GENERAL ASSEMBLY

SPATIAL AND SPATIOTEMPORAL DATA ANALYSIS

Matt Brems

Data Science Immersive, GA DC

LEARNING OBJECTIVES

- Define the three different types of spatial data.
- Create weight matrices.
- Describe the MAUP.
- Identify three common difficulties when working with spatiotemporal data.
- Describe three strategies for modeling with spatiotemporal data.

INTRODUCTION

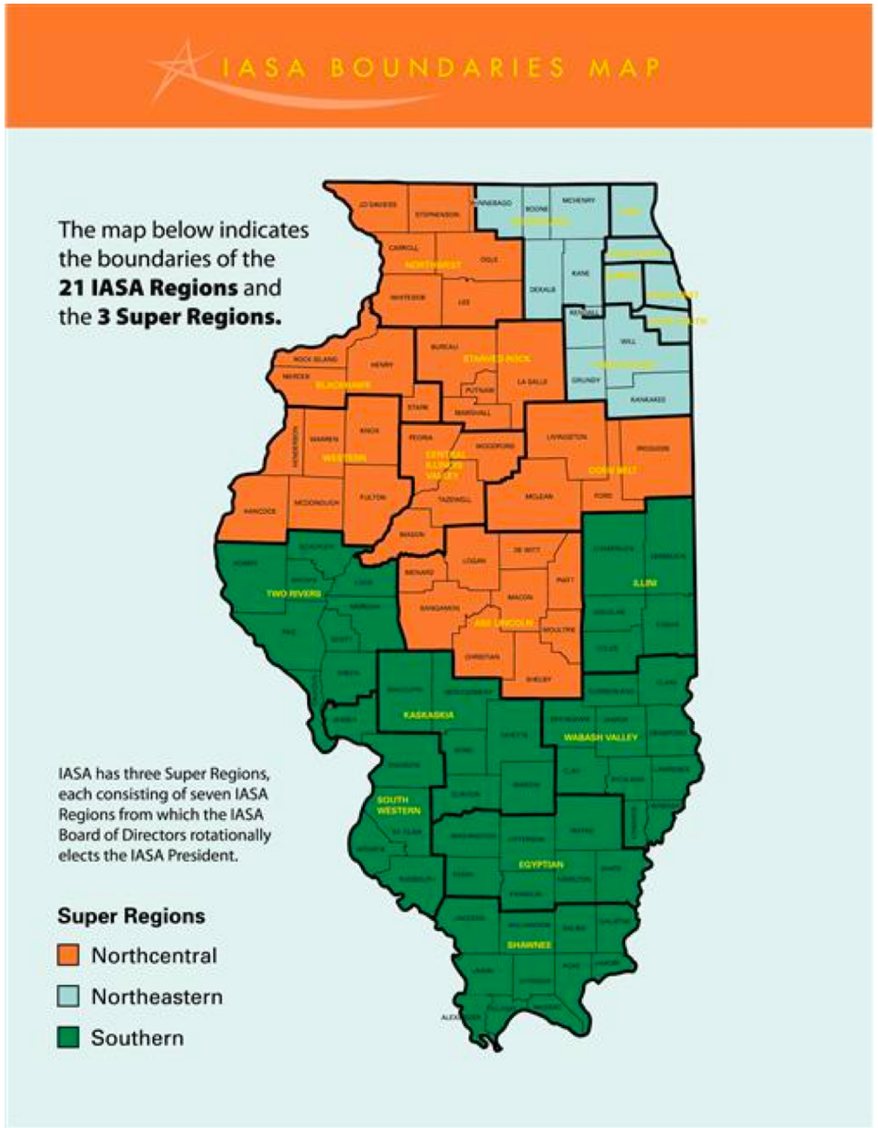
- Today, we'll continue our discussion of correlated data by jumping into spatial data, then how to integrate spatial data analysis with time series analysis.

INTRODUCTION

- Today, we'll continue our discussion of correlated data by jumping into spatial data, then how to integrate spatial data analysis with time series analysis.

- Why is it a problem to use our standard methods of analysis for spatiotemporal data?

EXAMPLE



- Suppose I want to measure the average amount of smog in the air in each of these three regions.

STRATEGY

- Instead of fitting this directly (and thus having more parameters than observations), we'll fit a spatial auto-regressive model where we explicitly account for the dependence in our observations.

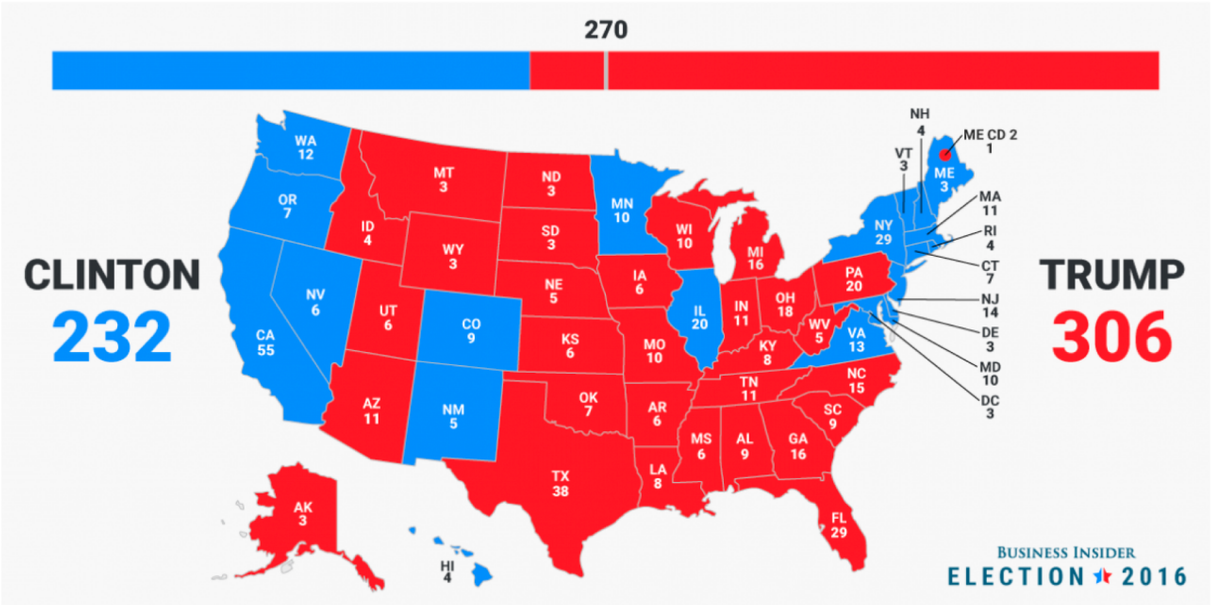
- This requires us to create W!

TOBLER'S FIRST LAW OF GEOGRAPHY

- One principle on which we'll rely in spatial data is known as Tobler's First Law of Geography:
  - "Everything is related to everything else, but near things are more related than distant things."

HOW DO WE DEFINE W?

- Our weight matrix W is what accounts for the dependencies in our observations.



SPATIOTEMPORAL DATA

- Spatiotemporal data is interpreted as a realization of a stochastic process. (This just means we view spatiotemporal data as a bunch of random variables and the values of these random variables.)

- We would formally write this $\{Y(s,t)|s \in D, t \in T\}$.

- The spatial domain D is what determines with what type of data we are working.

TYPES OF SPATIAL DATA

- If our spatial domain D is a set of non-overlapping regions, then we are working with an **areal process**. (i.e. states, ZIP codes)

- If our spatial domain D is continuous (likely 2-D or 3-D space), then we are working with a **geostatistical process**. (i.e. rainfall)

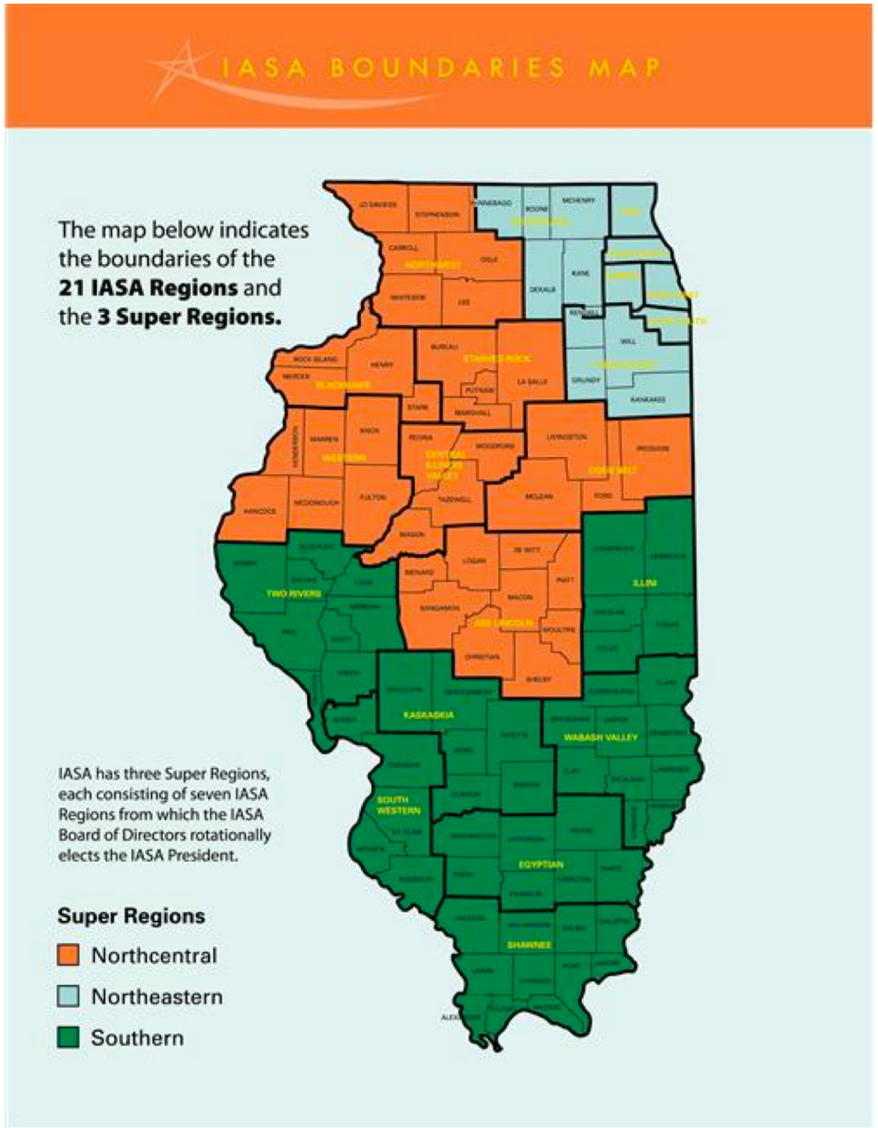
- If our spatial domain D is a collection of random points, then we are working with a **point pattern process**. (i.e. locations of car accidents)

AREAL DATA

- **Areal processes** are particularly annoying to handle because geography is messy and the regions we study might not have a specific rhyme or reason as to why they are structured the way they are.

- With areal data, we'll discuss:
  1. Weights
  2. Moran's *I* Statistic
  3. MAUP

AREAL PROCESSES: WEIGHTS



- Recall that, instead of fitting a full set of models, we want to fit a model that takes our spatial dependencies into account.
    - This requires us to define W.

AREAL PROCESSES: WEIGHTS

- We use a weight matrix W to measure how closely related two regions are. This describes how much weight to assign to region j when measuring region i.

- $w_{ii} = 0$ in all cases. (Why?)

AREAL PROCESSES: WEIGHTS

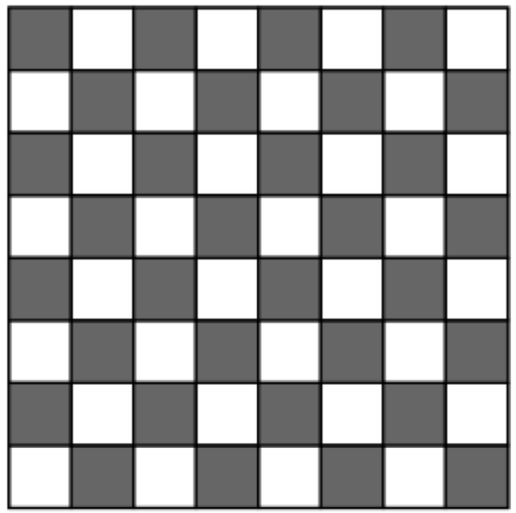
- We use a weight matrix W to measure how closely related two regions are. This describes how much weight to assign to region j when measuring region i.

- $w_{ii} = 0$ in all cases. (Why?)

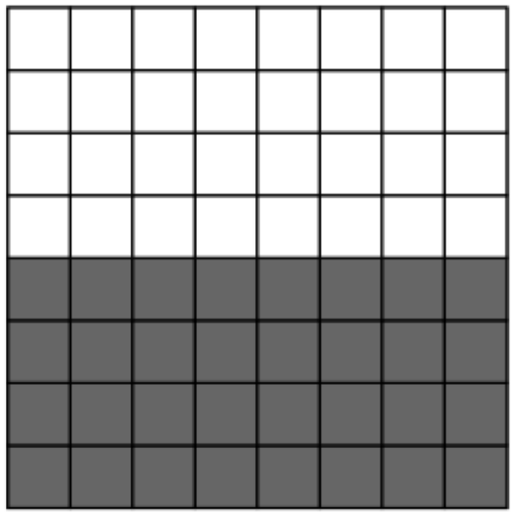
- While the choice is otherwise arbitrary, there are a few standard choices:
  - $w_{ij} = 1$ if region i borders region j (or is within a certain distance of j), otherwise 0.
  - $w_{ij} = \rho$ if region i borders region j, ρ^2 if i and j are one neighbor apart, and so on. (In this case, $0 < \rho < 1$ makes the most sense.)
  - $w_{ij} = \text{dist}\{centroid\ i, centroid\ j\}^{-1}$.

AREAL PROCESSES: MORAN'S I

- Suppose you want to detect whether or not spatial autocorrelation even exists!
  - We can use Moran's *I* statistic, which is a permutation test.



-1 (Perfect Dispersion)



+1 (Perfect Correlation)

AREAL PROCESSES: MORAN'S I

- Suppose you want to detect whether or not spatial autocorrelation even exists!
  - We can use Moran's I statistic, which is a permutation test.

- The most commonly used statistical test is the permutation test using Moran's I statistic. (https://pysal.readthedocs.io/en/latest/library/esda/moran.html)
  - H_0: no spatial correlation
  - H_A: spatial correlation exists.

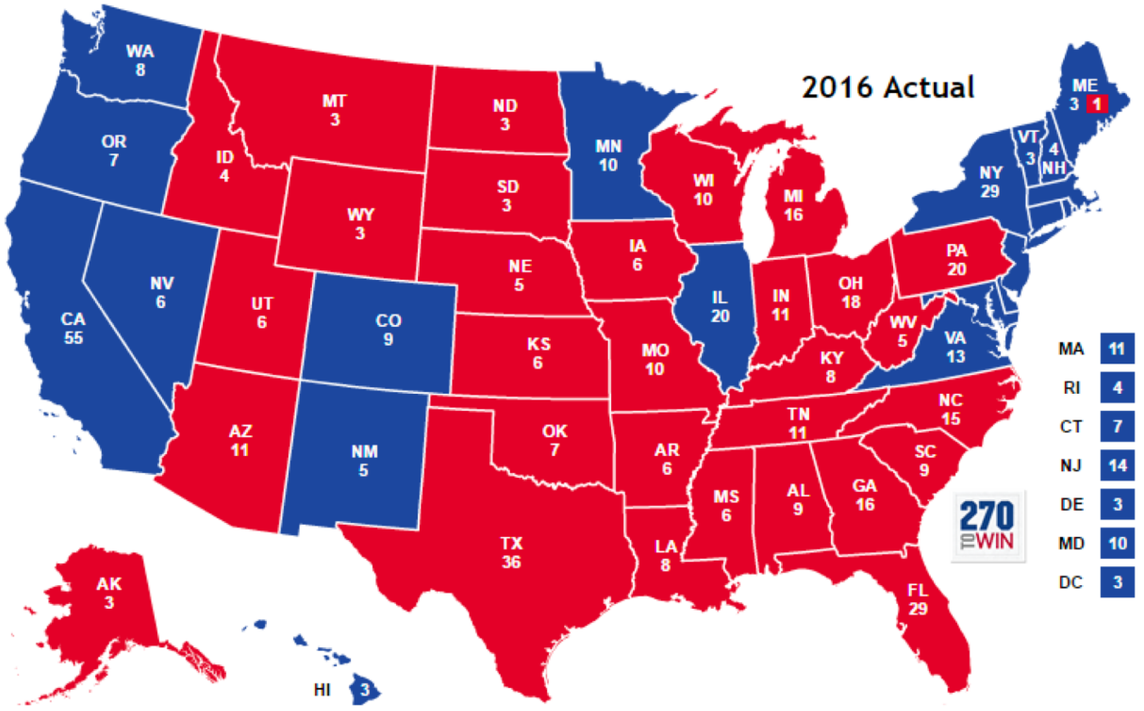
- **Note**: Small p-values provide evidence that that spatial correlation exists, but they don't provide information about the direction of the association!

AREAL PROCESSES: MAUP

- In **areal problems**, we have one massive problem we need to deal with.

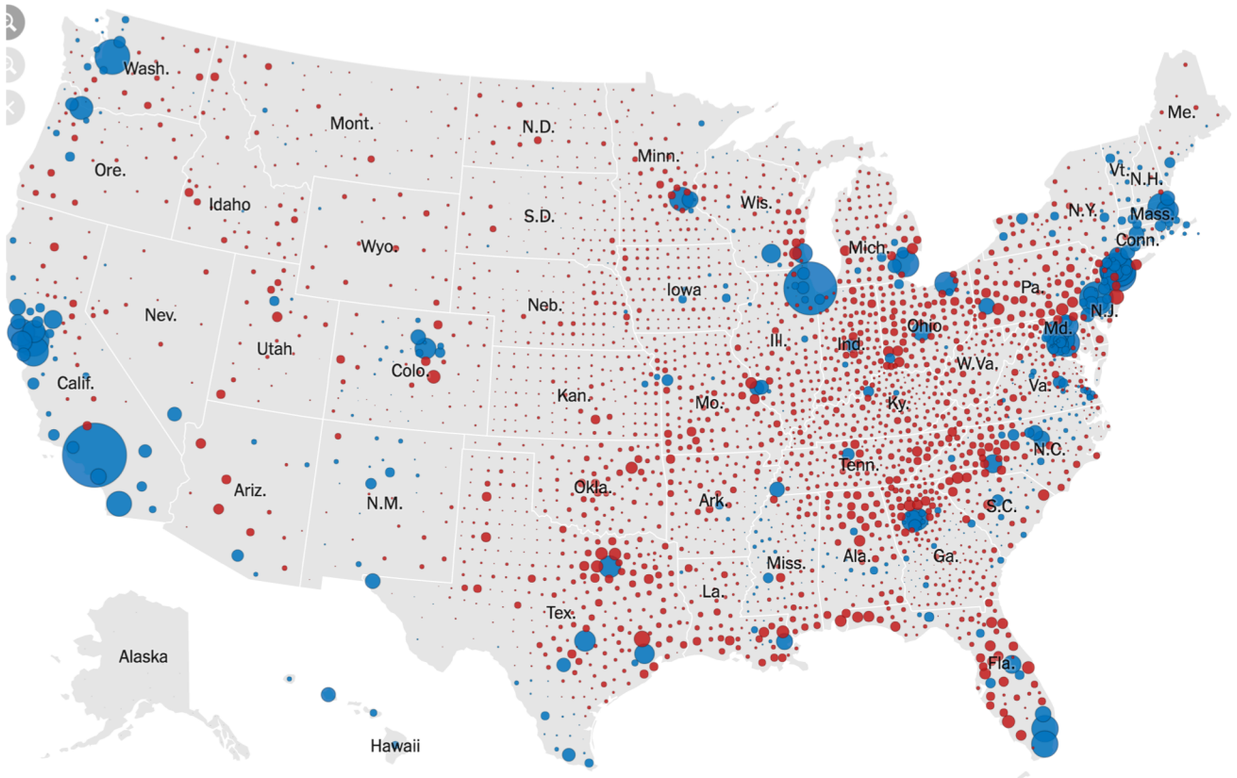
AREAL PROCESSES: MAUP

- In **areal problems**, we have one massive problem we need to deal with.



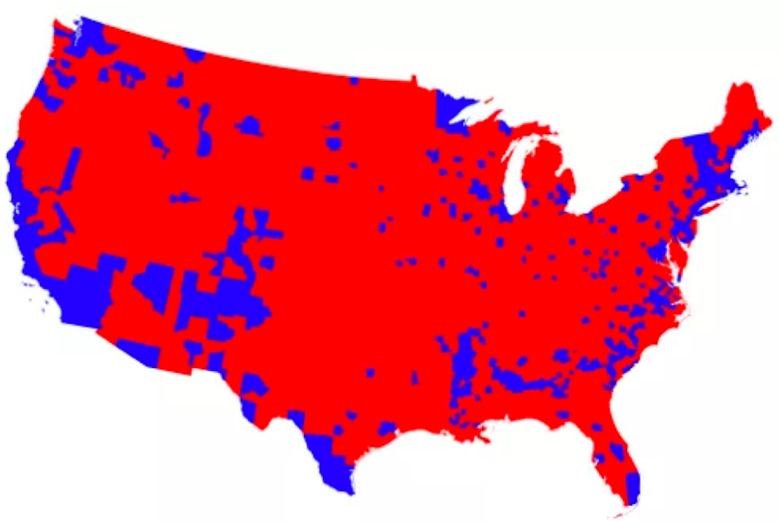
AREAL PROCESSES: MAUP

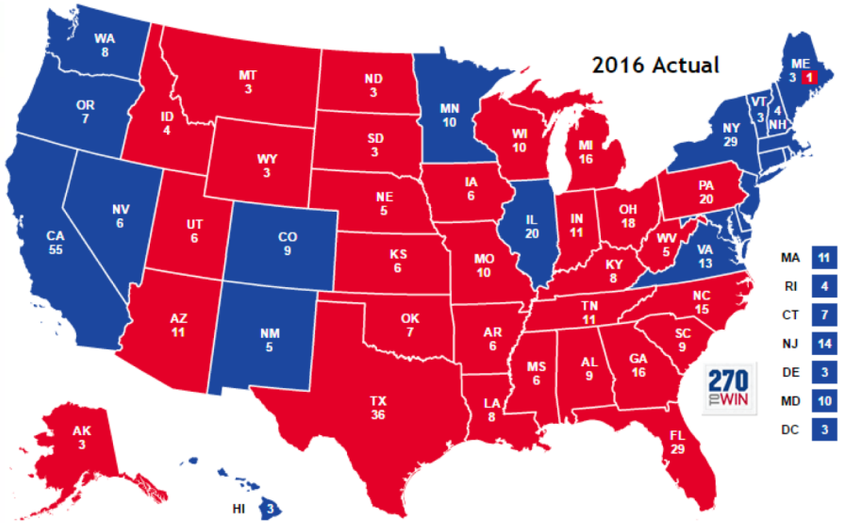
- In **areal problems**, we have one massive problem we need to deal with.

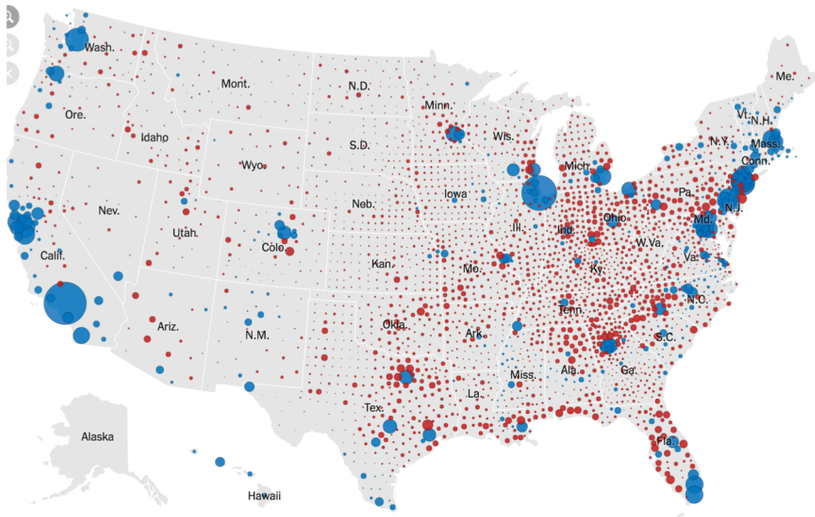


AREAL PROCESSES: MAUP

- In **areal problems**, we have one massive problem we need to deal with.

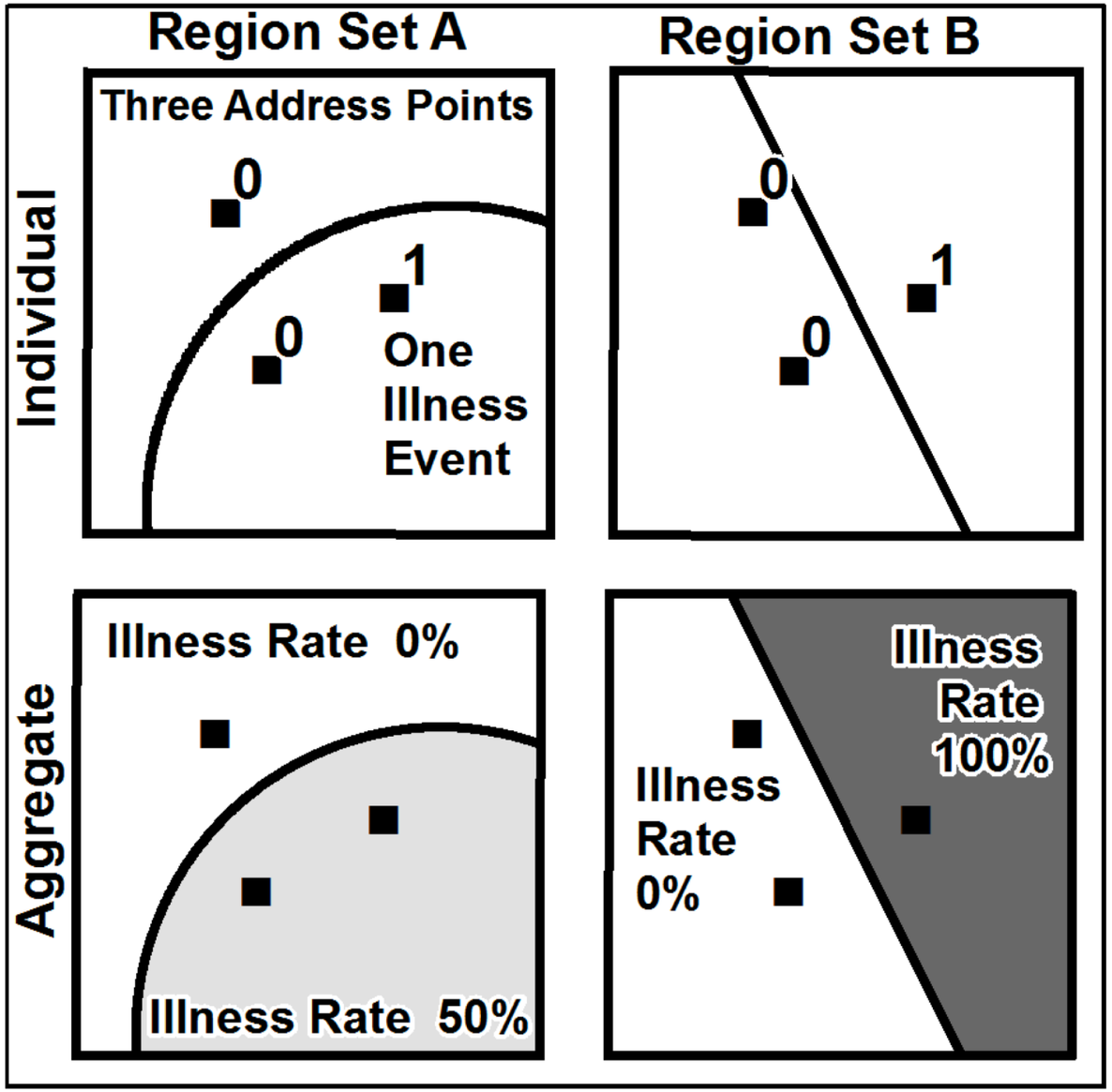






- This is known as the **Modified Areal Unit Problem** (MAUP).

AREAL PROCESSES: MAUP



- When we aggregate data into regions and have one measurement for that region, we're naturally going to lose information.
  - Depending on the shape and scale of our areal units, we can often get statistics to say what we want.
- This is closely related to the problem of gerrymandering,

GEOSTATISTICAL & POINT PATTERN

- With geostatistical processes and point pattern processes, we can generally rely on “natural distance,” so designing a weight matrix is not difficult.

- We also do not need to worry about the MAUP.

- When in doubt, defer to **nonparametric hypothesis tests**.
  - If you’re working with “irregularly placed data gatherers,” this might be particularly helpful.

MODELING WITH SPATIOTEMPORAL DATA

- Spatiotemporal data is commonplace, but working with it can be difficult.

- We'll outline three difficulties with spatiotemporal data, then get into modeling and assumptions.

DIFFICULTIES IN SPATIOTEMPORAL DATA ANALYSIS

- Train/Test Split
- Gathering Data
- Visualization

TRAIN/TEST SPLIT

- Why might using a simple random sample of 30% of our data for testing be improper?

TRAIN/TEST SPLIT

- Why might using a simple random sample of 30% of our data for testing be improper?
  - Use stratified random sampling to ensure a good cross-section of space and time data in both the training and testing sets. (Recommended.)
  - Alternatively, you may use certain representative locations as the test locations. (This is known as a cluster sample.)
  - If your goal is prediction/forecasting, it might make sense for you to use early time periods in your training and later time periods in your testing.

GATHERING DATA

- Gathering spatio-temporal data can be quite difficult:
  1. Generally need a significant amount of data.
  2. Differing reporting periods. (i.e. daily vs. weekly)
  3. Format of data.

- Munging spatio-temporal data can be quite time intensive and, once finished, you may not have enough data to build an accurate model!

VISUALIZATION

- Spatiotemporal data are notoriously difficult to visualize, because humans can't think in many dimensions at once!
  - Latitude, longitude, and time are already three dimensions – but we're likely interested in studying something else.
  - https://www.gapminder.org/tools/#_data_/_lastModified:1521724031820;&chart-type=bubbles

- Be smart about how you include "time" and "space" in your visuals:
  - https://fivethirtyeight.com/features/the-52-best-and-weirdest-charts-we-made-in-2016/
  - https://bokeh.pydata.org/en/latest/docs/gallery.html

SPATIOTEMPORAL MODELING STRATEGIES

- Recall that our data are given by $Y(s,t)$. Let's decompose our data into two components:
  - Model: $\mu(s,t)$
  - Noise: $\varepsilon(s,t)$

- There are three broad categories of things we might want to do with a model:
  - Inference
  - Prediction/Simulation
  - Adjustment

SPATIOTEMPORAL MODELING STRATEGIES: INFERENCE

- Recall that our data are given by $Y(s,t)$ and that $Y(s,t) = \mu(s,t) + \varepsilon(s,t)$.

- If our goal is to **conduct inference** on how space and time affect $Y(s,t)$, we should only include space and time components in $\mu(s,t)$.
  - For example, latitude, longitude, time, $Y(s,t-1)$ or $Y(s-1,t)$.

SPATIOTEMPORAL MODELING STRATEGIES: FORECASTING

- Recall that our data are given by $Y(s,t)$ and that $Y(s,t) = \mu(s,t) + \varepsilon(s,t)$.

- If our goal is to **predict/forecast** values of $Y(s,t)$, we should include space and time components in $\mu(s,t)$ **and other factors** that help predict $Y(s,t)$.
  - A commonly used method of forecasting is to **simulate** values of $Y(s,t)$ given a series of inputs.
  - The Markov chains example of weather the other day was a rudimentary form of this.

SPATIOTEMPORAL MODELING STRATEGIES: ADJUSTMENT

- Recall that our data are given by $Y(s,t)$ and that $Y(s,t) = \mu(s,t) + \varepsilon(s,t)$.

- If our goal is to **adjust for/negate** the effect of time/space on $Y(s,t)$, we should include only space and time components in $\mu(s,t)$, then study $Y(s,t) - \mu(s,t) = \varepsilon(s,t)$ instead.

SIMPLIFYING ASSUMPTIONS (USUALLY UNREALISTIC)

- **Stationary**: A stationary spatio-temporal process is one that has:
  1. A constant mean $\mu(s,t)$ that does not depend on space or time.
  2. A covariance that depends only on spatial lag h and temporal lag u, not the actual points themselves s and t: $\text{Cov}\big(\varepsilon(s+h,t+u),\varepsilon(s,t)\big) = \text{Cov}\big(\varepsilon(h,u)\big)$

- **Isotropy**: An isotropic spatio-temporal process is one where spatial distance matters, but spatial direction does not.

- **Separability**: A separable spatio-temporal process is one where the covariance in space is independent of the covariance in time.
  - Mantel Test, H_0: Space is independent of time. H_A: Dependency exists.

LEARNING OBJECTIVES

- Define the three different types of spatial data.
- Create weight matrices.
- Describe the MAUP.
- Identify three common difficulties when working with spatiotemporal data.
- Describe three strategies for modeling with spatiotemporal data.

REFERENCES

- This lecture draws heavily from Peter Craigmile's lectures. Peter is a professor of statistics at The Ohio State University and his lecture notes on spatial and spatiotemporal statistics can be found here:
  - http://www.stat.osu.edu/~pfc/teaching/Lyon/notes/5_spatio-temporal.pdf
  - https://www.stat.osu.edu/~pfc/teaching/5012_spatial_statistics/