Spatial Data Analysis

Week 1: Introduction

Meredith Franklin

Department of Statistical Sciences and School of the Environment

September 5th, 2025

Course Details

- Meeting time/place: Fridays 10 am - 1 pm, Victoria College 115
- Assignment and project based course, no exams.
- Prerequisites are for undergrads only: STA302, STA303
- Knowledge for graduate students: experience with regression, some coding R, Python

Course Details

- **Assignments** must be submitted individually but you can discuss with others. No copying! Late penalty of 10% 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 (midterm), submitting a final paper that details your analysis and findings, making a short recorded presentation.

Course Details

From the syllabus, the major topics are:

- Overview of spatial data types, working with spatial data, making maps.
- 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.
- Spatial and spatiotemporal machine learning methods.

Visualization will be a major component throughout the semester. We will do this in R but may also check out QGIS (open source GIS).

Course Details

Assignment 0

- We will mostly be using R in this course, and Python can be used if you would like.
- Download and install R (`http://www.r-project.org/`) and RStudio (`http://www.rstudio.com/`)
- You can also work through UofT's JupyterHub (`https://live-datatools.pantheonsite.io/`)
- Install packages `leaflet`, `tidyverse`, `ggmap` to start.
- Work through spatial1.Rmd and add a tab that repeats a map for a new location (we will step through this in class). Try an international location and a domestic one! Try different basemaps, too.

Spatial Analysis Objectives

- The key components of spatial analysis and spatial 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.
  - **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 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.

Spatial analyses: past and present

- Historical examples of spatial data were often stumbled upon "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

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

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

What are spatial data?

- Data that are location specific and that vary in space
- Referenced by a spatial location, s where $s = (x, y)$ are cartesian coordinates; often 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".

What are spatial data?

- Data that are location specific and that vary in space
- Referenced by a spatial location, s where $s = (x, y)$ are cartesian coordinates; often 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.

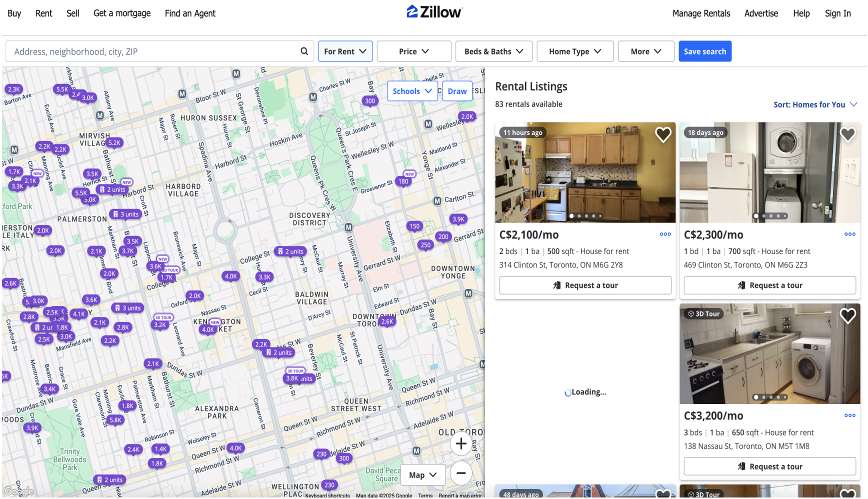
- **Geostatistical data**
- **Areal data** (sometimes called aggregate data or lattice data)
- **Point pattern data**

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

Types of Spatial Data

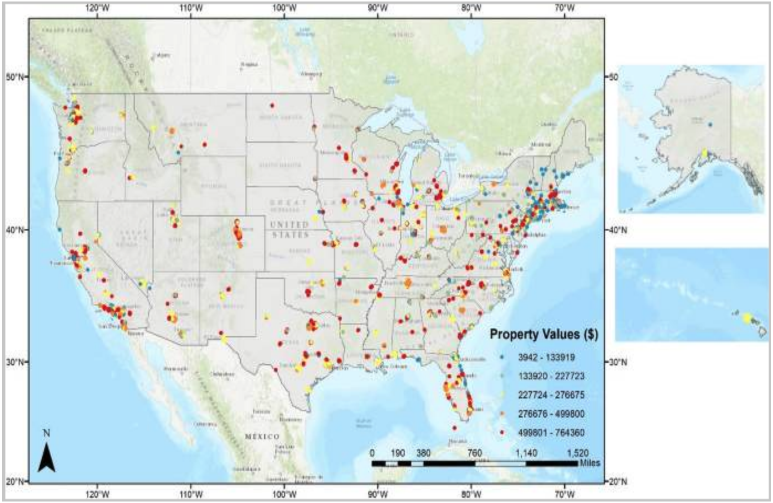
GEOSTATISTICAL DATA

Geostatistical Data: Example



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

Geostatistical Data: Example



Optimization of United States Residential Real Estate Investment through Geospatial Analysis and Market Timing, Gale and Roy, 2023

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, noise levels (environmental data)
- plant, animal abundance or disease prevalence
- 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

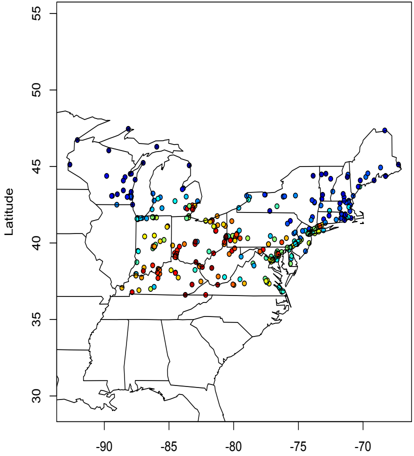
- The components of geostatistical data are the locations (s_i), and the measurements at each location (z_s)
- The best way to visualize these data is to display on a map, and differentiate the values of the measurements of interest by colour or size.

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)

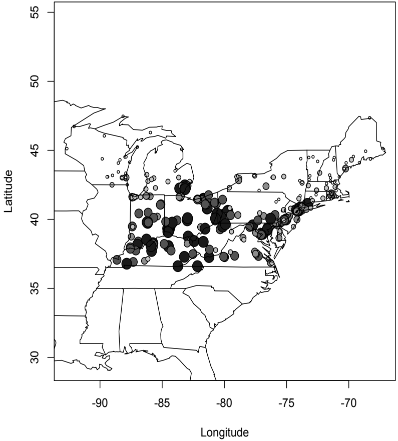


Visualizing Geostatistical Data

Alternatively we can display the points as gradients in size:

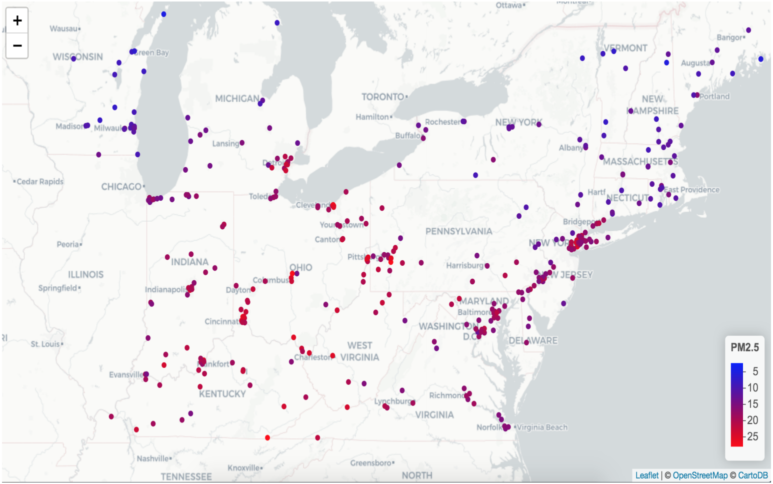
1. the points are air pollution monitors
2. the monthly average $PM_{2.5}$ concentration where larger circles represent higher concentrations, smaller circles are lower concentrations

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



Visualizing Geostatistical Data

- Adding a basemap helps provide a spatial reference frame
- Adding a legend helps provide context for what is being mapped



Visualizing Geostatistical Data in R

- The leaflet package is very useful for generating interactive maps
- Let's build a couple of maps for
  1. northeast US/Canada city populations
  2. Air Quality data during a wildfire in California
  3. Visualizing cycling fatalities using Toronto Open Data

Exploring and Modeling Geostatistical Data

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

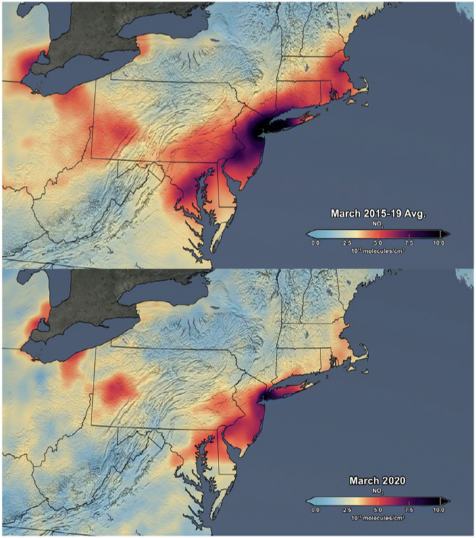
- Explore the spatial pattern in the observations. (Often called spatial "structure").
- Quantify the spatial pattern with a function.
- Model the spatial correlation/covariance in the observations.
- Make predictions at unobserved locations: interpolation, smoothing.

Additional considerations:

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

Exploring and Modeling Geostatistical Data

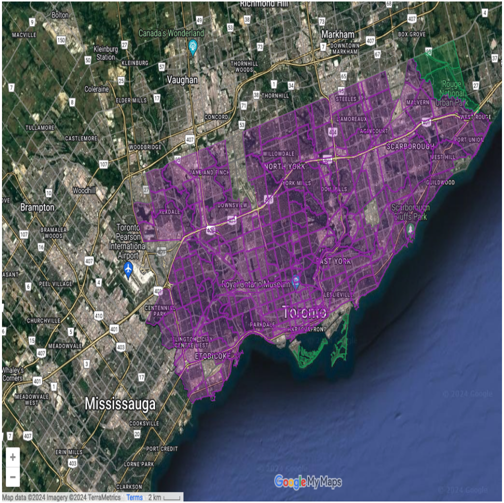
For example, interpolating the pollution data from the point measurements can give a map like this



Types of Spatial Data

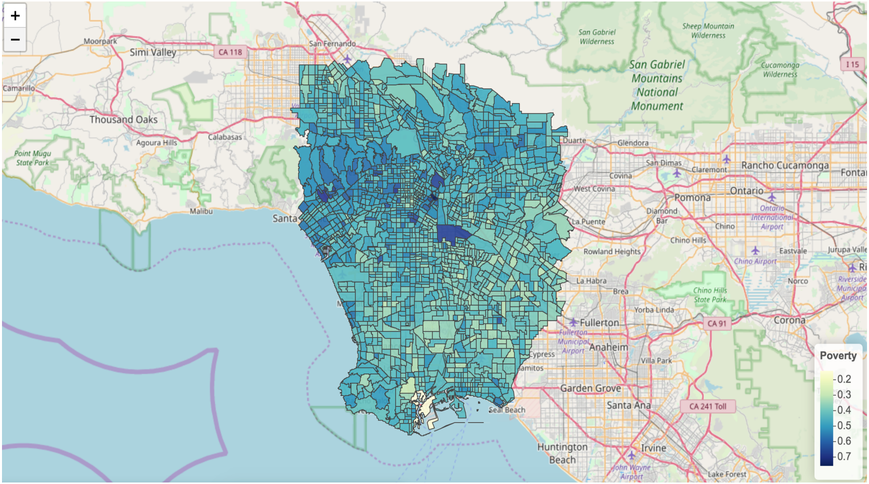
AREAL DATA

Areal Data: Example 1



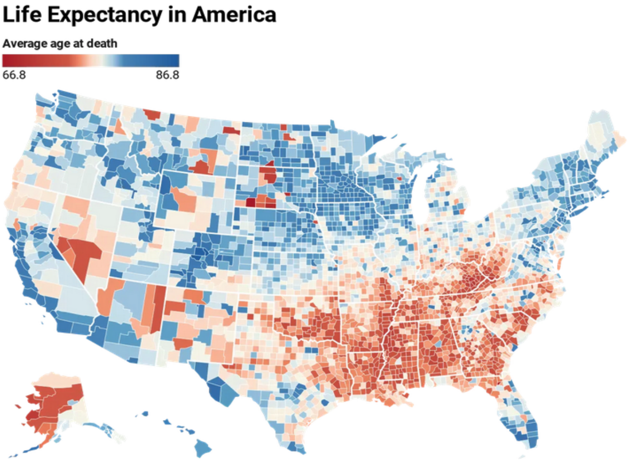
Neighborhood (purple), industrial (grey), parkland (green) https://metroscapes.ca/toronto/projects/neighbourhoods/

Areal Data: Example 2



Poverty Rates in LA County

Areal Data: Example 3



Life Expectancy USA https://americaninequality.substack.com/p/life-expectancy-and-inequality

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 Toronto neighborhoods
  - Poverty rates in Los Angeles county census tracts
  - State-, county-, census tract-, postcode-, zipcode-specific census data or election results
  - County-specific hospital admission rates for a particular disease

Visualizing Areal Data

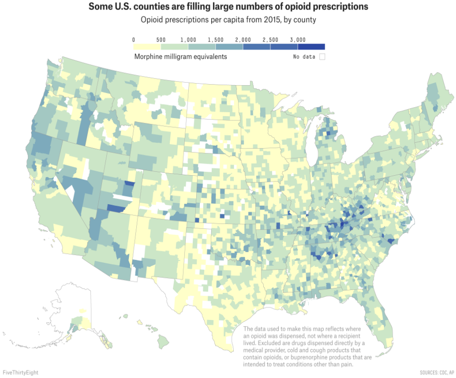
- Areal units are referenced as polygons.
- The centroids of the areal units may be useful for a spatial reference, in combination with the area of the polygon.
- The best way to visualize these data is to display as a map, differentiating the areal units by color.

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.

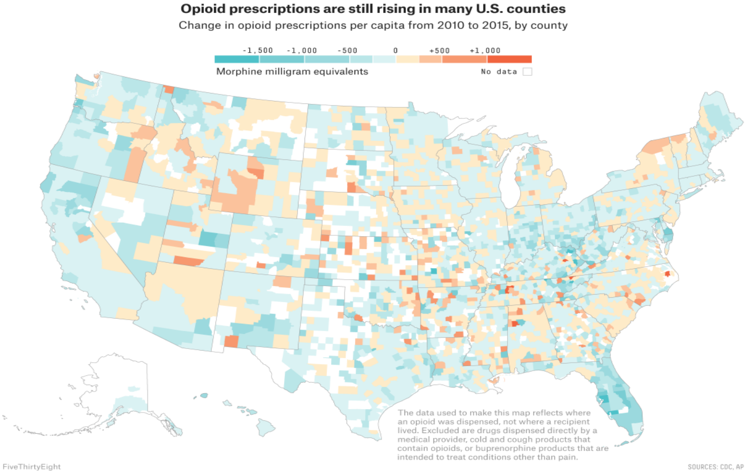
Note that we often look at "per capita" information when examining areal data.



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

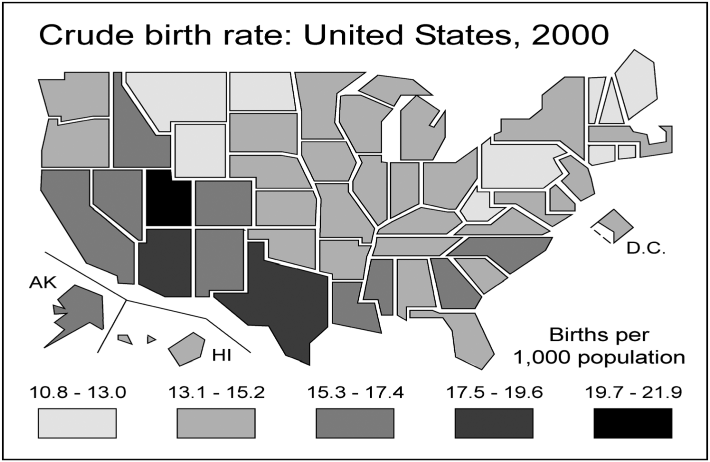
We notice that our visual conclusions can be different with different size areal units of the same thing. There often isn't anything we can do about it, other than to use statistical tools to analyze the data rather than just relying on visual conclusions. There are a couple of names of problems such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

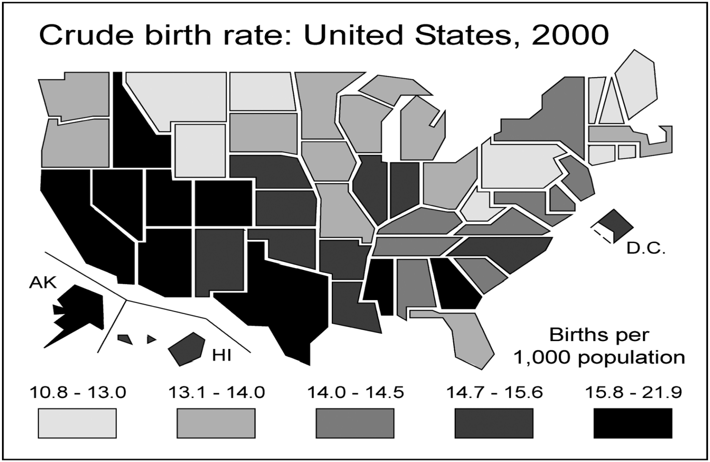
Crude birth rates by state based on equal-interval cut points



Monomier, N. Lying with Maps. Statistical Science 2005, 20(3) 215–222.

Visualizing Areal Data

Crude birth rates by state based on quantile cut points



Monomier, N. Lying with Maps. Statistical Science 2005, 20(3) 215–222.

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?*

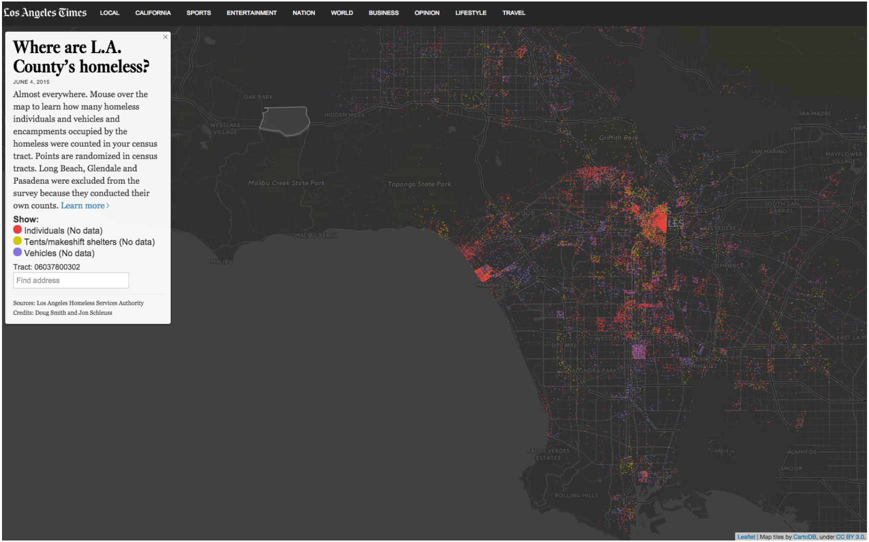
Additional Considerations

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

Types of Spatial Data

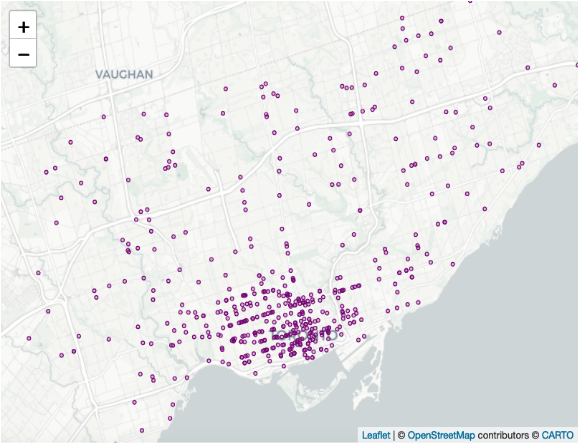
POINT PATTERN DATA

Point Pattern Data: Example 1



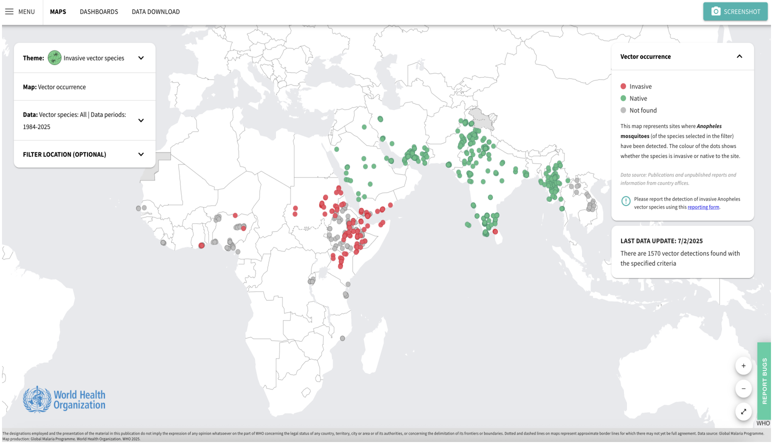
http://graphics.latimes.com/homeless-los-angeles-2015/

Point Pattern Data: Example 2



Cycling Accidents in Toronto, 2024

Point Pattern Data: Example 3



Malaria Map WHO

Point Pattern Data: Description

- A spatial point pattern 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
  - Locations of bicycle thefts in Toronto
  - Global cases of
  - Locations of a specific tree species in a forest

Visualizing Point Pattern Data

- The components of point pattern data are the locations (s_i), again a cartesian reference in x,y coordinates.
- If there are different categories of a point pattern, such as with the homeless data, then these categories may be "marked" 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?

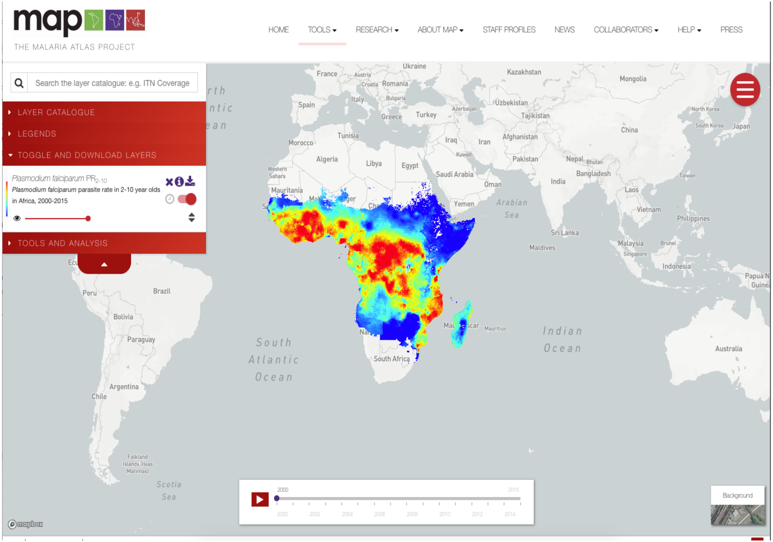
Exploring and Modeling Point Pattern Data

- 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/

The Limits and Intensity of Plasmodium falciparum Transmission: Implications for Malaria Control and Elimination Worldwide

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 2020 census).
  Point pattern: Changes in spatial clustering of Malaria cases 2000 to 2025.

Machine Learning in Spatial Data Analysis

- **Geostatistical data (continuous spatial processes)**
  - Predicting pollutant concentrations, soil nutrients, or temperature at unsampled locations using Random Forests, XGBoost, or neural networks.
  - Enhancing interpolation (e.g., regression-kriging + ML; deep learning for spatiotemporal prediction).
  - Dimension reduction of high-dimensional sensor/monitoring data (e.g., PCA, NMF, autoencoders).
- **Areal data (lattice or polygonal units)**
  - Modeling disease incidence, crime counts, or socioeconomic indicators across regions using ML regression/classification.
  - Integrating spatial autocorrelation into tree-based models or spatially regularized neural networks.
  - Feature extraction from remote sensing to generate covariates for areal units (e.g., land cover proportions).
- **Point pattern data (event locations in space)**
  - Classifying and clustering event locations (e.g., wildfire ignitions, traffic accidents) to detect hotspots.
  - Learning intensity surfaces using kernel methods, Poisson processes augmented with ML predictors, or graph neural networks.
  - Identifying covariate relationships driving point occurrence (e.g., vegetation, proximity to roads).

Linear Regression Including Variance-Covariance

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

Review: Linear Regression Form

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 errors 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), however $\mathbf{X}$ has dimension $n \times (p+1)$ (1 is for the intercept). 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 Matrix Notation

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

$$\text{SSE} = \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})\right)^2$$

  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 homoscedastic variance, is expressed in matrix form via

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

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

$$s^2 = \frac{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{n - p - 1}$$

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. $\mathrm{Cov}(\epsilon_i, \epsilon_j)$
  - The variance-covariance matrix is $\boldsymbol{\Sigma} = \mathrm{Var}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{V}$ where $\mathbf{V}$ is not the identity matrix $\mathbf{I}$, but rather specifies the variance σ^2 and covariances $\mathrm{Cov}(\epsilon_i, \epsilon_j)$

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 $\boldsymbol{\Sigma}$. Now our parameter estimates have the form:

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

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

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