Spatial Statistics

PM599, Fall 2013

Lecture 1: August 26th, 2013

Course Details

- Meeting time/place: Mondays 1-4pm, Soto 303
- No class: September 2 (Labor Day)
- Last class: December 2 (We will have project presentations that day)
- Prerequisites: PM 510, 511a (or equivalent intro statistics and regression modeling)
- Grading: 6 assignments (10% each), 1 final project/presentation small group or individual (40%)

Course Details

- **Assignments** must be submitted individually but you can discuss with others. No copying! Late penalty of 20% per day.
- The **final project** can be done individually or in pairs. Consideration in terms of grading will be given to students wishing to work solo vs with a partner. There will be several components to the project: deciding a suitable topic, writing a brief proposal, making a presentation, and submitting a final paper that details your analysis.

Course Details

General outline of course

- Theory and analysis of geostatistical data (3 lectures, 2 assignments)
- Theory and analysis of areal data (4 lectures, 2 assignments)
- Theory and analysis of point pattern data (3 lectures, 1 assignment)
- Specialized topics (2 lectures, 1 assignment)

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 geoR, lattice, maps, maptools, spdep, spatstat, splancs

Review of non-spatial statistics

Important background concepts that play a role in this course

- Linear Regression: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
- The response variable y_i (predictand) is modeled as a linear function of one or more predictor variables x_i (covariates).
- The error, ϵ_i is associated with the response variable y_i
- The predictor variables are assumed to have no error (ignored, or x_is measured without error)

Matrix Notation

- The general 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).

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{p1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & \cdots & x_{pn} \end{bmatrix}$$

$$\mathbf{y} = [y_1, \cdots, y_n]$$

$$\boldsymbol{\epsilon} = [\epsilon_1, \cdots, \epsilon_n]$$

$$\beta = [b_0, b_1, \cdots, b_n]$$

Review: Linear Regression

OLS

- To estimate $\hat{\beta}$, we minimize the squared error $\sum_i \epsilon_i^2$
- Take the derivative, set to 0 to get the normal (or score) equations for $\hat{\beta}$
- In matrix notation, $\hat{\beta}_{OLS} = (X'X)^{-1}X'Y$

Review: Linear Regression

Important Assumptions

- The residuals are identically and independently distributed (iid) from a normal distribution $N \sim (0, \sigma^2)$
- In matrix notation, the residuals have mean zero and variance-covariance matrix $\Sigma = var(\sigma) = \sigma^2 I$ where I is the identity matrix.
- The residuals have constant variance (homoscedastic)

Review: Linear Regression

Important Assumptions

- The 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 previous assumptions are met.
- What if the assumptions fail?
- The variance-covariance matrix is not $\sigma^2 I$
- There is covariance between errors, i.e. $Cov(\sigma_i, \sigma_j)$
- Our variance-covariance matrix is $\Sigma = var(\sigma) = \sigma^2 \mathbf{V}$

Review: Covariance

- Covariance measures linear associations.

$$Var(X) = Cov(X, X) = E[(X - E(X))^2] = E(X^2) - (E(X))^2$$
$$Cov(X, Y) = E[(X - E(X))(Y - E(Y))]$$
$$Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$$

Review: Linear Regression

Generalized Least Squares

- We need maximum likelihood methods to provide estimates for generating the score equations for $\hat{\beta}$
- In matrix notation $\hat{\beta}_{GLS} = (X'V^{-1}X)^{-1}X'V^{-1}Y$
- Familiar with MLE? We will go over in more detail when we talk about spatial regression.

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

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

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

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).
- May also be referenced by an area such as a zip code, county, state.
- **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".

John Snow: Early spatial analysis

- In August 1854 there was a major Cholera outbreak in the Soho neighbourhood 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.

Introduction to spatial analysis

Historical examples

John Snow: Early spatial analysis

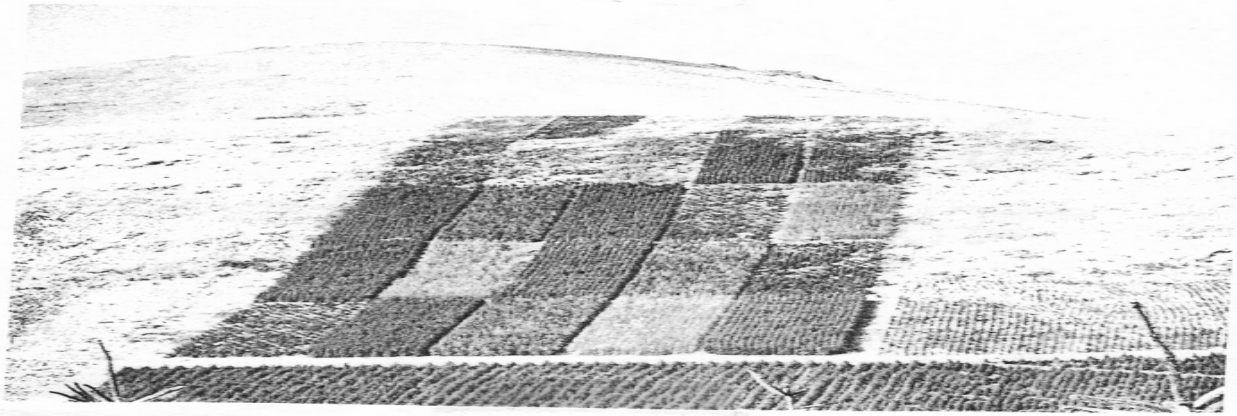
- Dr. Snow used statistics to find a relationship between water quality and cholera cases.
- He found that the waterworks company supplying water to Broad Street pump was taking water from the sewage polluted area of the Thames river.

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.

Introduction to spatial analysis

Historical examples



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 Data

Types of spatial data

- Geostatistical (point referenced)
- Areal (lattice)
- Point process
- References:
  N. Cressie *Statistics for Spatial Data* (1993).
  L.A. Waller and C.A. Gotway *Applied Spatial Statistics for Public Health Data* (2004).

Spatio-temporal data: All three types

- Data that are location specific but replicated 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: Examining birth rates from year to year in US states
  Point process: Changes in spatial clustering of cholera in Kenya over time

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

Examples:

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

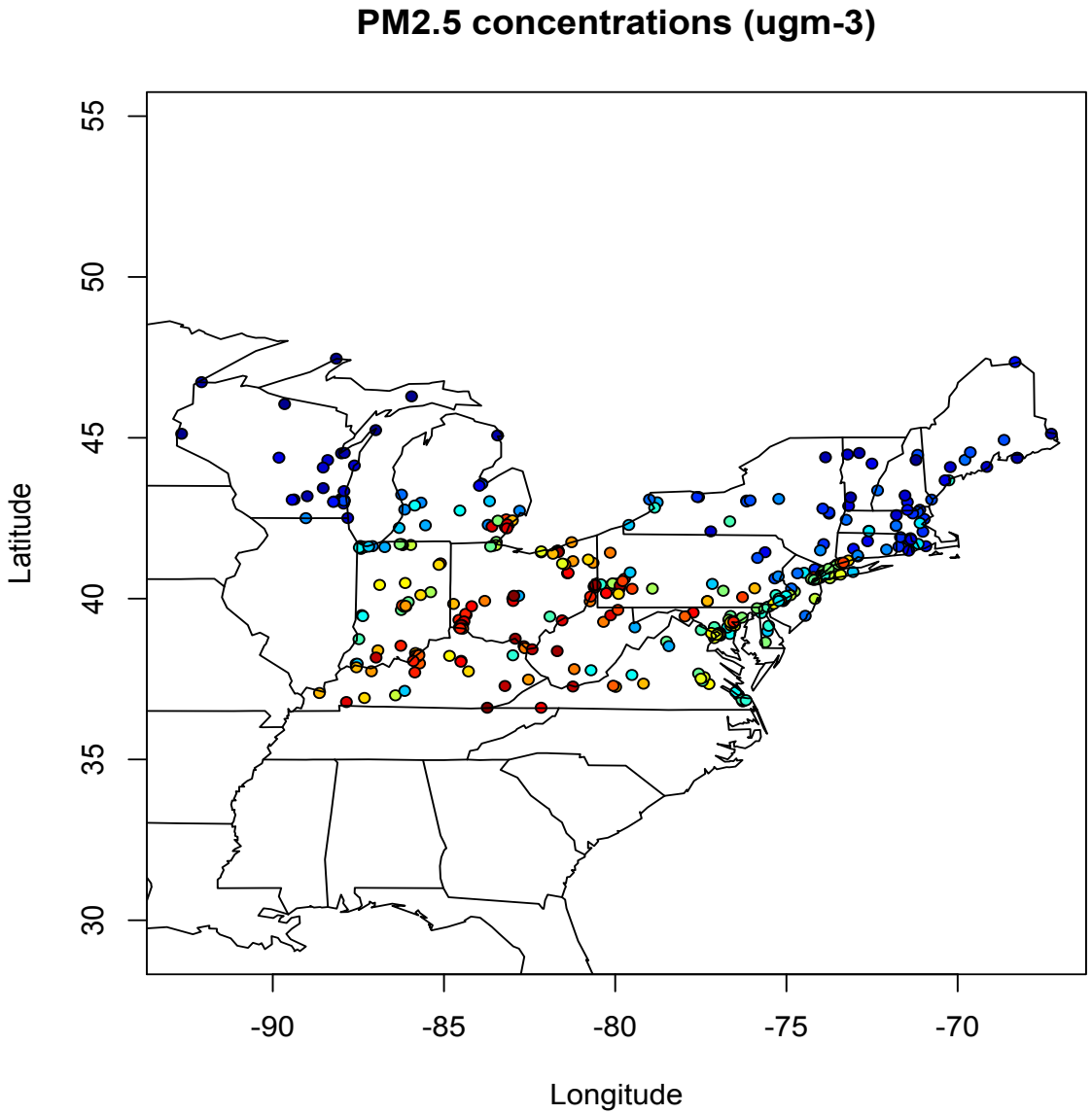
The common thread that links the data is a random process (also called stochastic process or random field)

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

where D is a domain in $\Re^d$ (d typically 2)

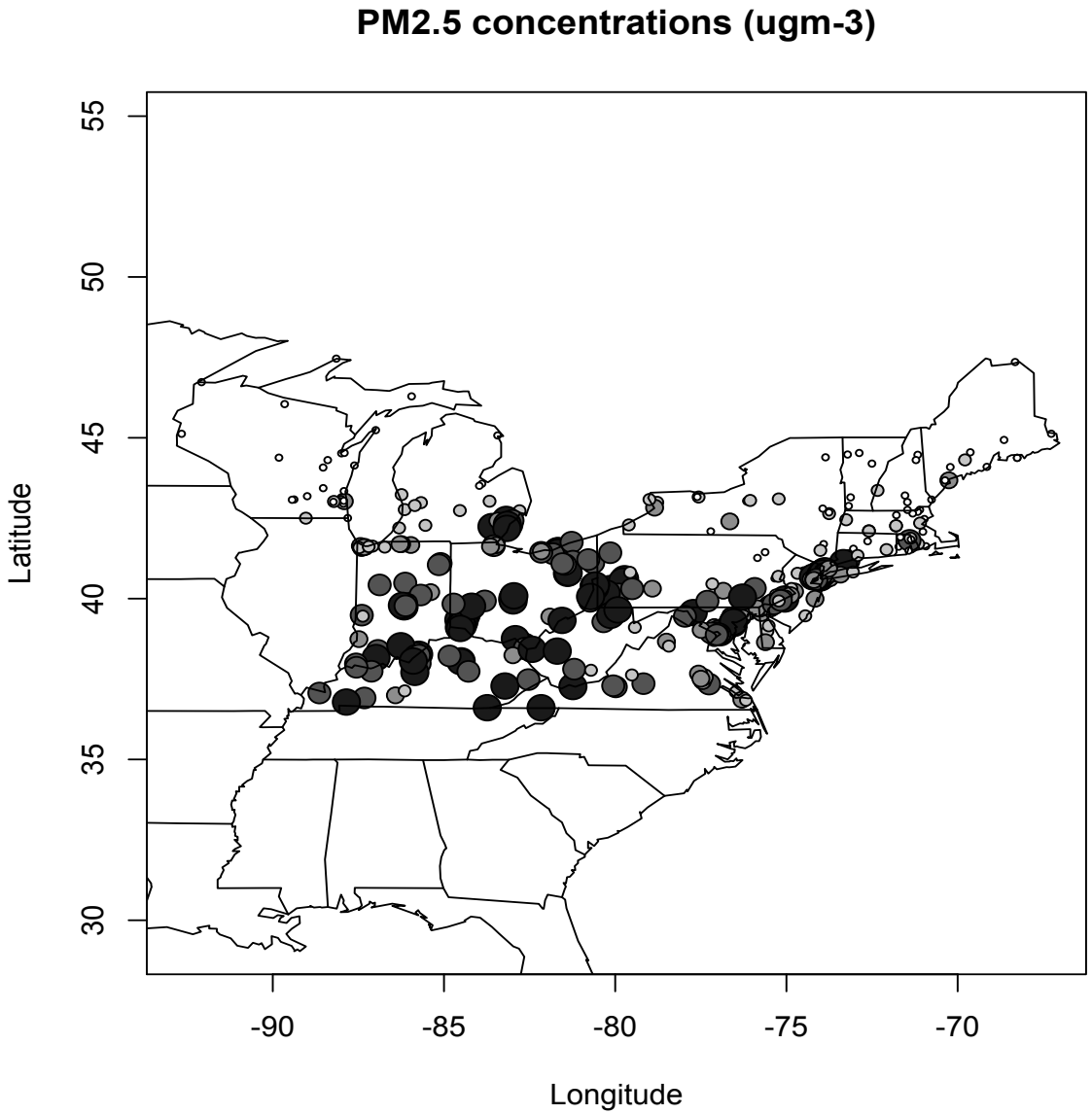
Geostatistical Data

Example



Geostatistical Data

Example



Goals of spatial statistics applied to point referenced data

- Determining if there is a spatial pattern in the observations. (Often called spatial "structure")
- Modeling the spatial correlation/covariance in the observations.
- Making predictions at unobserved locations: interpolation, smoothing.
- Accounting for spatial structure in regression models.
- Testing null hypothesis of no spatial structure.

Areal Data

Description and examples

- Analysis of data associated with an area
- Want to determine spatial patterns in data collected by zones or regions
- Areal units can be irregular (e.g. zip code, county) or 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)
- Areal data (lattices) use neighbour relationships

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.
- If there is a spatial pattern, *how strong is it?*

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

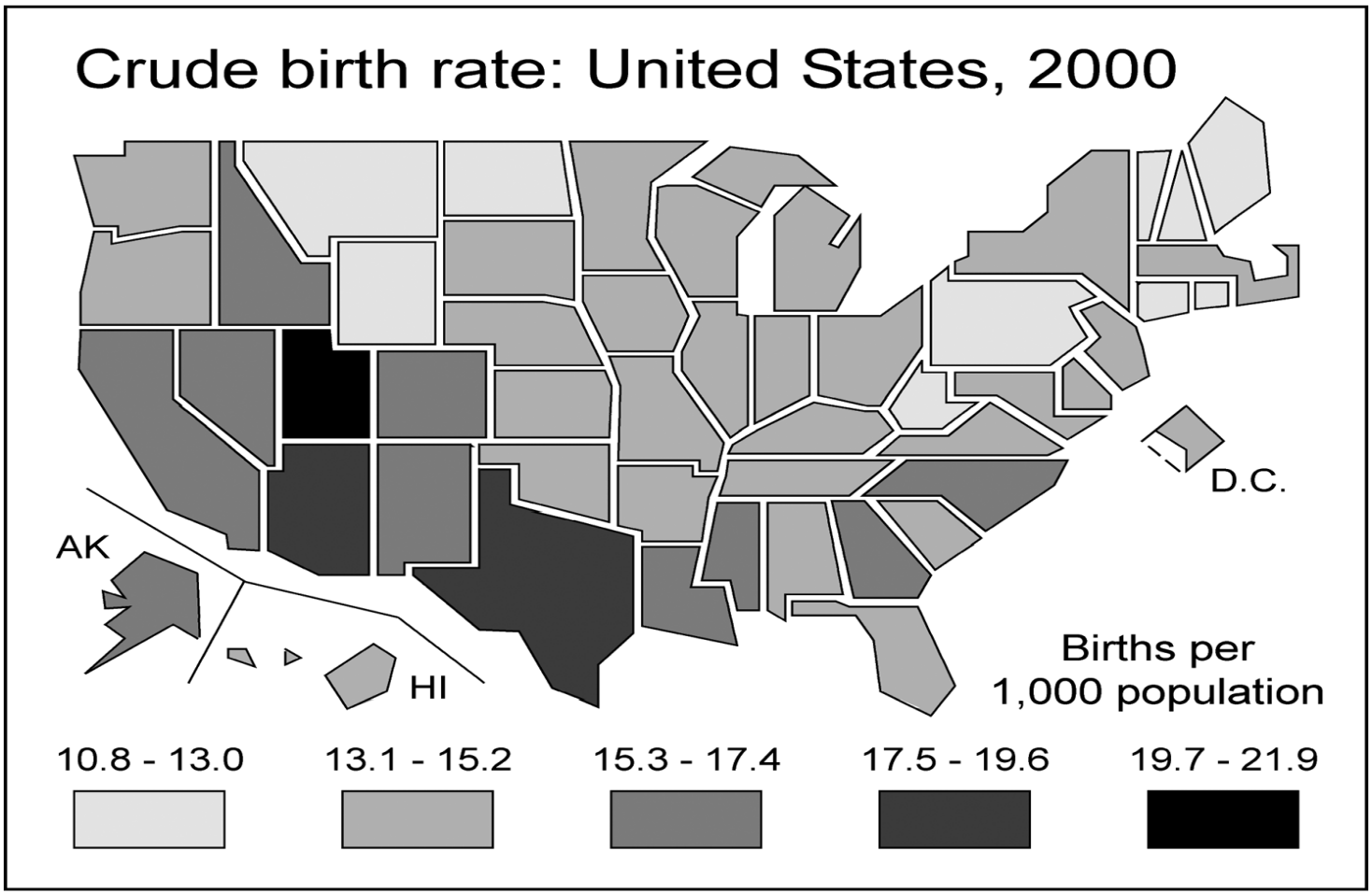


Figure: Monomier, N. Lying with Maps. Statistical Science 2005, 20(3) 215222.

Crude birth rates by state based on quantile cut points

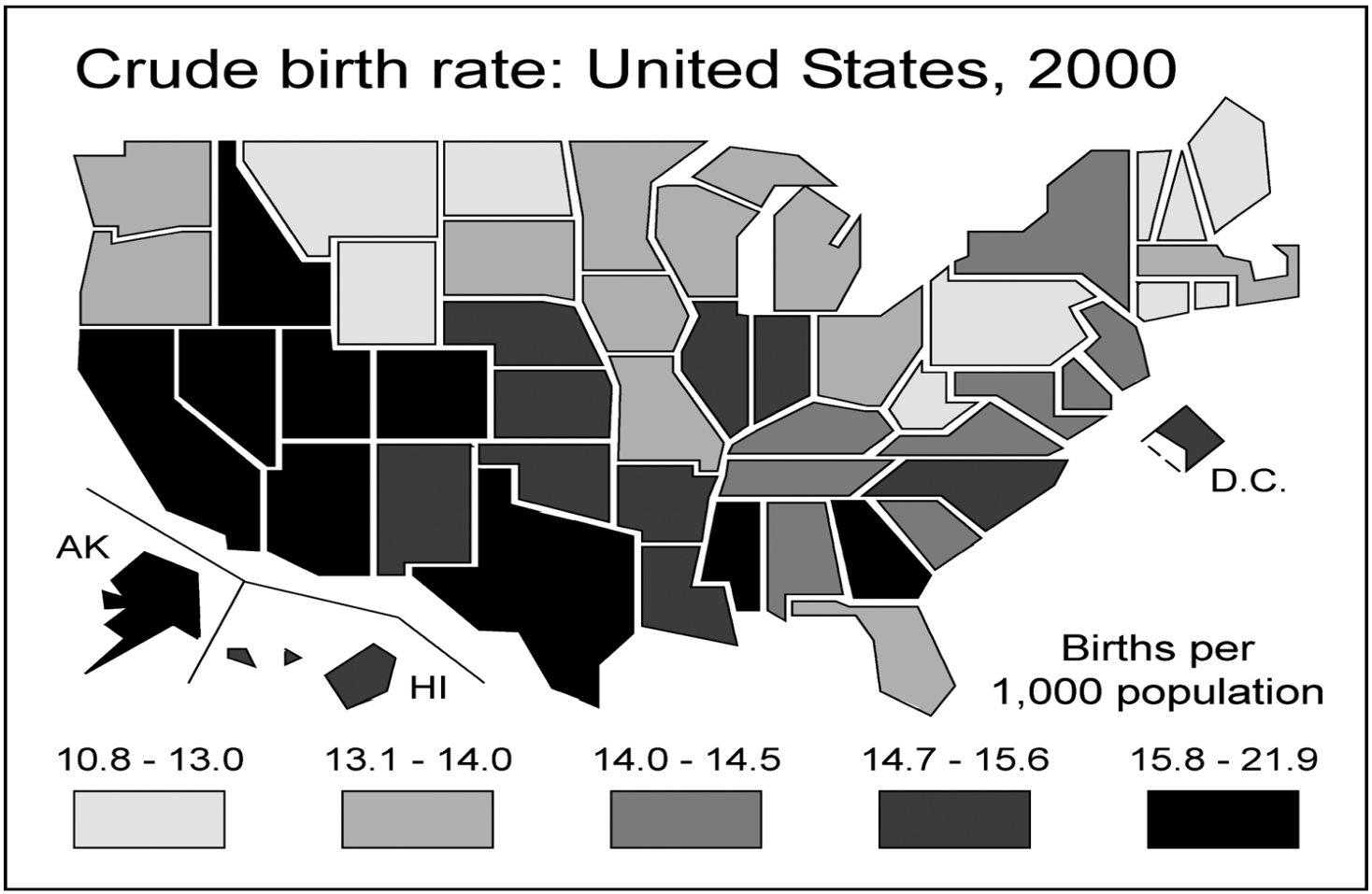


Figure: Monomier, N. Lying with Maps. Statistical Science 2005, 20(3) 215222.

Areal Data

Visualization

Figure: NASA MODIS satellite retrieval, August 2009

Areal Data

Visualization

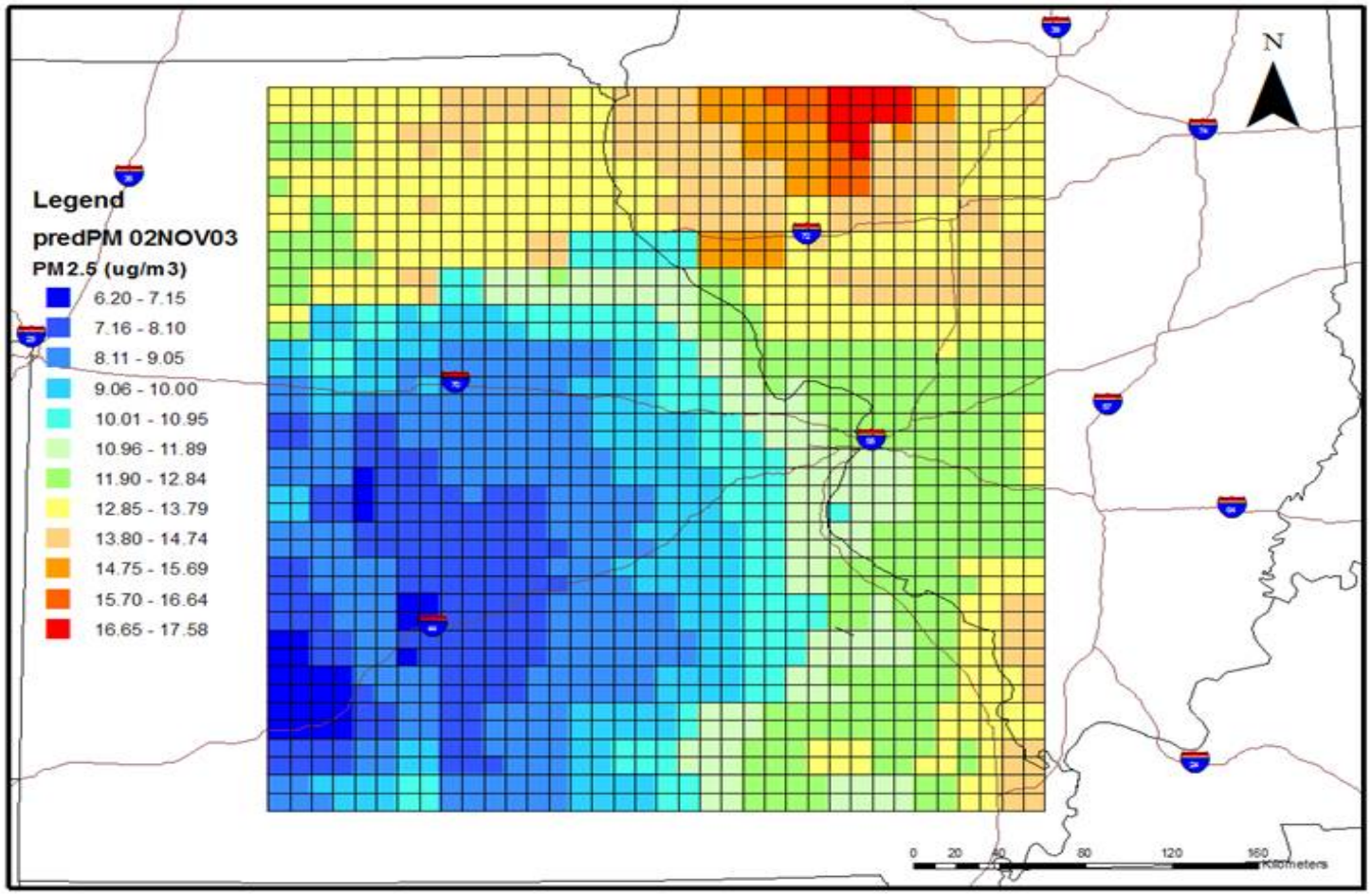


Figure: NASA MODIS AOT

Areal Data

Visualization

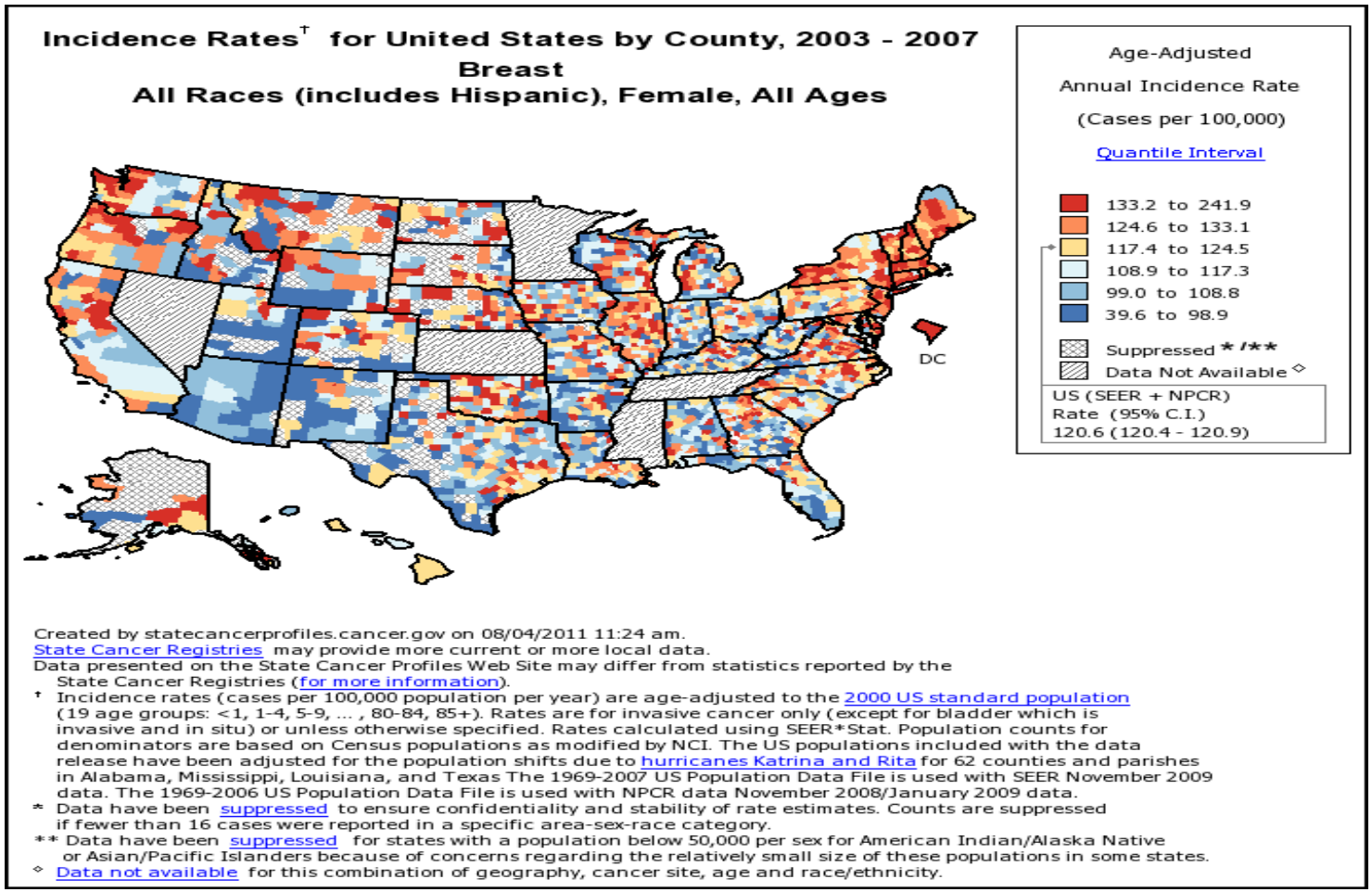
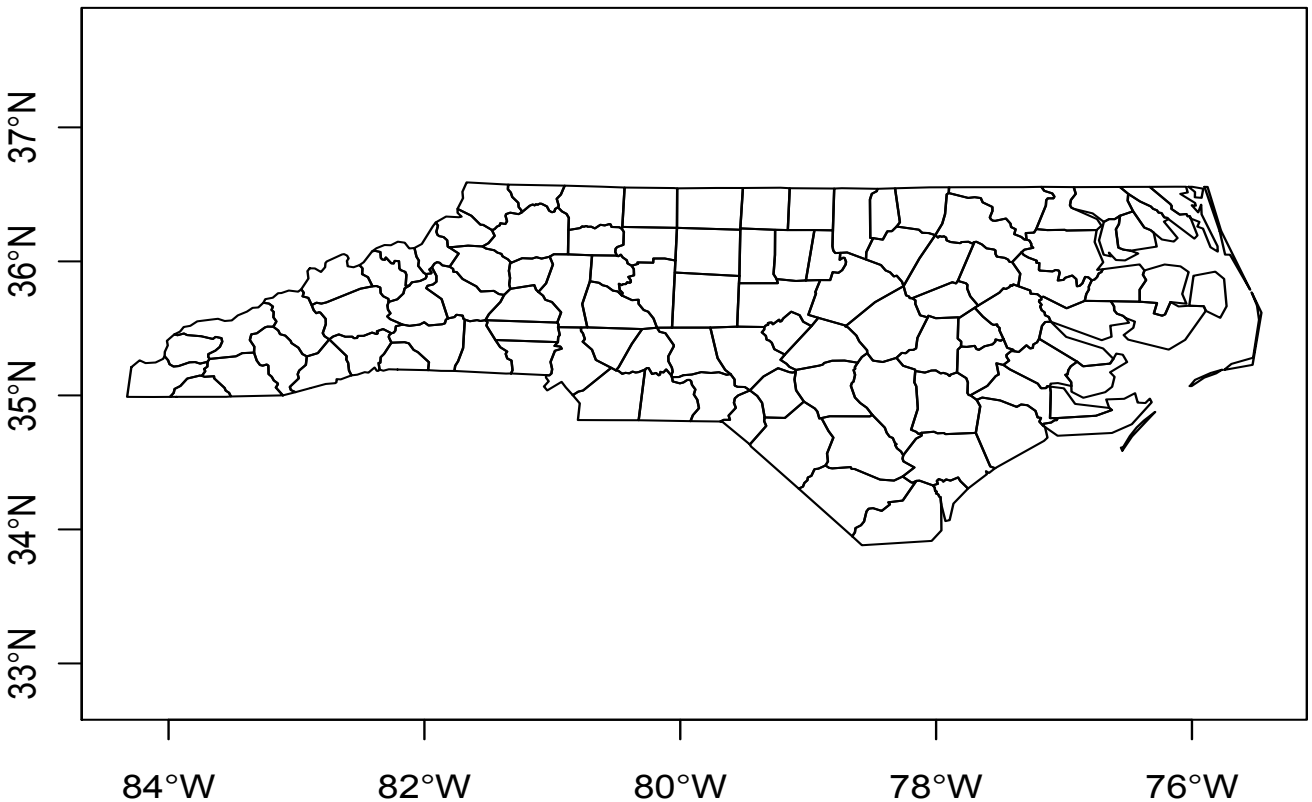
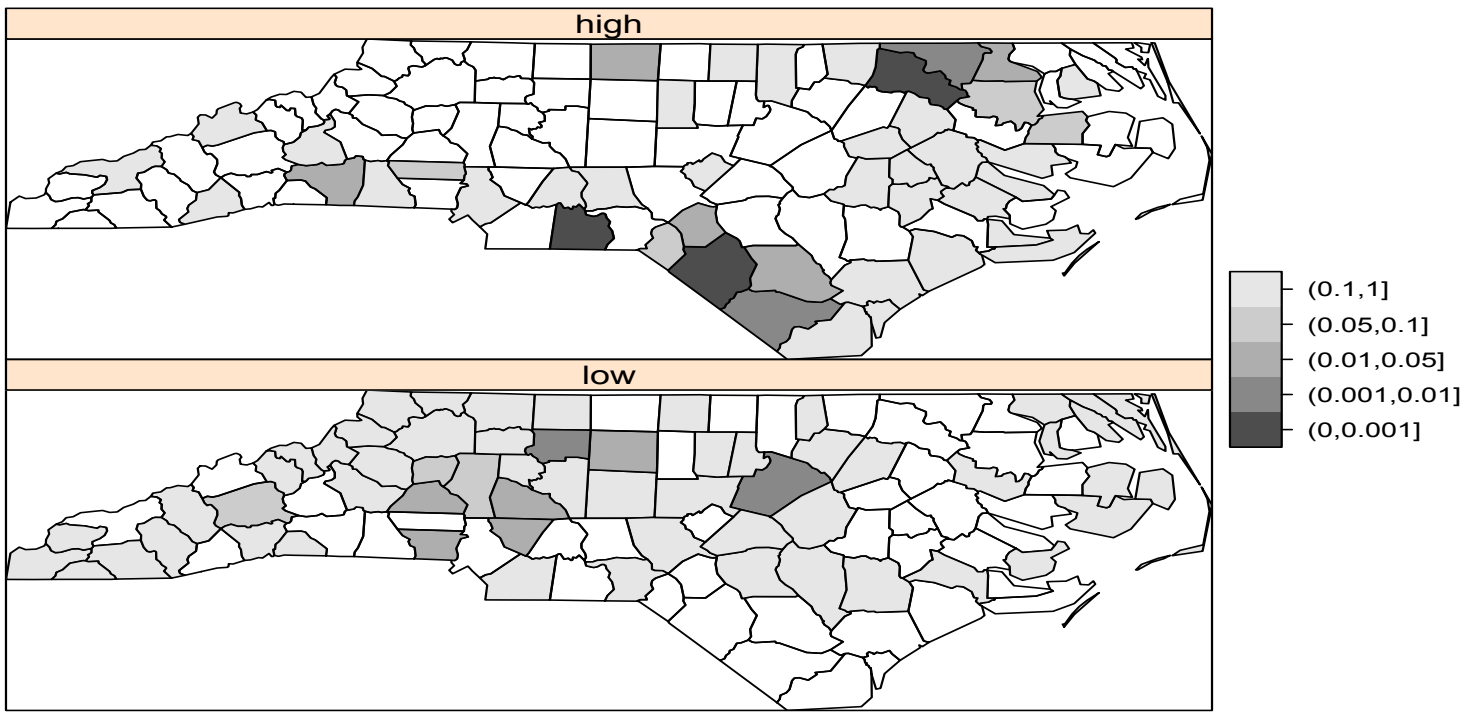


Figure: statecancerprofiles.cancer.gov

Sudden Infant Deaths in North Carolina



Sudden Infant Deaths in North Carolina



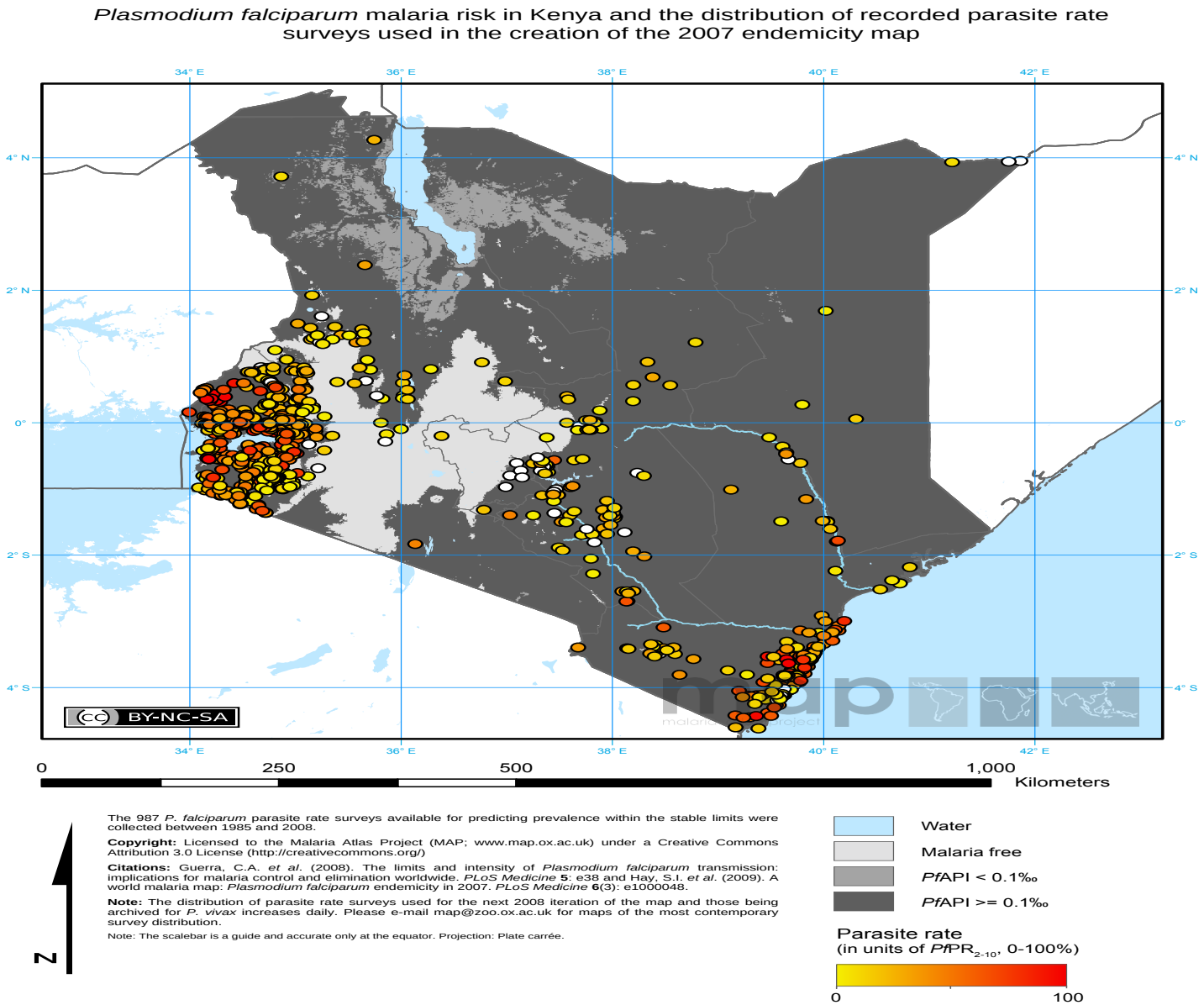
Point Pattern Data

Description and examples

- 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

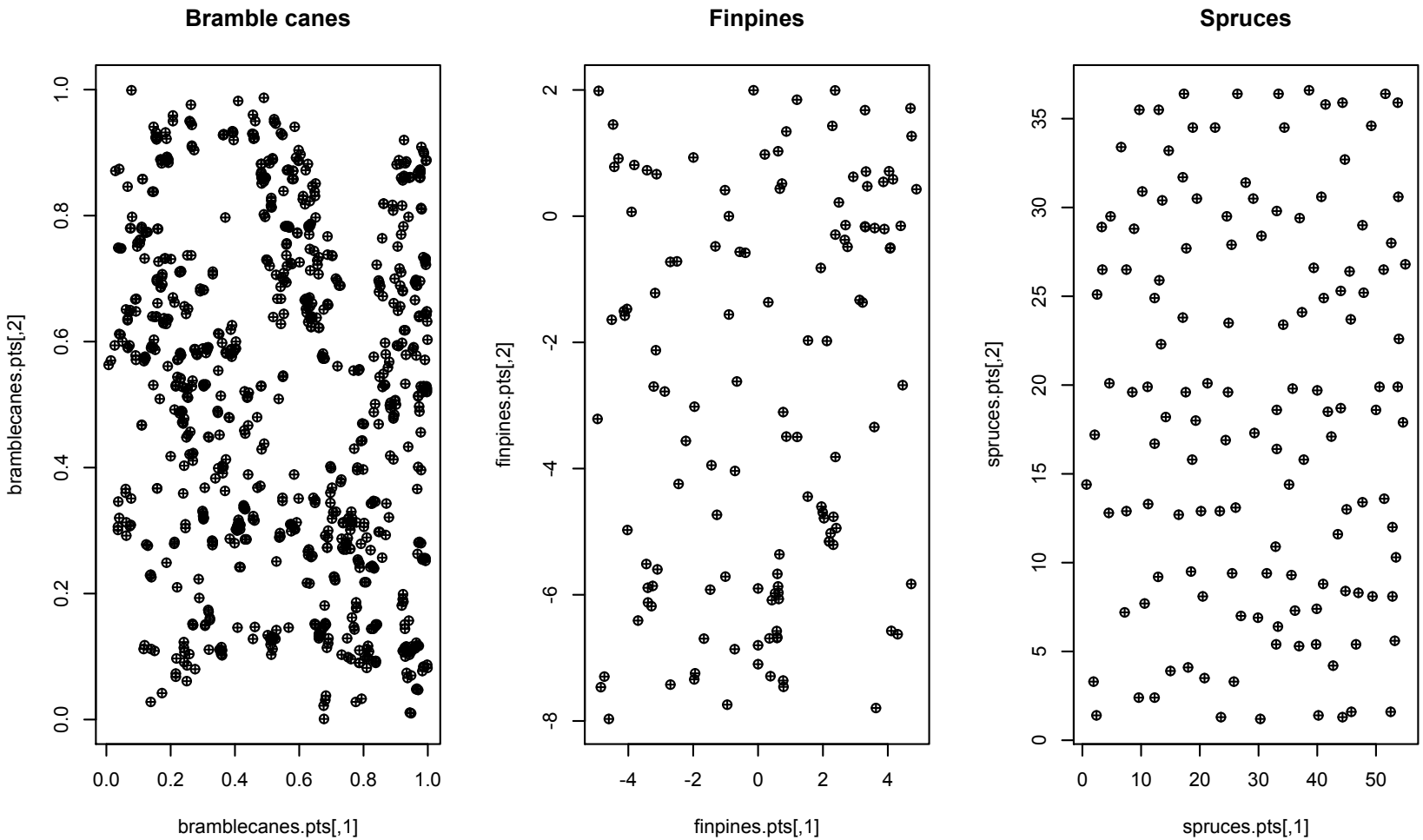
Point Pattern Data

Visualization



Point Pattern Data

Visualization



Point Pattern Data

Visualization

- Is there an underlying population distribution from which events arise in a region?
- Are events clustering in areas of high population?
- May need to account for features such as population when visualizing point patterns
- Often conclusions cannot be drawn from visual inspection alone

Point Pattern Data

Exploration

- 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?

- 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 (Ripleys K).
- Model some spatial pattern

Course Goals

- Three steps in analyzing spatial data
- Visualization
- Exploration
- Modeling
- Three types of spatial data
- Point process (Geostatistical)
- Areal
- Point pattern