Spatial Statistics

PM569 Lecture 1: Introduction

Meredith Franklin

Division of Biostatistics, University of Southern California

August 24th, 2018

Course Details

- Meeting time/place: Fridays 10a-1p, Soto 117
- Last class: November 30th
- Final "Exam" is a project. Presentations will be during exam period (Friday, December 7th).
- Prerequisites: PM 511a (or equivalent intro statistics and regression modeling)
- Grading: 6 assignments (10% each), 1 individual final project (35%)
- Grading: Weekly reading assignments and discussion (5%)

Course Details

- **Assignments** must be submitted individually but you can discuss with others. No copying! Late penalty of 20% per day.
- The **final project** will also be done individually. There are several components to the project: deciding a suitable topic (making sure you have some data), writing a brief proposal, making a presentation, and submitting a final paper that details your analysis.

Course Details

From Syllabus, the major topics are:

- Geostatistical data analysis (point referenced), including variograms, kriging, spatial regression.
- Areal data analysis (block or polygon referenced), including Moran's I, SAR, spatial lag and CAR regression models.
- Point pattern data analysis (points with no attributes), including Poisson processes, cluster methods.

Visualization will be a major component throughout the semester. We will also do a "workshop" lecture in ArcGIS.

Course Details

Assignment 0

- We will be using R *a lot* in this course
- Download and install R (`http://www.r-project.org/`)
- You may want to use a nice IDE called RStudio (`http://www.rstudio.com/`)
- Install packages maps, proj4, ggplot2, ggmap, geoR to start
- Work through Lecture1Intro.R and repeat for a new location (we will step through this at the end of lecture). Try an international location and a domestic one! Try different basemaps, too.

Course Goals

- Through this course you will gain an understanding of what are spatial data, and how to work with these data.
- The key components of spatial analysis and statistics are:
  - **Visualization**. "A major pleasure in working with spatial data is their visualization. Maps are amongst the most compelling graphics, because the space they map is the space we think we live in, and maps can show things we cannot see otherwise". Bivand, Pebesma, Gomez-Rubio, Applied Spatial Data Analysis with R, Springer 2008.
  - **Exploration**. This involves looking for patterns in data such as clusters and the behavior of events that are close in space or very distant.
  - **Modeling**. We incorporate what we learn from data visualization and exploration into a formal statistical setting. This allows for estimation and inference.

Historical examples of spatial analysis

John Snow: Early spatial analysis

- In August 1854 there was a major Cholera outbreak in the Soho neighbourhood of London, UK. There were 127 cholera related deaths around the area.
- At the time, germ theory (microorganisms causing disease) was not really known.
- Dr. John Snow spoke to local residents and mapped where cholera cases occurred. As a result of his map, he was able to pinpoint the public water pump on Broad Street as the source of contaminated water causing the cholera outbreak.

Historical examples of spatial analysis

Historical examples of spatial analysis

John Snow: Early spatial analysis

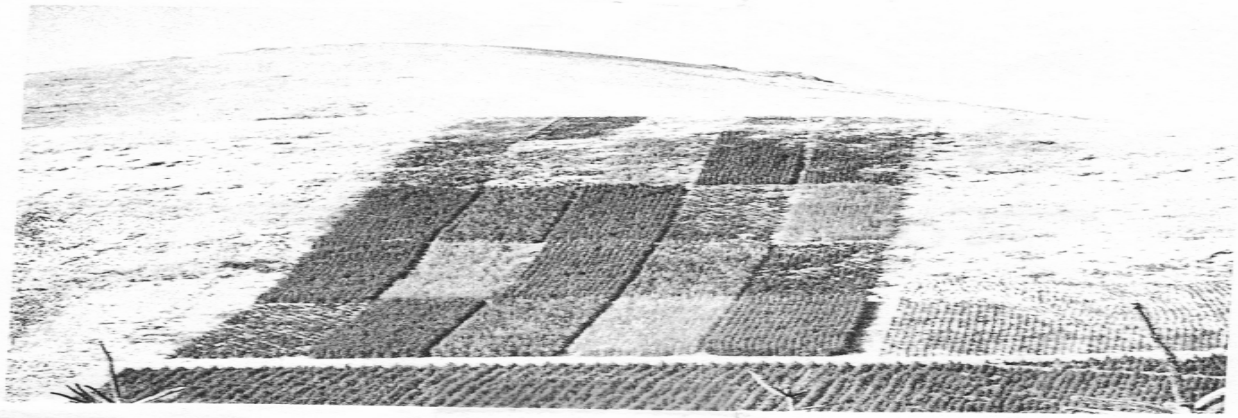
- Dr. Snow used statistics to find a relationship between water quality and cholera cases.
- He found that the waterworks company supplying water to Broad Street pump was taking water from the sewage polluted area of the Thames river.

Historical examples of spatial analysis

R.A. Fisher: Early spatial analysis

- R.A. Fisher was probably the first to recognize the implications of spatial dependence.
- In his work on design of experiments in agricultural science, he wrote (Fisher, 1935, p. 66):
  After choosing the area we usually have no guidance beyond the widely verified fact that patches in close proximity are commonly more alike, as judged by the yield of crops, than those which are further apart.
- Spatial variability, i.e. plot-to-plot variability, was largely due to physical properties of the soil and environmental properties of the field. He avoided the confounding of treatment effect with plot effect with the introduction of randomization
- Basically his solution was to eliminate spatial dependence by localizing into blocks.

Historical examples of spatial analysis



BETTGELERT FOREST
LAYOUT

N

Elevation:					
1730—1800′	*B*	*A*	*E*	*D*	*C*
1530—1730′	*C*	*E*	*B*	*A*	*D*
1460—1590′	*A*	*C*	*D*	*E*	*B*
1340—1460:	*D*	*B*	*A*	*C*	*E*
1250—1340′	*E*	*D*	*C*	*B*	*A*

A. Sitka spruce
B. Japanese larch
C. Sitka spruce/Japanese larch 50/50
D. Sitka spruce/Pinus contorta 50/50
E. Norway spruce/European larch 50/50

Two rows of Beech planted on each side of the series.

Plate 7. Layout of Bettgelert Experiment.

Spatial analyses: past and present

The historical examples were spatial analyses "by chance". Today, spatial data are seemingly everywhere. Not only do we see countless examples of spatial data in a variety of fields of research (public health, economics, sociology, earth sciences, etc.), we often find them used in news articles, apps and pop culture.
Let's first explain what is a spatial analysis and what are spatial data.

Introduction to spatial analysis: general concepts

What is spatial analysis?

- The quantification of phenomena referenced in space.
- The study of methods to describe and explain a process that operates in space based on a sample of observations taken at particular locations.
- Quantitative spatial analysis: Methods
- **Visualization**
  Maps, graphical display
- **Exploration**
  Tools to broadly look at spatial patterns
- **Modeling**
  Fitting models, testing hypothesis, formalizing spatial dependence

Introduction to spatial analysis: general concepts

What is spatial analysis?

- The quantification of phenomena referenced in space.
- The study of methods to describe and explain a process that operates in space based on a sample of observations taken at particular locations.
- Quantitative spatial analysis: Methods
- **Visualization**
  Maps, graphical display
- **Exploration**
  Tools to broadly look at spatial patterns
- **Modeling**
  Fitting models, testing hypothesis, formalizing spatial dependence

Introduction to spatial analysis: general concepts

What is spatial analysis?

- The quantification of phenomena referenced in space.
- The study of methods to describe and explain a process that operates in space based on a sample of observations taken at particular locations.
- Quantitative spatial analysis: Methods
- **Visualization**
  Maps, graphical display
- **Exploration**
  Tools to broadly look at spatial patterns
- **Modeling**
  Fitting models, testing hypothesis, formalizing spatial dependence

Introduction to spatial analysis: general concepts

What is spatial analysis?

- The quantification of phenomena referenced in space.
- The study of methods to describe and explain a process that operates in space based on a sample of observations taken at particular locations.
- Quantitative spatial analysis: Methods
- **Visualization**
  Maps, graphical display
- **Exploration**
  Tools to broadly look at spatial patterns
- **Modeling**
  Fitting models, testing hypothesis, formalizing spatial dependence

Spatial Data

What are spatial data?

- Data that are location specific and that vary in space
- Referenced by a spatial location, s where $s = (x, y)$; x is longitude (easting) and y is latitude (northing). We typically have multiple spatial locations to examine, and they would be referenced as $s_1, s_2, ..., s_n = (x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$.
- Spatial locations may alternatively be referenced by an area such as a zip code, county, state. In such cases, we might use the centroid of the area as a reference point, and define the area and boundaries that encompass the area.
- In terms of thinking about how to approach the analysis of spatial data, the motto taken is:
- **Data that are close together in space (time) are often more alike than those that are far apart.**
  Often labeled as Tobler's first law of geography - "everything is related to everything else, but near things are more related than distant things".

Spatial Data

What are spatial data?

- Data that are location specific and that vary in space
- Referenced by a spatial location, s where $s = (x, y)$; x is longitude (easting) and y is latitude (northing). We typically have multiple spatial locations to examine, and they would be referenced as $s_1, s_2, ..., s_n = (x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$.
- Spatial locations may alternatively be referenced by an area such as a zip code, county, state. In such cases, we might use the centroid of the area as a reference point, and define the area and boundaries that encompass the area.
- In terms of thinking about how to approach the analysis of spatial data, the motto taken is:
- **Data that are close together in space (time) are often more alike than those that are far apart.**
  Often labeled as Tobler's first law of geography - "everything is related to everything else, but near things are more related than distant things".

Types of Spatial Data

We can dissect the broad definition of spatial data into three sub-categories.

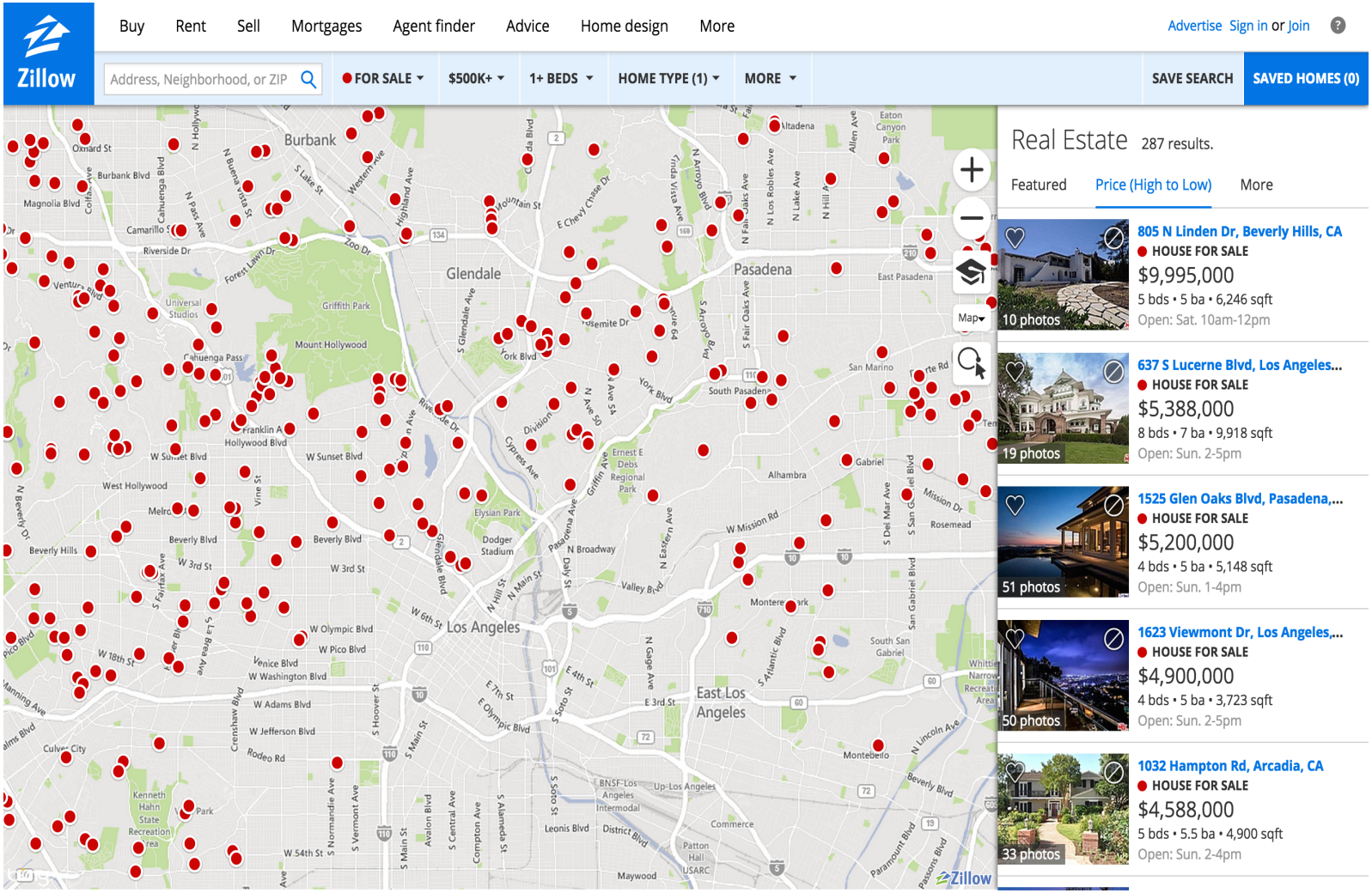
- **Point process data** - **Geostatistical data**
- **Areal data** (sometimes called aggregate data or lattice data)
- **Point pattern data**

There are specific statistical analyses for each of these types of spatial data. We will go into specifics of each.

Types of Spatial Data

GEOSTATISTICAL DATA (POINT PROCESS)

Geostatistical Data: Example



https://www.zillow.com/research/data/

Geostatistical Data: Description

Data that vary continuously over space, but measured only at discrete locations

Examples:

- housing prices in a metropolitan area
- field observations such as soil samples, air pollution measurements (environmental exposures)
- meteorological and climate data

The common thread that links the data is a random process (also called stochastic process or random field)

$$Z(\mathbf{s}) : \mathbf{s} \in D$$

where D is a domain in $\mathbb{R}^d$ (d typically 2)

Visualizing Geostatistical Data

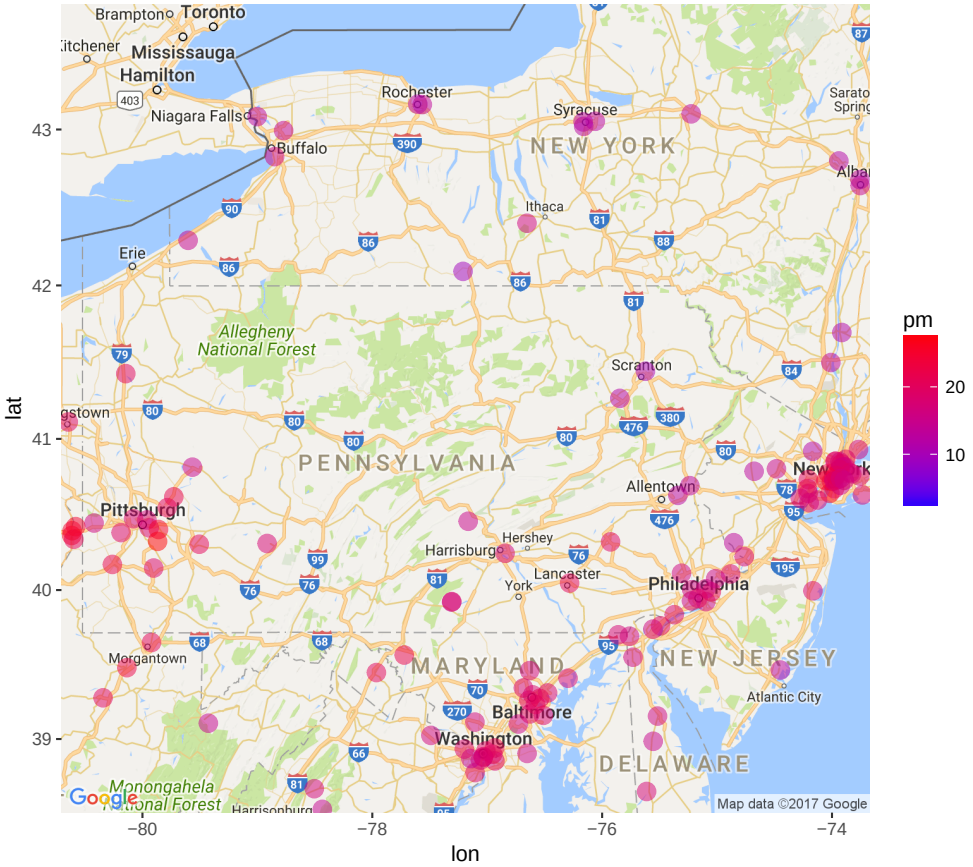
- The components of geostatistical data are the locations (s_i), and the measurements at each location (z_s)
- The best way to visualize these data is to display on a map, and differentiate the values of the measurements of interest by colour or size.

Let's examine field observations of air pollution measurements in the northeast US.

Visualizing Geostatistical Data

We display:

1. the points, which are air pollution monitors
2. the monthly average $PM_{2.5}$ concentration colour coded using a gradient from blue (low) to red (high)
3. a base map (here Google maps), which helps provide a frame of reference

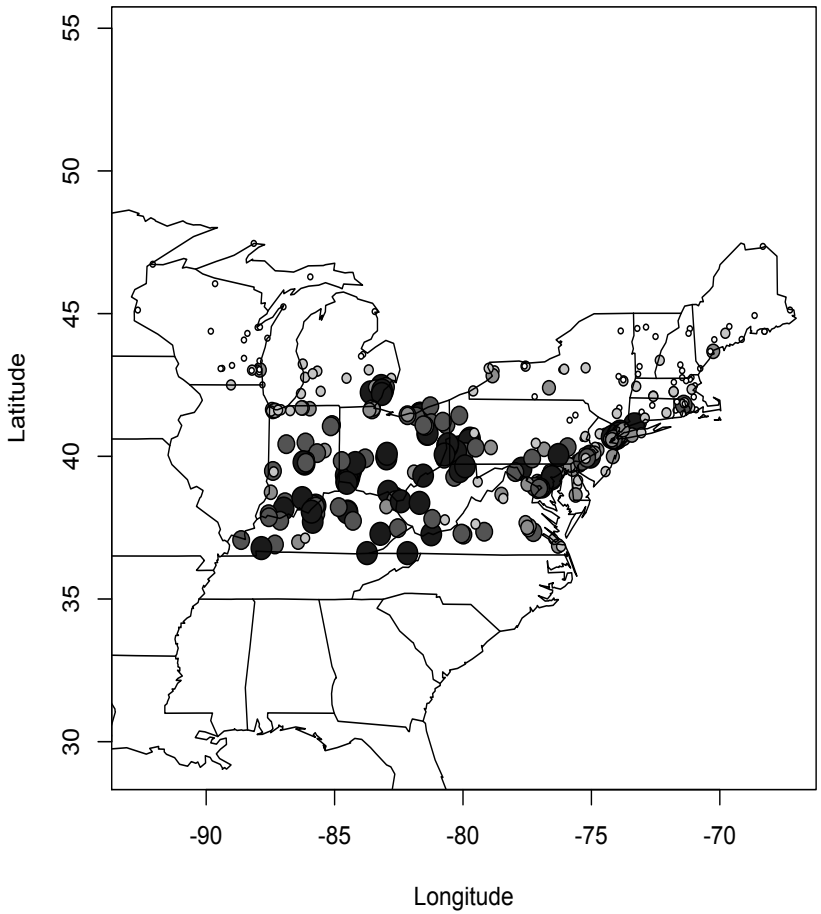


Visualizing Geostatistical Data

Alternatively we can display the points as gradients in size:

1. the points are air pollution monitors
2. the monthly average $PM_{2.5}$ concentration where larger circles represent higher concentrations, smaller circles are lower concentrations

* Note the choice of colour and size gradient of the points can lead do different conclusions!



Exploring and Modeling Geostatistical Data

Goals of spatial statistics applied to geostatistical (point referenced) data

- Explore the spatial pattern in the observations. (Often called spatial "structure").
- Quantify the spatial pattern with a function.
- Model the spatial correlation/covariance in the observations.
- Make predictions at unobserved locations: interpolation, smoothing.

Additional considerations:

- Account for spatial structure in regression models.
- Test a null hypothesis of no spatial structure.

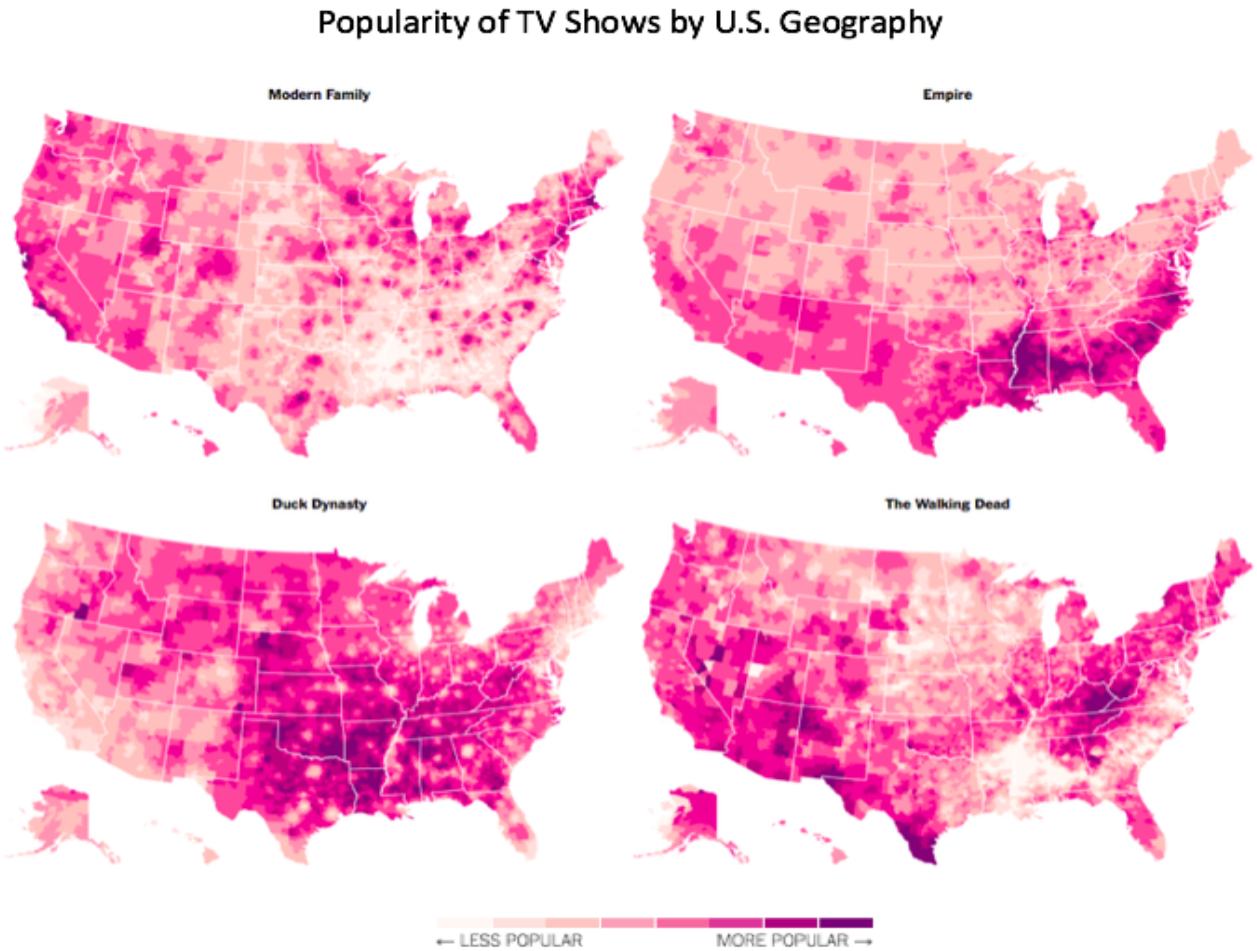
Types of Spatial Data

AREAL DATA

Areal Data: Example 1

http://maps.latimes.com/neighborhoods/

Areal Data: Example 2



https://www.nytimes.com/interactive/2016/12/26/upshot/duck-dynasty-vs-modern-family-television-maps.html

Areal Data: Description

- Data are associated with an area, so are aggregate in nature.
- Areal units tend to be be irregular in shape (e.g. zip code, county) but can be regular grids (e.g. remote sensing data).
- Information collected in areal units may be census related, health related, environmental (satellite estimates of pollution, land cover).
- We want to determine spatial patterns of areal units within a region.
- Areal data (lattices) use neighbour relationships.
- Examples:
  - Median household income in Los Angeles neighborhoods
  - State-specific (or county-, census tract-, zip code-specific) election results
  - County-specific TV viewing preferences
  - County-specific hospital admission rates for a particular disease

Visualizing Areal Data

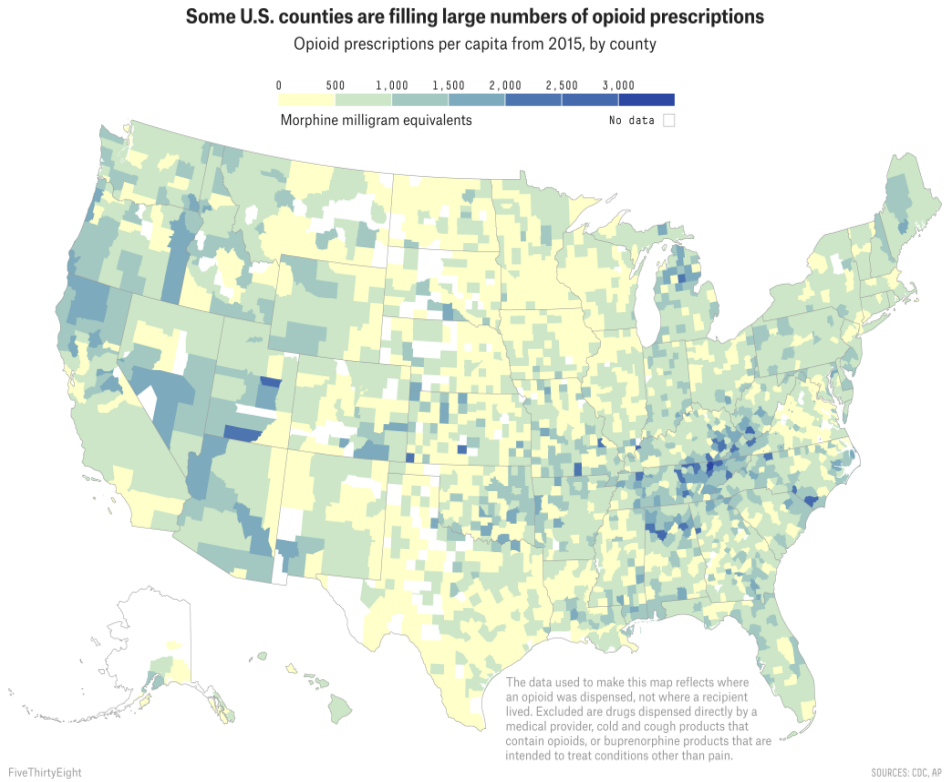
- Areal units are referenced as polygons.
- The centroids of the areal units may be useful for a spatial reference, in combination with the area of the polygon.
- The best way to visualize these data is to display as a map, differentiating the areal units by colour.

Visualizing Areal Data

We display:

1. the areal units (polygons), in this case, counties.
2. the colour representing a quantity, in this case Morphine (mg) categorized into 7 ordinal groups.
3. labels and legends to relate the data to the areal units.

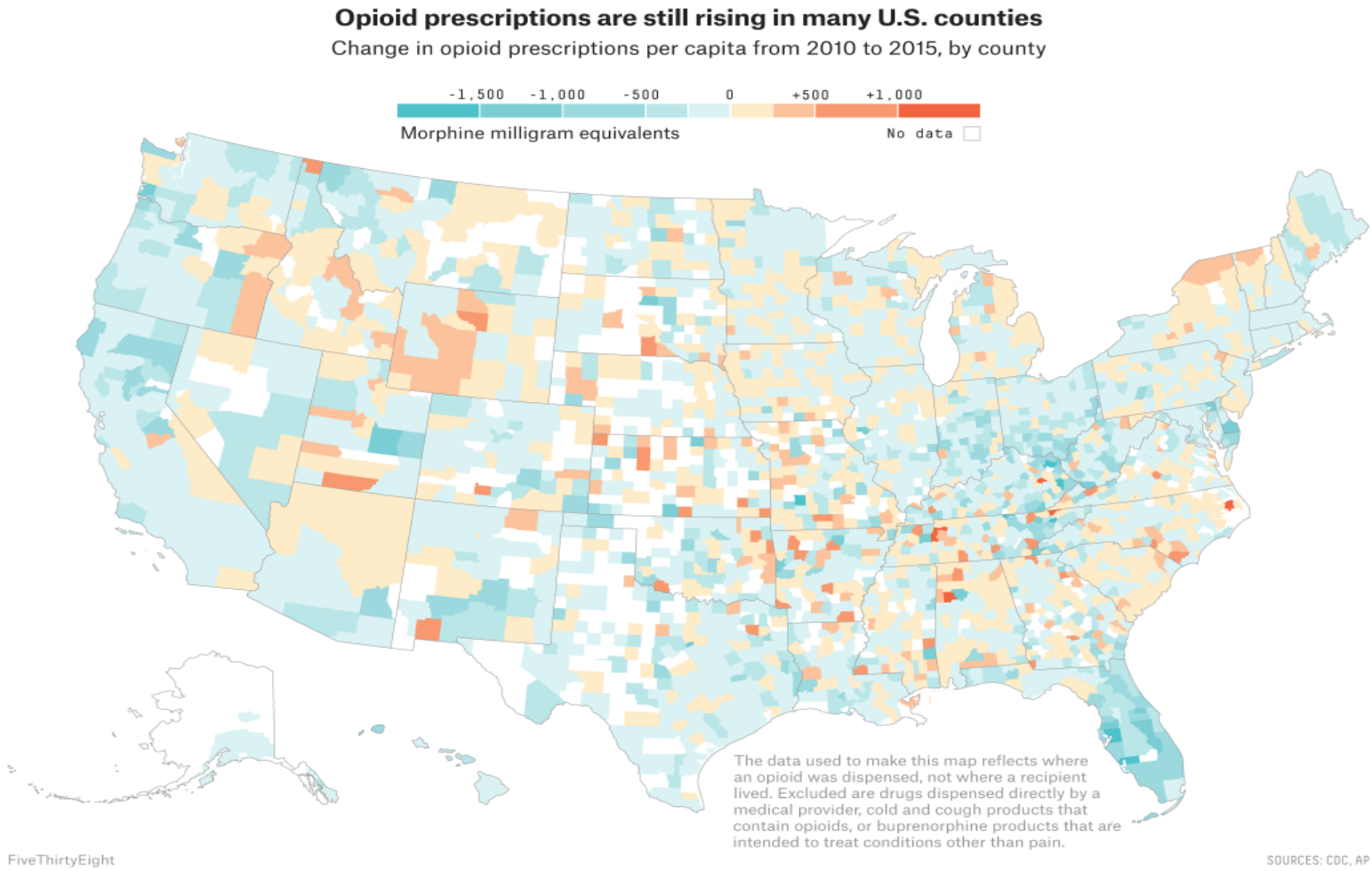
Note that we often look at "per capita" information when examining areal data.



https://fivethirtyeight.com/features/opioid-prescriptions-across-the-u-s/

Visualizing Areal Data

We can look at the data in different ways, such as by changes over time (we will have a similar example later in the semester)



Visualizing Areal Data

Spatial scale is also very important! Let's look at this interactive map: http://www.justicemap.org/

- Smaller areas give us more spatial information.
- Visualizing the tracts versus counties should refine our view of the data.
- Do we expect our map to show the same spatial pattern at different scales?

Visualizing Areal Data

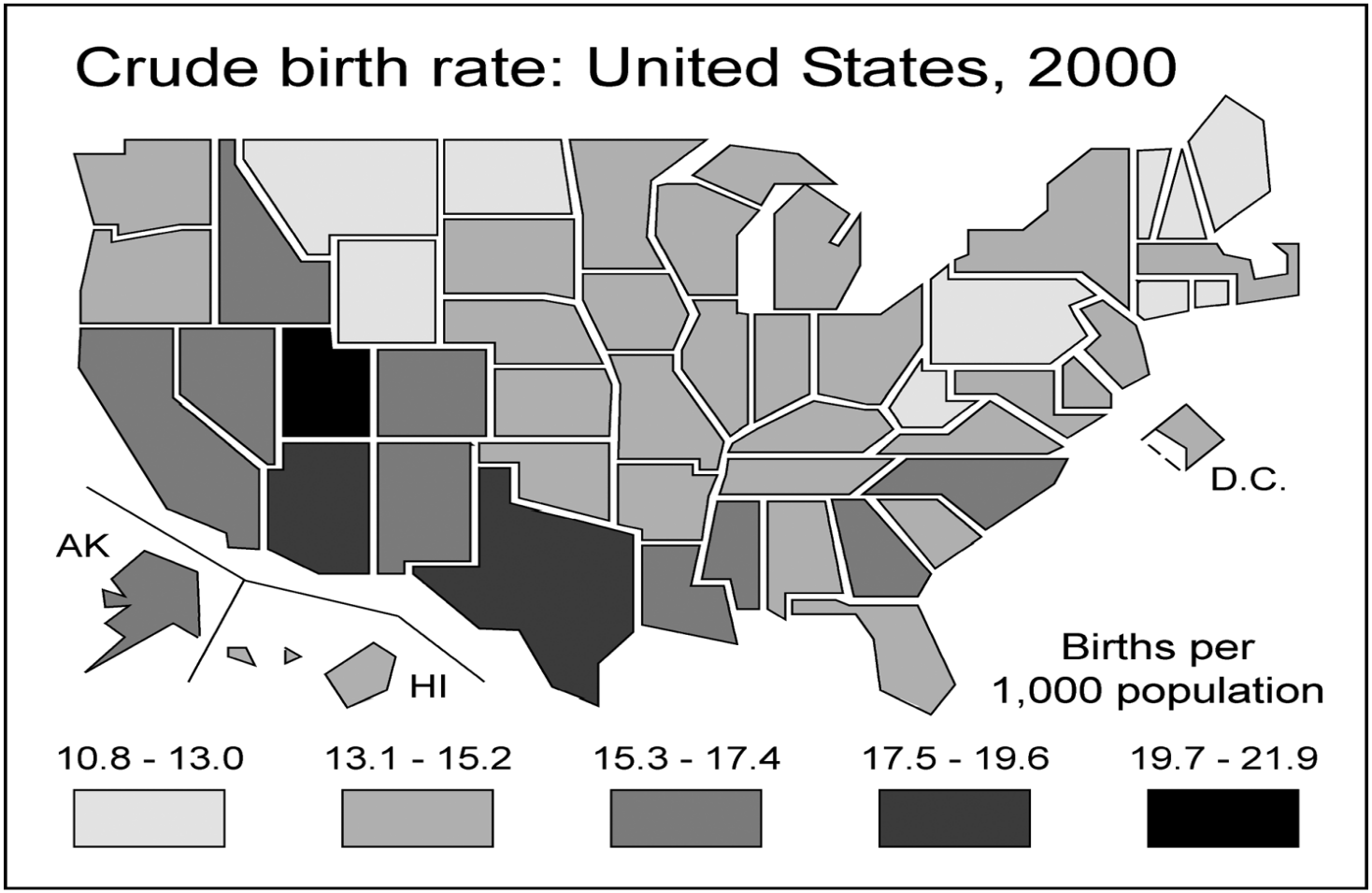
We notice that our visual conclusions can be different with different size areal units of the same thing. There often isn't anything we can do about it, other than to use statistical tools to analyze the data rather than just relying on visual conclusions. There are a couple of names of problems such as this that arise from areal data:

- The "modifiable areal unit problem" (MAUP)
- Ecological fallacy

Another problem we have is in the colour gradient of the areal units. Let's examine mortality rates in the US to illustrate this.

Visualizing Areal Data

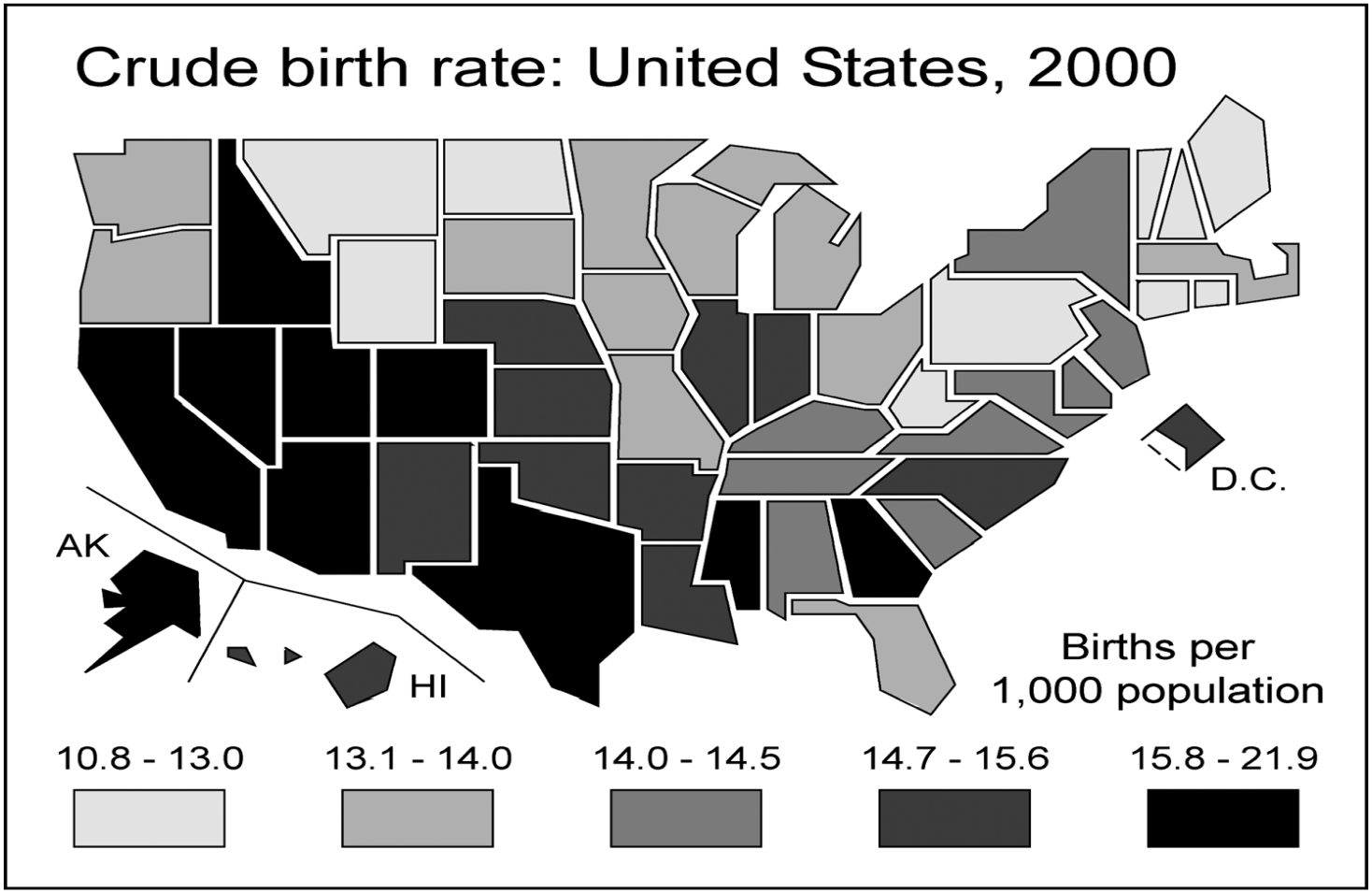
Crude birth rates by state based on equal-interval cut points



Monomier, N. Lying with Maps. Statistical Science 2005, 20(3) 215222.

Visualizing Areal Data

Crude birth rates by state based on quantile cut points



Monomier, N. Lying with Maps. Statistical Science 2005, 20(3) 215222.

Exploring and Modeling Areal Data

Is there a spatial pattern?

- Spatial pattern suggest that observations close to each other have more similar values than those far from each other.
- You might think that there is a pattern through visualization, but this is often subjective.
- Independent measurements will have no pattern, and would look completely random, but there may actually be an underlying pattern.

Goals of spatial statistics applied to areal data

- Understand the linkages between areal units and determine if there if areas that are closer to each other are more related than those that are farther apart.
- If there is a spatial pattern, *how strong is it?*

Additional Considerations

- Incorporate areal unit correlation in regression models (spatial autoregressive models).
- Test the null hypothesis of no spatial autocorrelation.

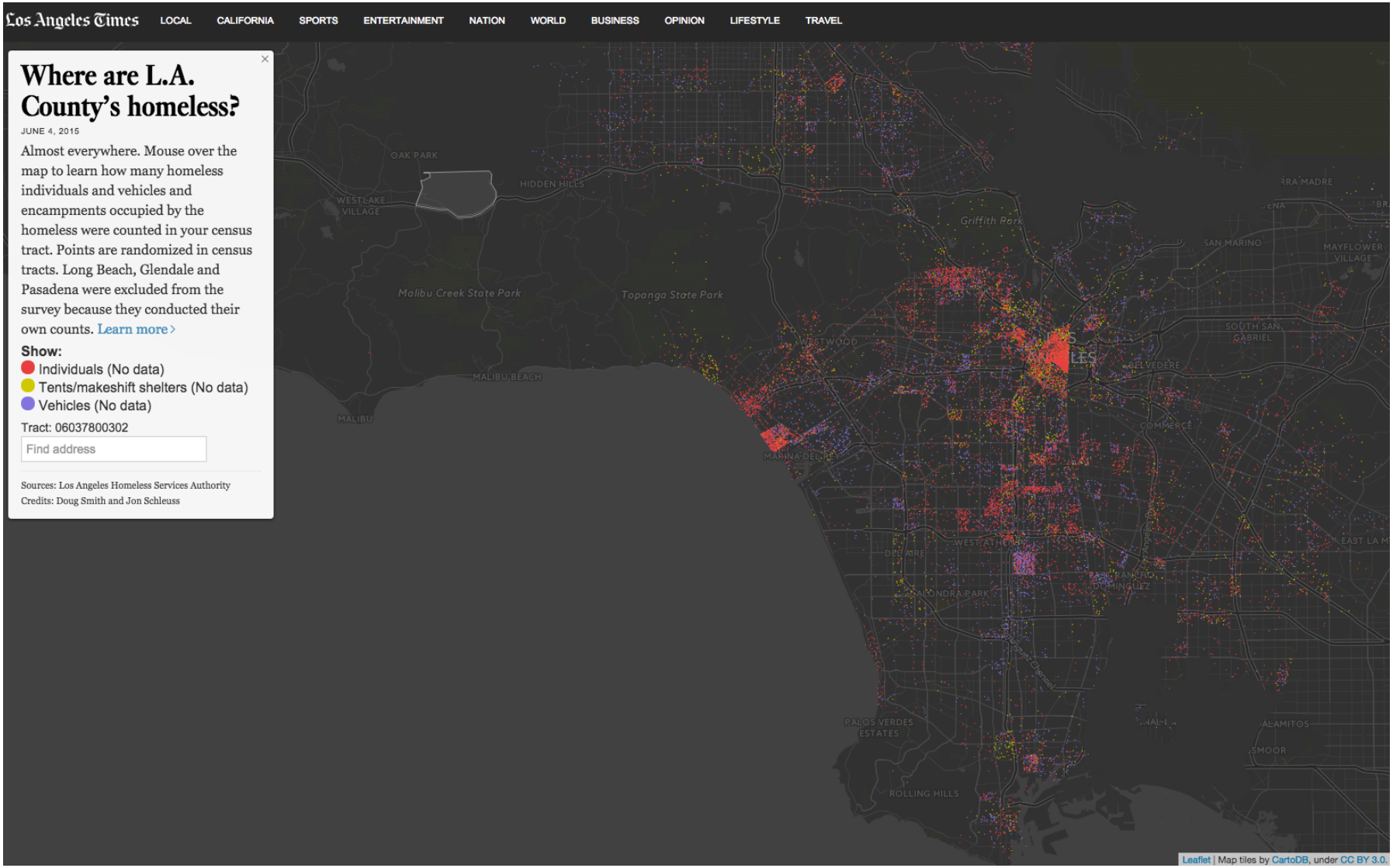
Areal Data

The first reading assignment will be
Mortimer, M (2005) Lying with maps. Statistical Science. 20(3) 215-222.
It is available on the course Blackboard. We'll discuss it next class.

Types of Spatial Data

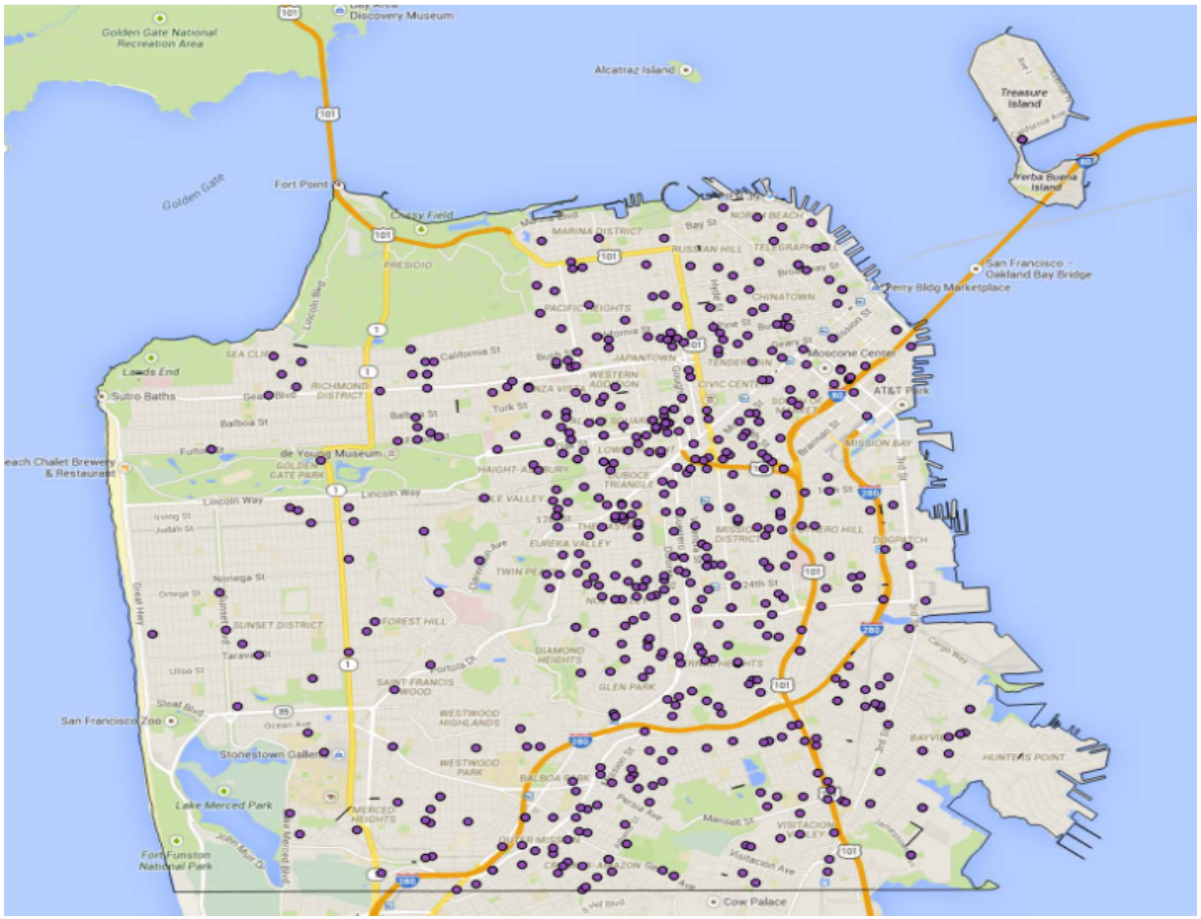
POINT PATTERN DATA

Point Pattern Data: Example 1



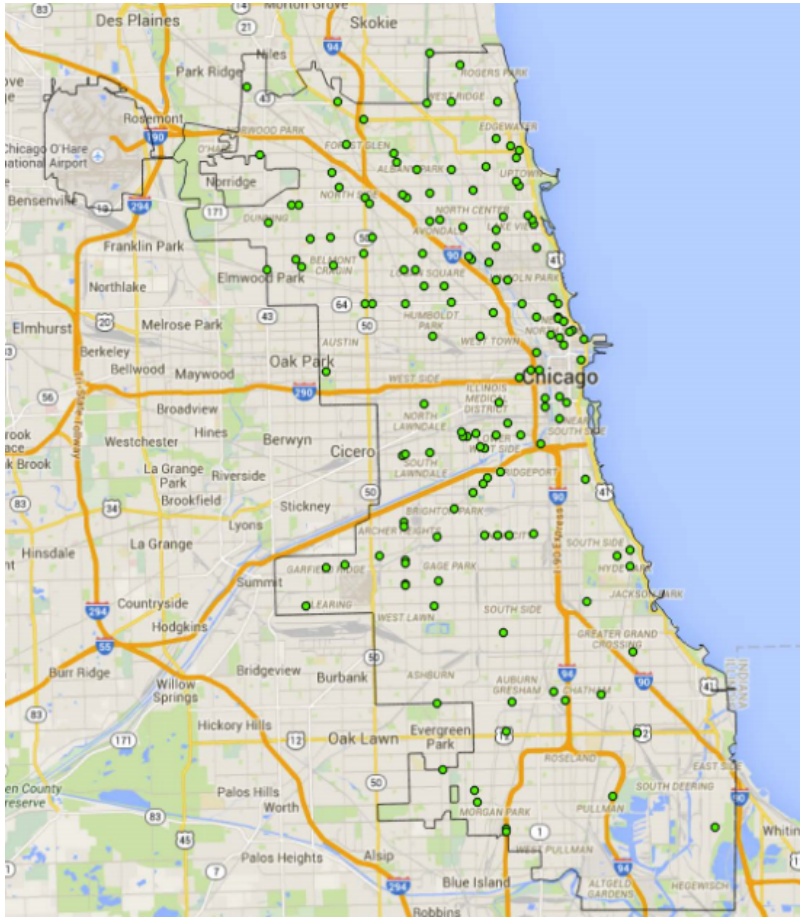
http://graphics.latimes.com/homeless-los-angeles-2015/

Point Pattern Data: Example 2



Car thefts in San Francisco

Point Pattern Data: Example 3



Grocery stores in Chicago

Point Pattern Data: Description

- A spatial point process is a stochastic mechanism that generates events in 2D.
- Event is an observation (presence/absence), point is the location.
- Mapped point pattern: Events in a study area D have been recorded.
- Sampled point pattern: Events are recorded after taking samples in an area D.
- Examples
  - Locations of homeless in Los Angeles
  - Cases of malaria in Kenya
  - Locations of a specific tree species in a forest

Visualizing Point Pattern Data

- The components of point pattern data are the locations (s_i).
- If there are different categories of a point pattern, such as with the homeless data, then these categories may be coloured separately.
- Often conclusions cannot be drawn from visual inspection alone

Exploring and Modeling Point Pattern Data

Questions about point pattern data that we would like to answer are:

- Is there a regular pattern in the points?
- Is there clustering of the points?
- Can we define a point process that our events follow?

There may be additional features that we need to take into account:

- Is there an underlying population distribution from which events arise in a region?
- Are events clustering in areas of high population?
- If there are underlying features that could affect the presence/absence of a point, such as population density, we need to account for this.

Exploring and Modeling Point Pattern Data

- Measure of intensity: mean number of events per unit area
- Are there differences between point process and a simple random process?
- Are points closer together than they would be by chance?
- Are the points more regularly spaced than they would be by chance?

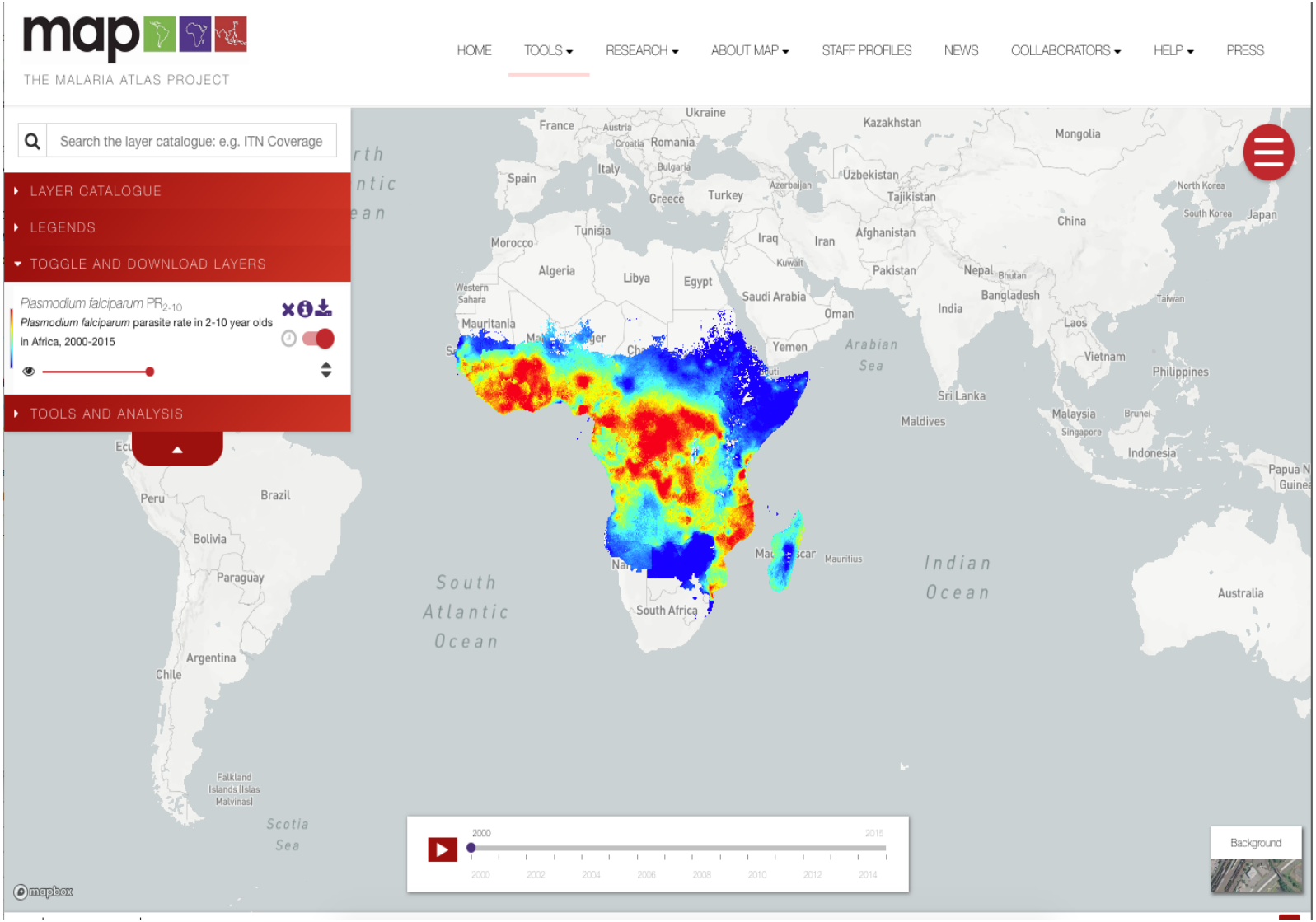
Exploring and Modeling Point Pattern Data

- Spatial location s
- Presence/Absence modeled by $Y, Y(s) = 1$ if there is a case and $Y(s) = 0$ otherwise
- Define a null hypothesis: no pattern (complete spatial randomness)
- Find a statistic to test whether the data is clustered, or regular.
- Model some spatial pattern and determine if our observed point pattern fits this model.

Exploring and Modeling Point Pattern Data

- Density based cluster models can detect clusters of points. These methods bridge into the world of machine learning.
- Density based smoothers (e.g. Gaussian kernels), can smooth out the intensity of the points to create a surface
- Example: mapping cases of Malaria in Africa (the Malaria Atlas Project)

Point Pattern Data: Smoothers



https://map.ox.ac.uk/explorer/

Types of spatial data: spatio-temporal

All three types of data we have described may be referenced in space and in time. That is, data that are location specific can have replicates in time:

- Each observation has a location, time and value
- Similar methods for analysis, with an added dimension
- Often encountered in environmental epidemiology:
  Geostatistical: Relationship between daily air pollution measured at discrete locations in the US Northeast and hospital admissions
  Areal: Changes in birth rates in census tracts in US states (2000 to 2010 census).
  Point process: Changes in spatial clustering of Malaria cases 2000 to 2015.

Review

Multiple Linear Regression
Matrix Representation of MLR
Residual Variance-Covariance Matrix and How Spatially Correlated Data are Represented

Review: Linear Regression

Recall, in non-matrix terms multiple linear regression is represented by:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$$

We have $i = 1, ..., n$ observations, so expanding gives

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} + \epsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p} + \epsilon_2 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_p x_{np} + \epsilon_n \end{aligned}$$

Important assumption of linear regression is that the residuals are identically and independently distributed (iid) from a normal distribution $\epsilon_i \sim N(0, \sigma^2)$.

Linear Regression: Matrix Notation

The matrix form of the linear model is

$$\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\epsilon}$$

Where $\mathbf{y}$ is the vector of responses (dependent variable) and $\mathbf{X}$ is the "design matrix" of p explanatory variables (independent variables). The elements of the regression are:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Linear Regression: Least Squares

- To estimate $\hat{\beta}$, we minimize the sum of squared error:

$$\text{SSE} = \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} +, ..., + \hat{\beta}_p x_{ip}))^2$$

  which is the same as minimizing $\sum_i \epsilon_i^2$

- Take the derivative, set to 0 to get the normal (or score) equations for $\hat{\beta}$.
- In matrix notation, $\hat{\beta} = (X^T X)^{-1} X^T Y$

Linear Regression: Variance-Covariance Matrix

The assumption of iid errors, namely that the residuals $\boldsymbol{\epsilon}$ have mean zero and heteroscedastic variance is expressed in matrix form via

$$\boldsymbol{\Sigma} = \text{Var}(\boldsymbol{\epsilon}) = \sigma^2 \mathbf{I}$$

where $\mathbf{I}$ is the identity matrix

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

and

$$\sigma^2 \mathbf{I} = \begin{bmatrix} \sigma^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma^2 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma^2 \end{bmatrix}$$

This is the variance-covariance matrix.

Linear Regression: Variance-Covariance Matrix

The variance parameter σ^2 is estimated using the unbiased estimator s^2

$$s^2 = \frac{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{n - p}$$

where p is the number of parameters in the regression model.
Under the assumption that $E(\boldsymbol{\epsilon}) = 0$, $\hat{\boldsymbol{\beta}}$ is an unbiased estimator, i.e. $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$.
Under the assumption that the residuals are uncorrelated with homogeneous variance, $\text{Var}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$, the variance-covariance matrix for $\hat{\boldsymbol{\beta}}$ is given by $\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$.

Linear Regression: Variance-Covariance Matrix

- The ordinary least squares (OLS) estimators of our regression parameters are unbiased (and the confidence intervals on the estimates are correct) when the model is correctly specified. Our covariates correctly specify the model and the iid assumption of $\boldsymbol{\epsilon}$ is met.
- What if the assumptions fail?
  - The variance-covariance matrix is not $\sigma^2\mathbf{I}$
  - There is covariance between errors, i.e. $\text{Cov}(\sigma_i, \sigma_j)$
  - The variance-covariance matrix is $\boldsymbol{\Sigma} = \text{Var}(\sigma) = \sigma^2\mathbf{V}$ where $\mathbf{V}$ is not the identity matrix $\mathbf{I}$, but rather specifies the variance σ^2 and covariance $\text{Cov}(\sigma_i, \sigma_j)$

The regression parameters obtained when $\boldsymbol{\epsilon} \sim N(0, \boldsymbol{\Sigma})$ are called the Generalized Least Squares (GLS) estimators. OLS is a special case of GLS.

Linear Regression: Generalized Least Squares

The GLS model equation is in the same form as OLS with the main difference that we must account for covariance with $\mathbf{\Sigma}$. Now our parameter estimates have the form:

$$\hat{\beta} = (X^T \mathbf{V}^{-1} X)^{-1} X^T \mathbf{V}^{-1} Y$$

Where $\mathbf{\Sigma} = \text{Var}(\sigma) = \sigma^2 \mathbf{V}$. Also recall that $\text{Var}(\mathbf{y}) = \mathbf{\Sigma}$.
We often also express the GLS model as $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{\Sigma})$.

In this course we spend a lot of time specifying $\mathbf{\Sigma}$, the spatial variance-covariance matrix, using a variety of methods for a variety of spatial data types.