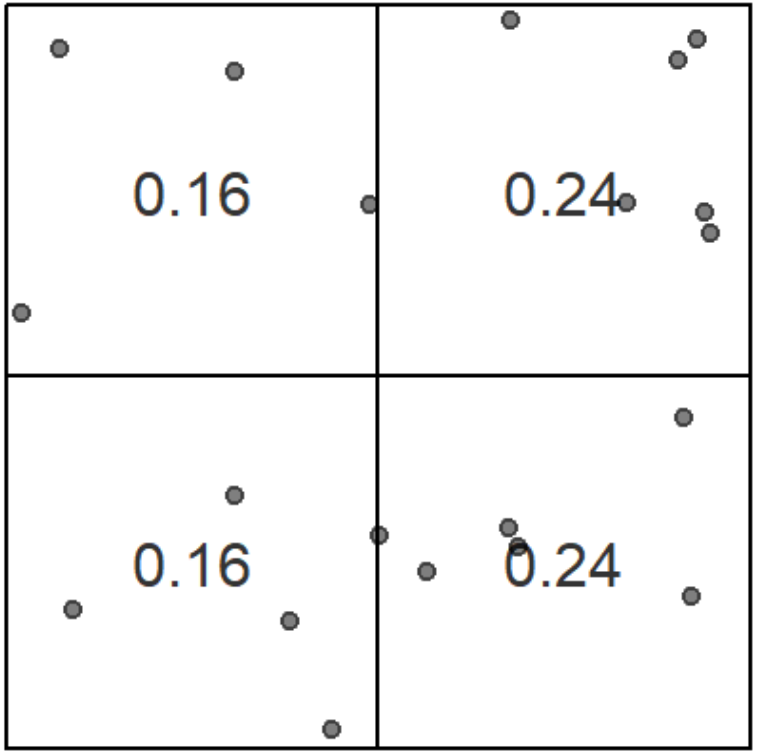


https://mgimond.github.io/Spatial/introGIS.html

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.

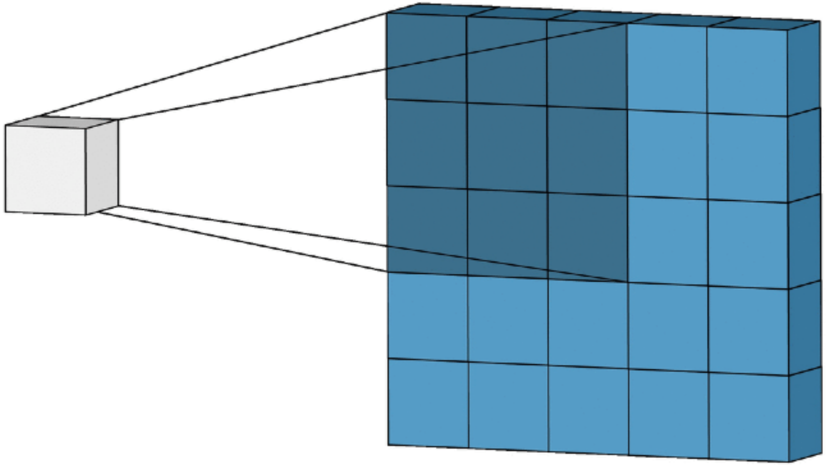


https://mgimond.github.io/Spatial/introGIS.html

Kernel Density (local)

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

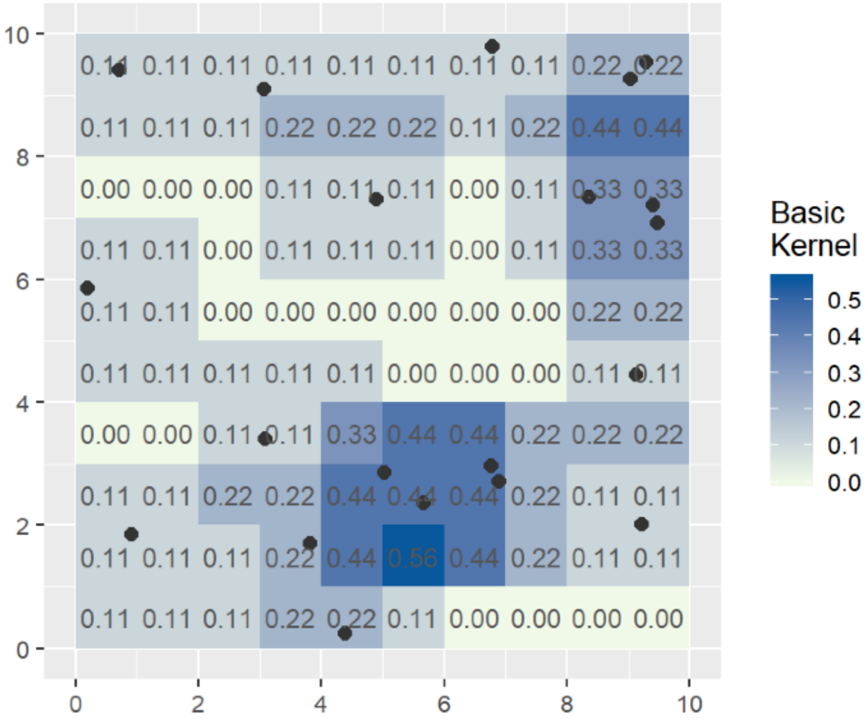
Note: kernel will affect measurement estimate, but this is less susceptible to MAUP.



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.



0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.22	0.22
0.11	0.11	0.11	0.22	0.22	0.22	0.11	0.22	0.44	0.44
0.00	0.00	0.00	0.11	0.11	0.11	0.00	0.11	0.33	0.33
0.11	0.11	0.00	0.11	0.11	0.11	0.00	0.11	0.33	0.33
0.11	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.22	0.22
0.11	0.11	0.11	0.11	0.11	0.00	0.00	0.00	0.11	0.11
0.00	0.00	0.11	0.11	0.33	0.44	0.44	0.22	0.22	0.22
0.11	0.11	0.22	0.22	0.44	0.44	0.44	0.22	0.11	0.11
0.11	0.11	0.11	0.22	0.44	0.56	0.44	0.22	0.11	0.11
0.11	0.11	0.11	0.22	0.22	0.11	0.00	0.00	0.00	0.00

https://mgimond.github.io/Spatial/introGIS.html

Modeling these data: Poisson Point Process

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

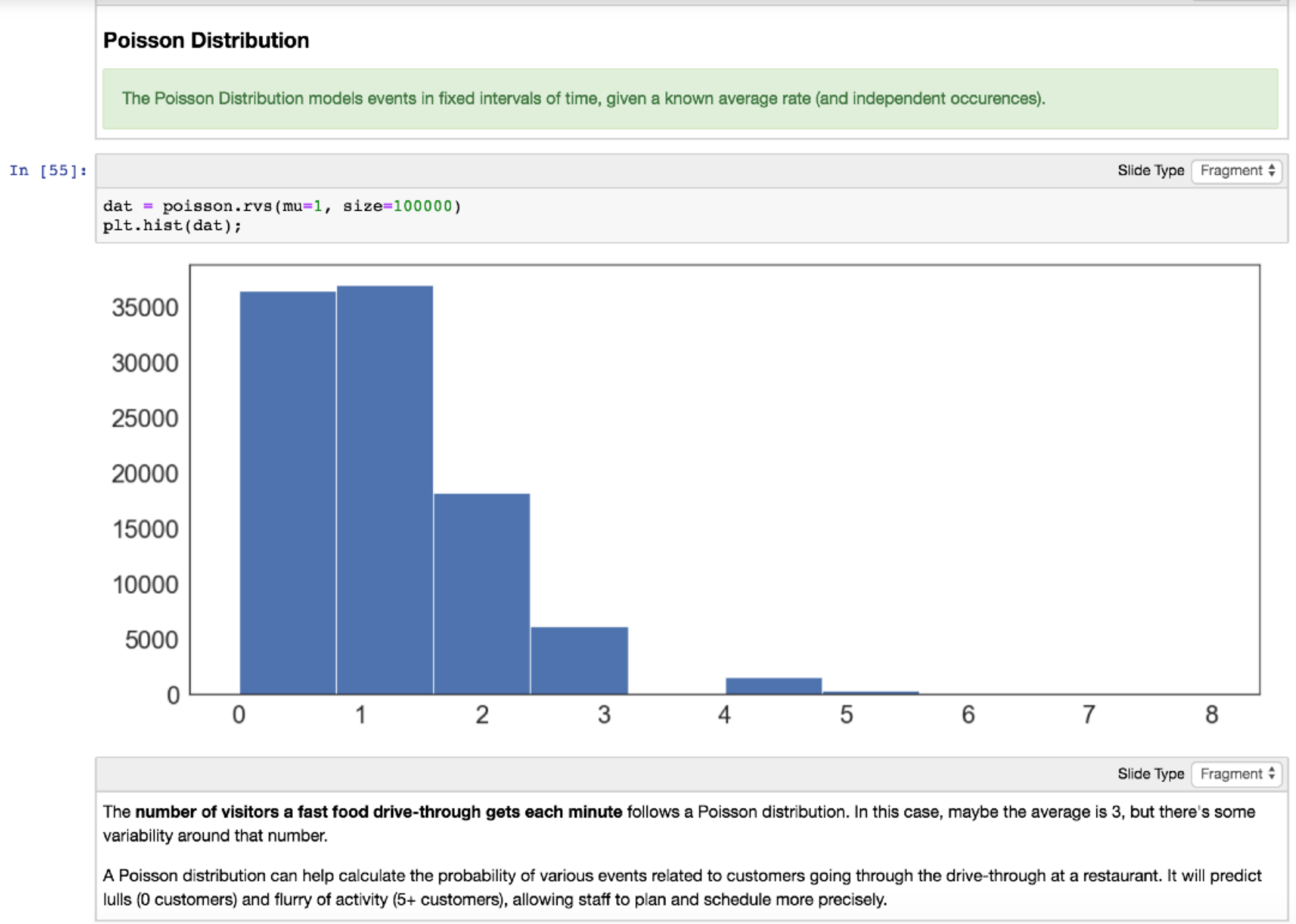
$$\lambda(i) = e^{\alpha+\beta Z(i)}$$

$\lambda(i)$ 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

https://mgimond.github.io/Spatial/introGIS.html



Poisson Distribution

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

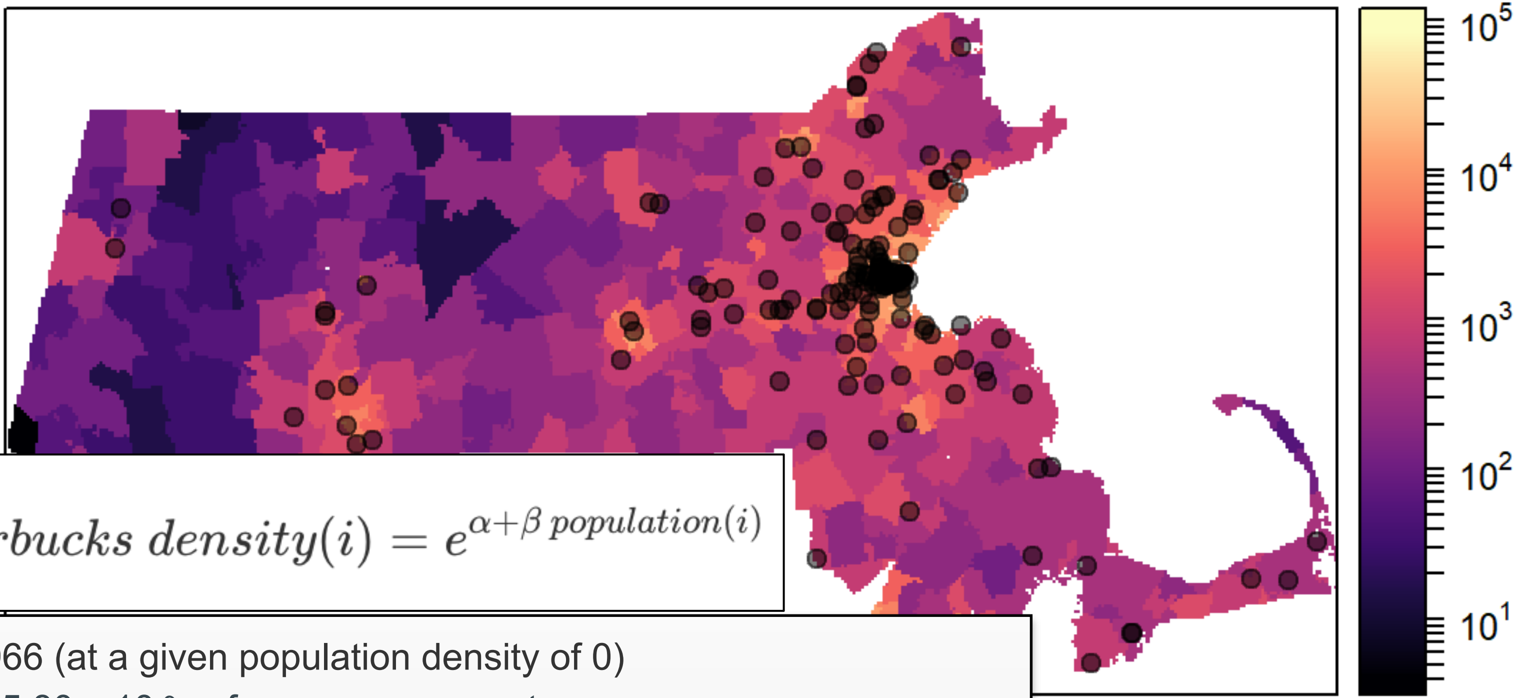
In [55]:

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

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

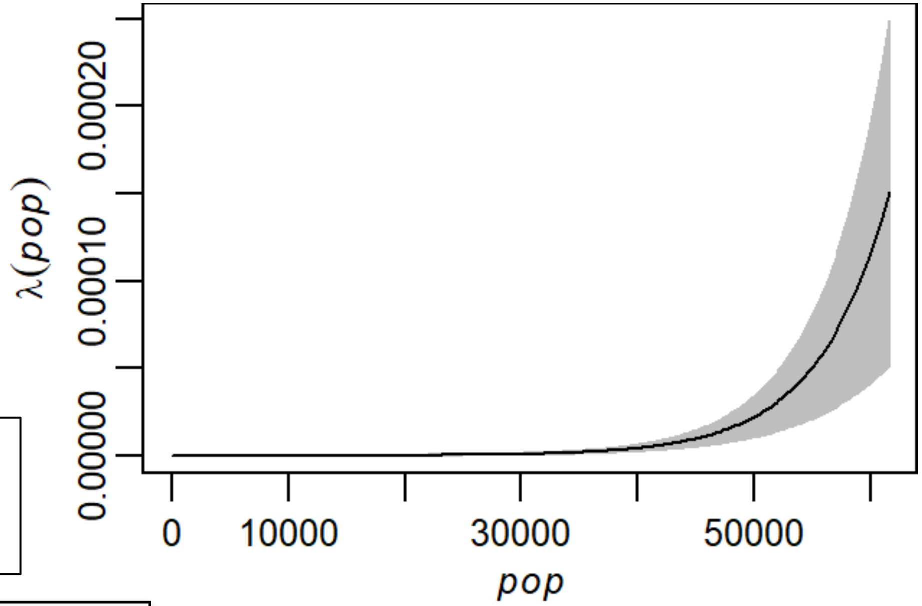


$$Starbucks\ density(i) = e^{\alpha+\beta\, population(i)}$$

α = -18.966 (at a given population density of 0)
$e^{-18.966}$ = 5.80 x 10^{-9} cafes per square meter
β = 0.00017; $e^{0.00017}$ or 1.00017

https://mgimond.github.io/Spatial/introGIS.html

Location of Starbucks relative to population density in MA



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

α = -18.966 (at a given population density of 0)

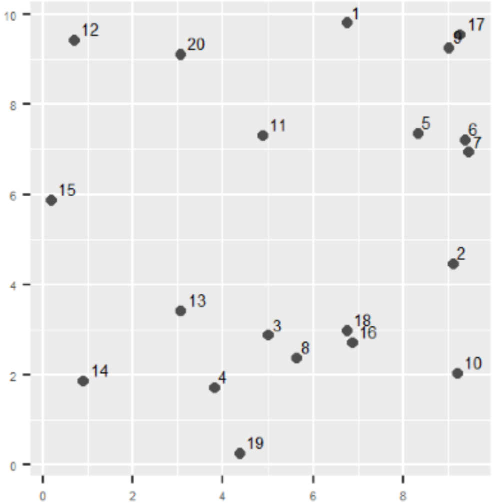
$e^{-18.966}$ = 5.80 x 10^{-09} cafes per square meter

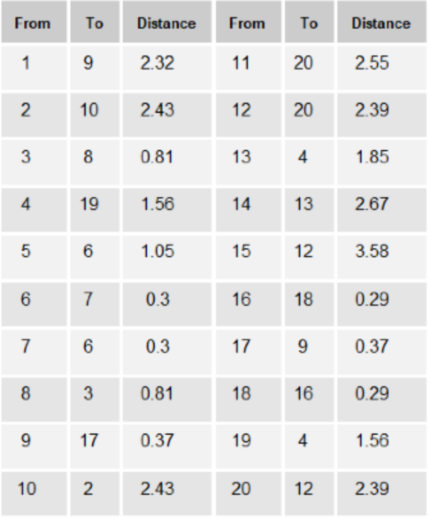
β = 0.00017; $e^{0.00017}$ or 1.00017

https://mgimond.github.io/Spatial/introGIS.html

Modeling these data: Average Nearest Neighbor

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





From	To	Distance	From	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

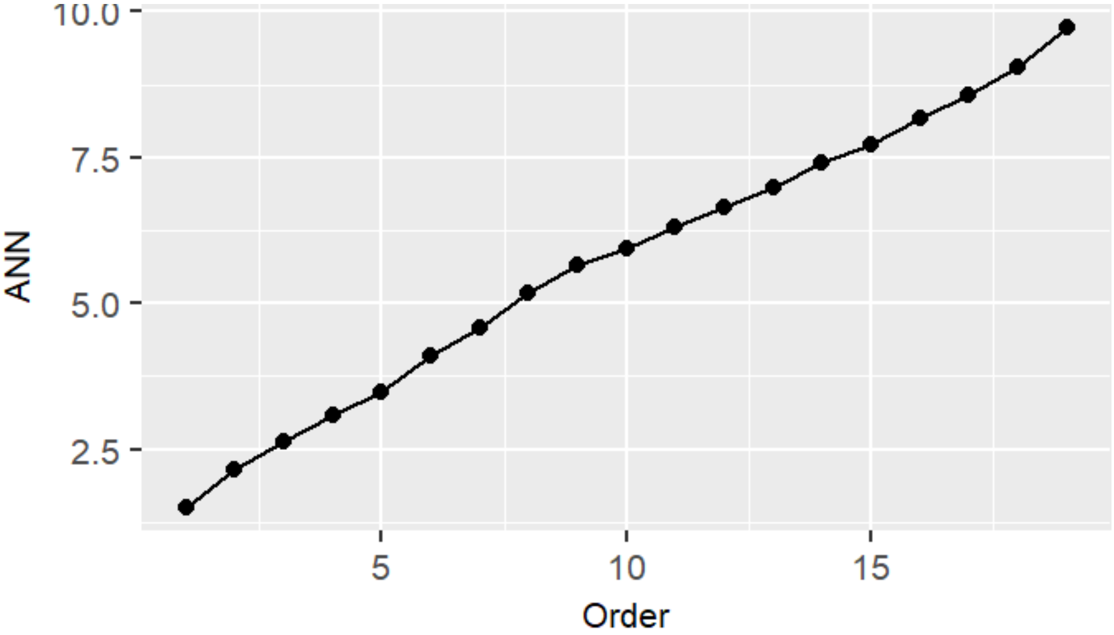
ANN = 1.52 units

https://mgimond.github.io/Spatial/introGIS.html

Modeling these data: Average Nearest Neighbor

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

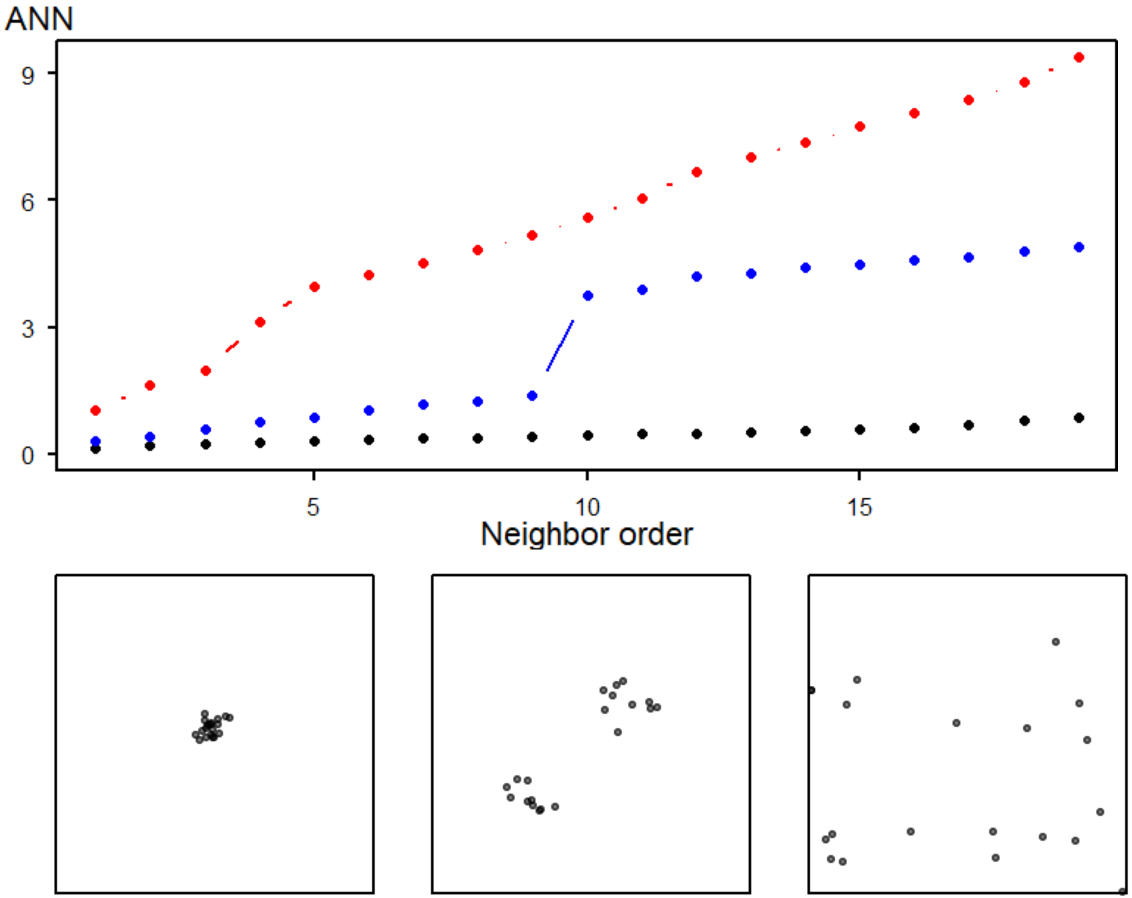
plot the ANN values for different order neighbors, that is for the first closest point, then the second closest point, and so forth.



https://mgimond.github.io/Spatial/introGIS.html

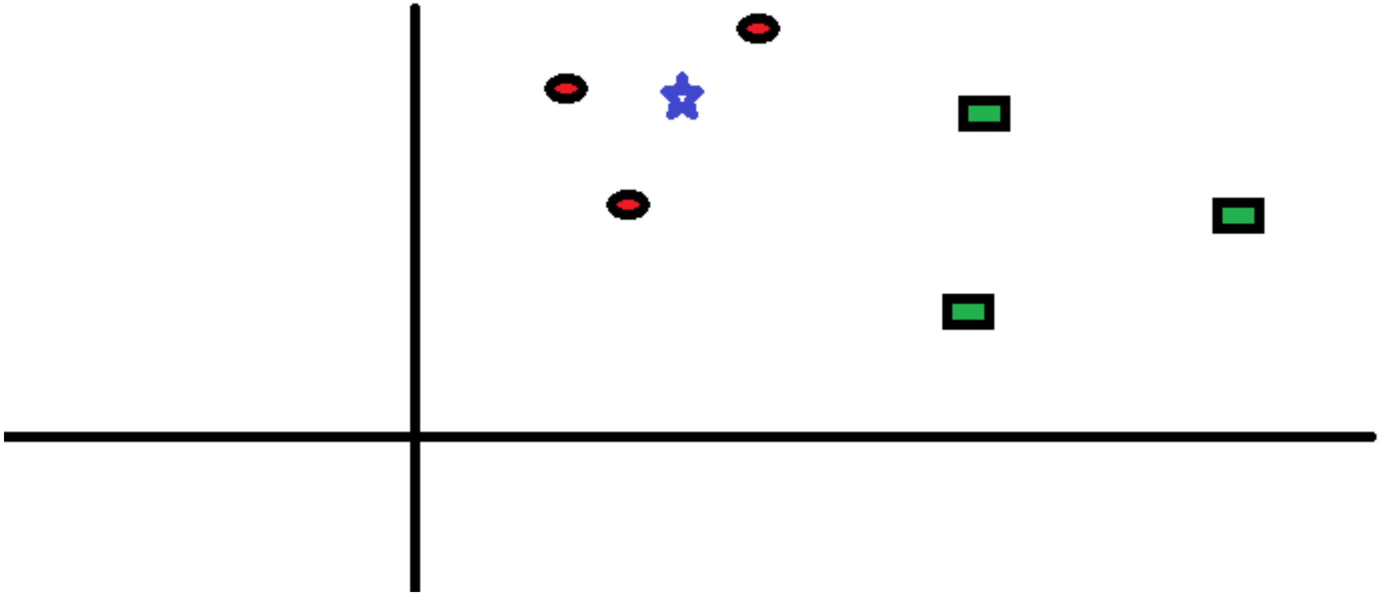
ANN vs neighbor order offers insight into underlying spatial relationship

Note: study space definition affects this measure



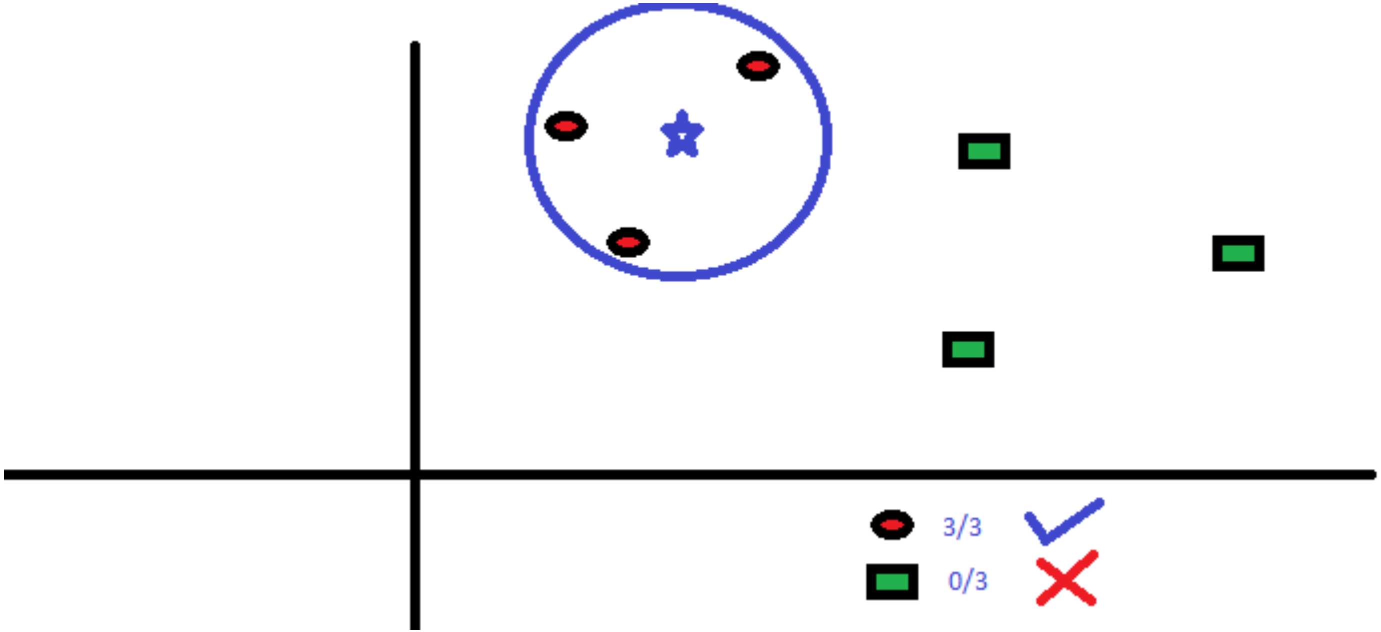
https://mgimond.github.io/Spatial/introGIS.html

KNN: K Nearest Neighbor for Classification



https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/

KNN: To which class does the blue star belong?

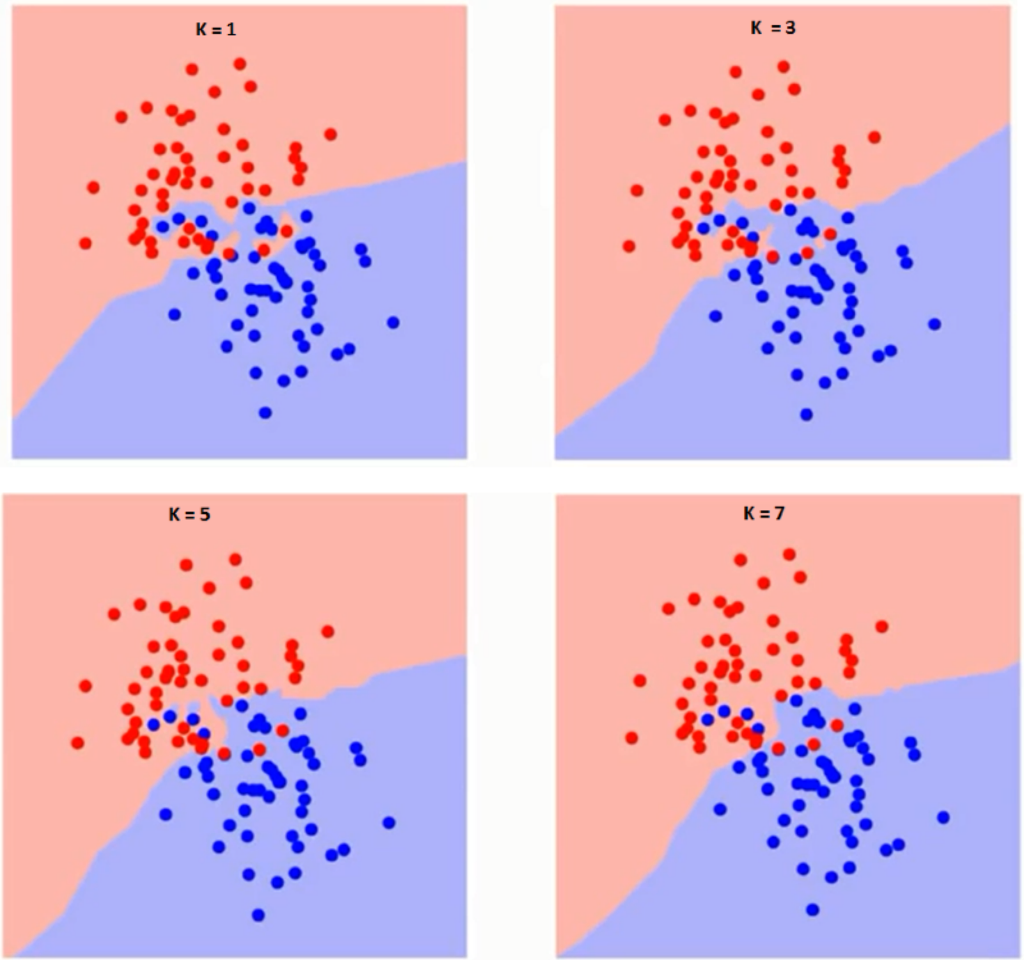


https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/

KNN: Choosing K

K specifies how many neighbors to consider.

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



https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/

KNN: Pros & Cons

Pros:

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

Cons:

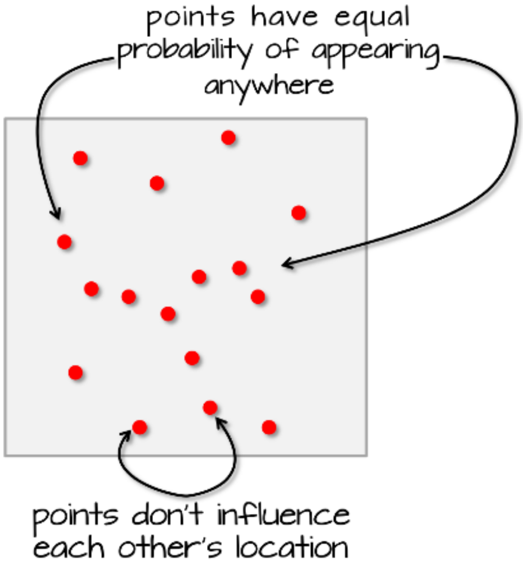
- Computationally intensive
- High Memory requirements
- Stores all (or most) of training data
- Prediction slow with large N
- Sensitive to outliers/irrelevant features

https://blog.usejournal.com/a-quick-introduction-to-k-nearest-neighbors-algorithm-62214cea29c7

Hypothesis Testing: CSR/IPR

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

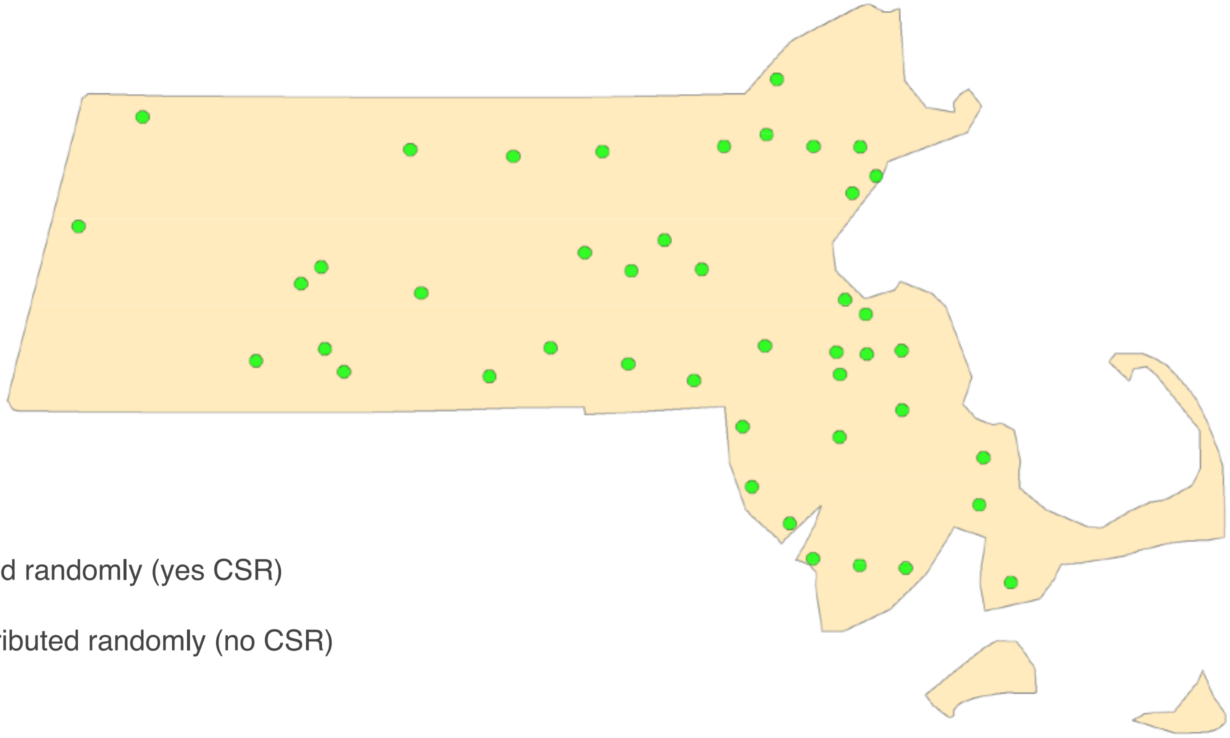
Compare observed point patterns to ones generated by an independent random process (IRP), aka complete spatial randomness (CSR).


points have equal probability of appearing anywhere
points don't influence each other's location

CSR/IRP 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_o: Distributed randomly (yes CSR)

H_a: NOT distributed randomly (no CSR)

https://mgimond.github.io/Spatial/introGIS.html

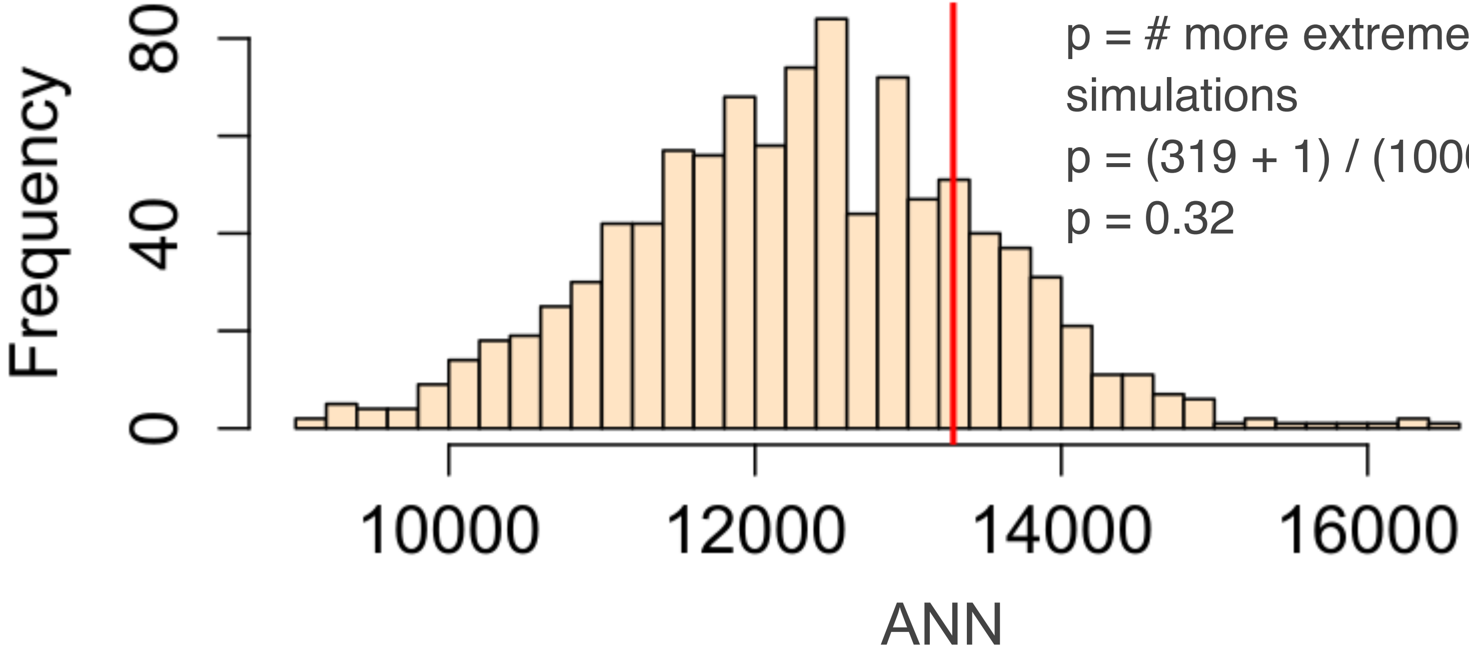
Hypothesis Testing: A Monte Carlo Test

1. First, we postulate a process–our null hypothesis, H_o.

   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!

https://mgimond.github.io/Spatial/introGIS.html

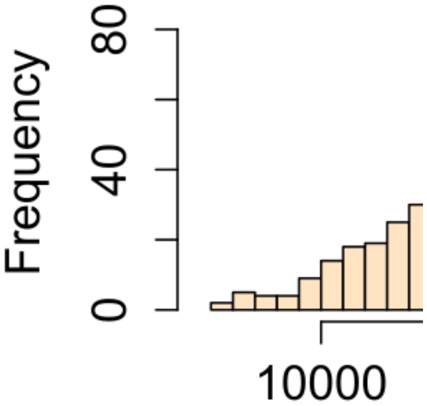


p = # more extreme / # simulations

p = (319 + 1) / (1000 + 1)

p = 0.32

https://mgimond.github.io/Spatial/introGIS.html

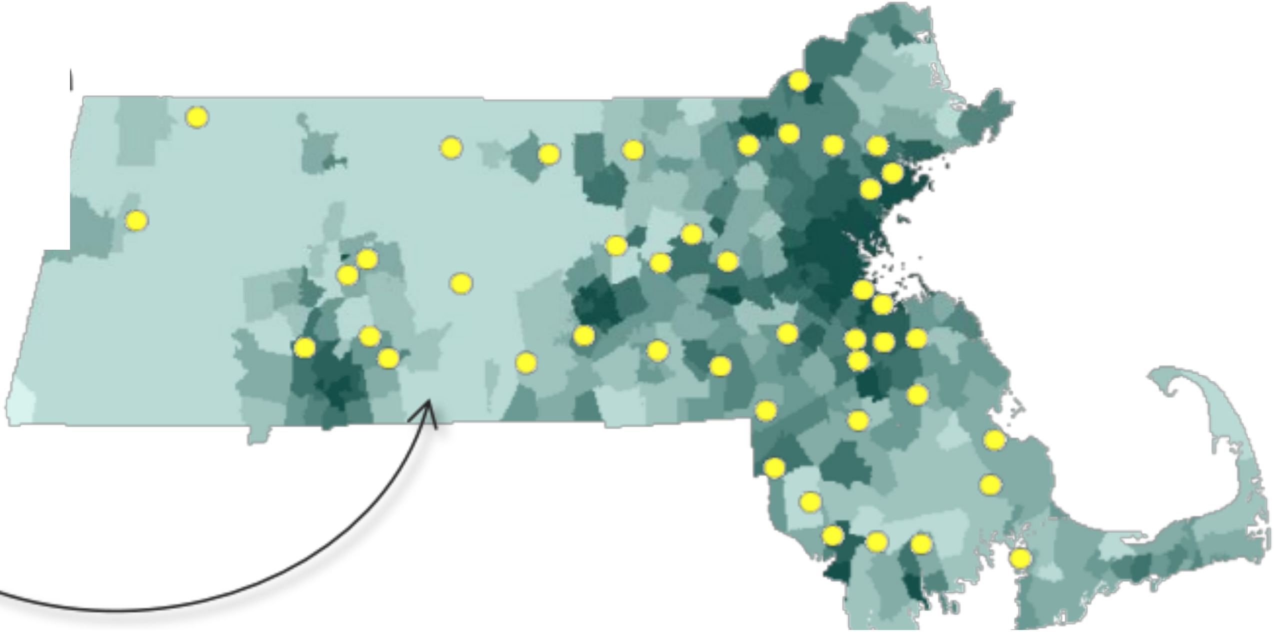


What does the histog

A
Alternative
Hypothesis

B
Null
Hypothesis

When controlling for population density, are Walmarts randomly distributed?

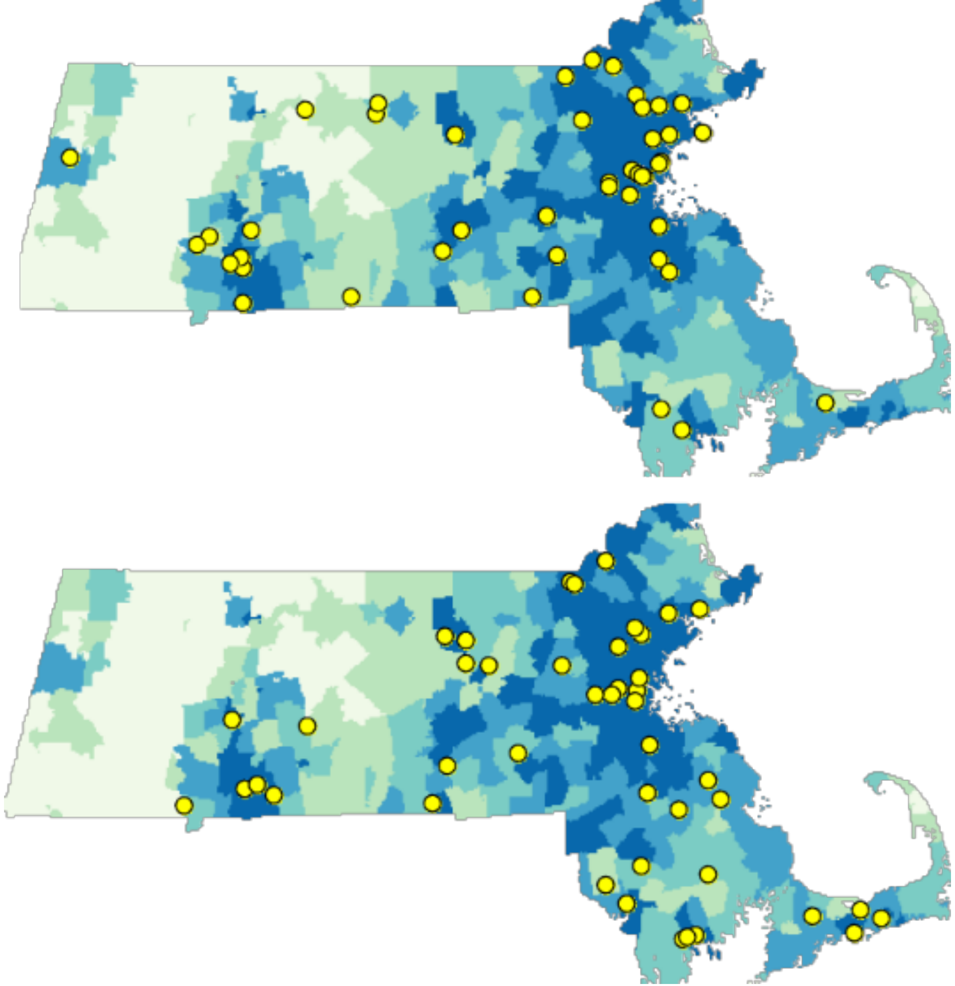


H_o: Walmarts are distributed according to population density alone

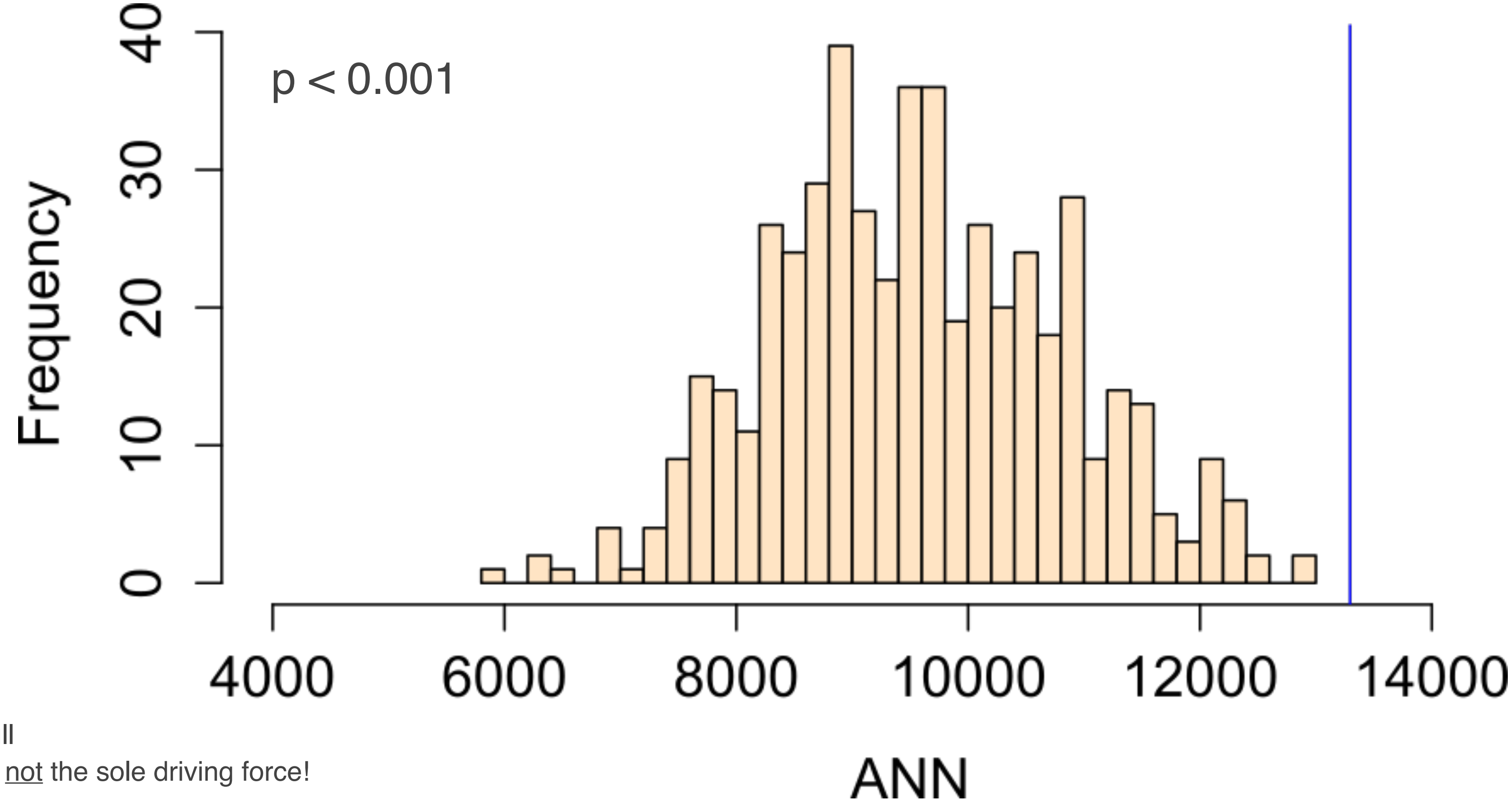
H_a: Walmarts are *not* distributed based on population density alone

https://mgimond.github.io/Spatial/introGIS.html

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



https://mgimond.github.io/Spatial/introGIS.html

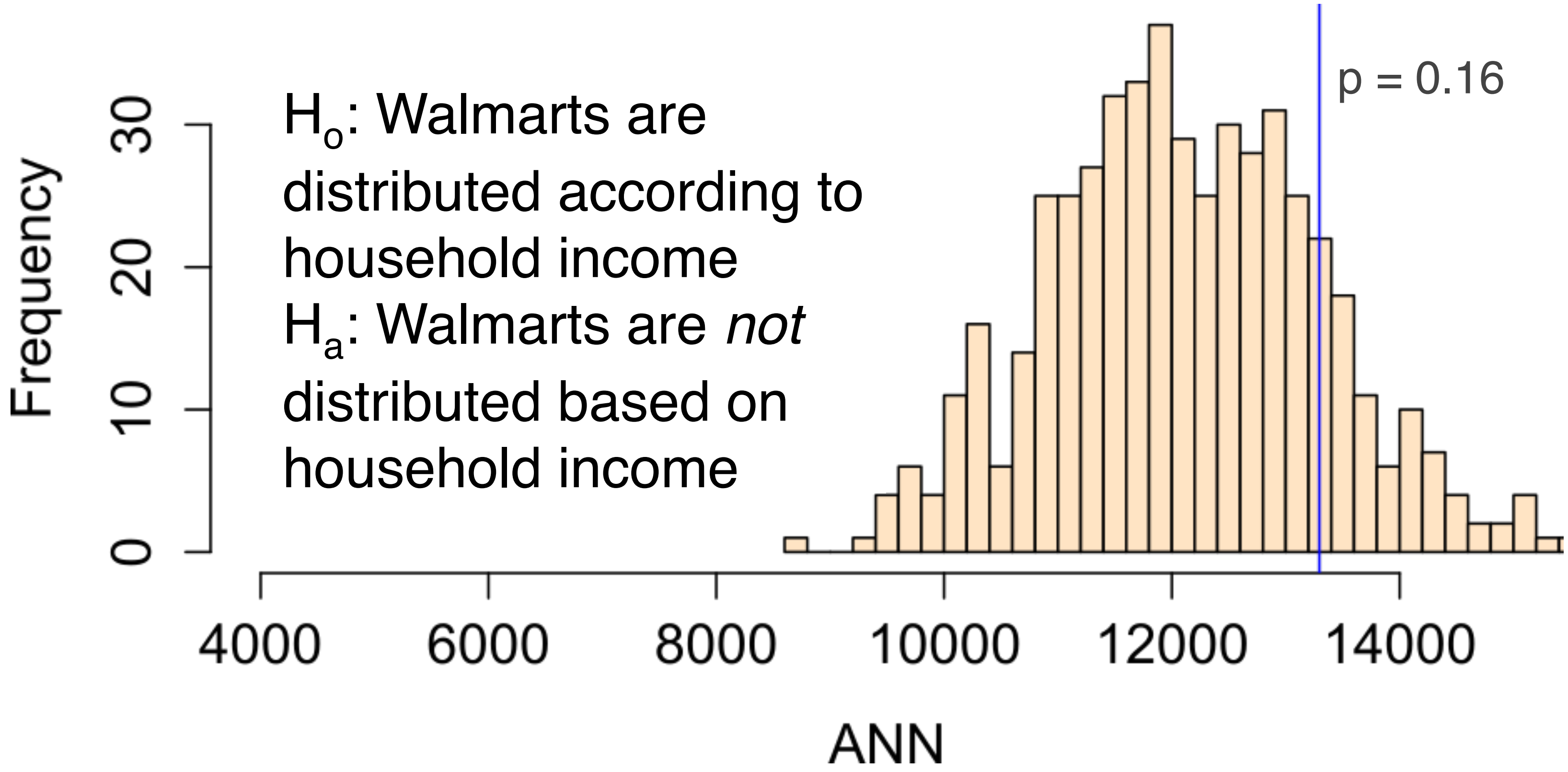


Reject the null

Population is not the sole driving force!

https://mgimond.github.io/Spatial/introGIS.html

Maybe median household income is the driving force…?



p = 0.16

H_o: Walmarts are distributed according to household income

H_a: Walmarts are *not* distributed based on household income

https://mgimond.github.io/Spatial/introGIS.html

...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)