COA 616 Geostatistics in Environmental Sciences

Lecture 1 Introduction

Wei Wu

August 30, 2016

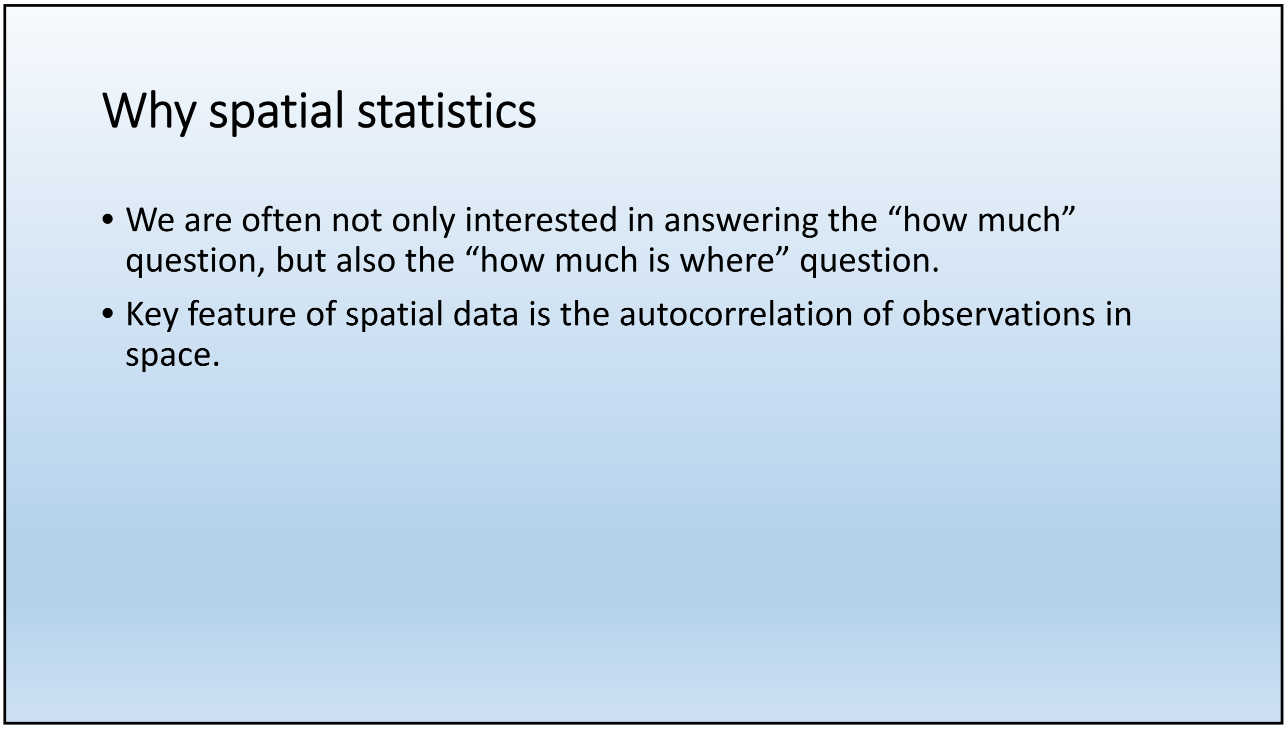
Spatial data analysis in R

https://cran.r-project.org/web/views/Spatial.html

- Your name, your advisor
- Research area
- Thesis or dissertation title if you know
- Your expectations of the course

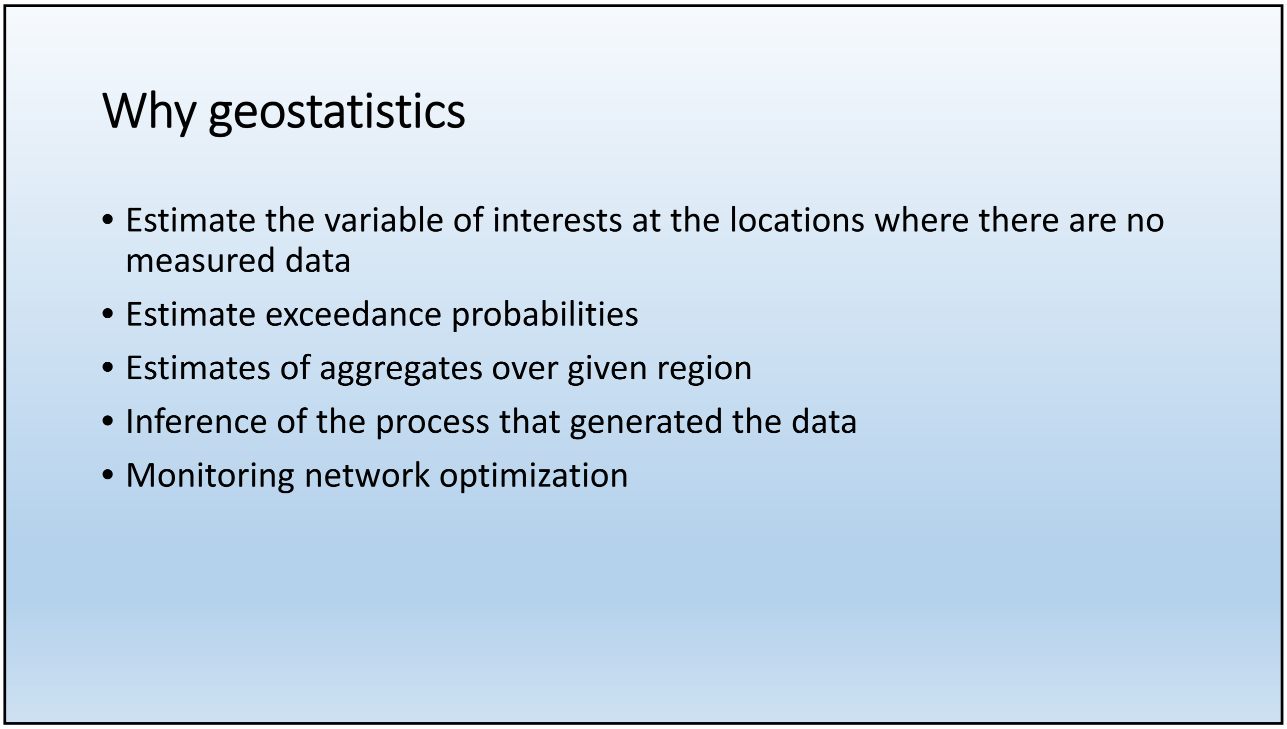
A little bit history on geostatistics

- Developed largely outside of statistical mainstream
- Originally, the term *geostatistics* was coined by Georges Matheron and colleagues at Fontainebleau, France, to describe their work addressing problems of spatial prediction arising in the mining industry.



Why spatial statistics

- We are often not only interested in answering the “how much” question, but also the “how much is where” question.
- Key feature of spatial data is the autocorrelation of observations in space.



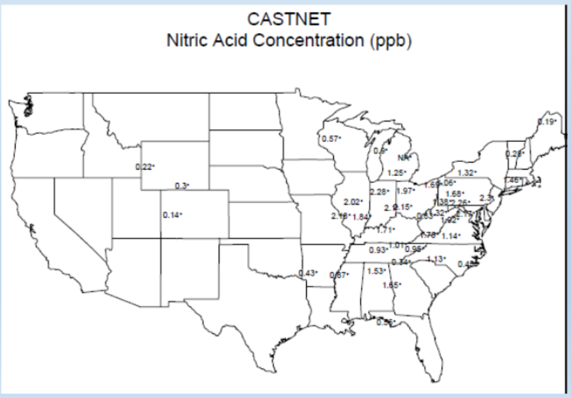
Why geostatistics

- Estimate the variable of interests at the locations where there are no measured data
- Estimate exceedance probabilities
- Estimates of aggregates over given region
- Inference of the process that generated the data
- Monitoring network optimization

Types of spatial data

We denote a spatial process in d dimensions as $\{Z(s): s \in D \subset R^d\}$

1) Geostatistical data: The domain D is continuous and fixed set.



Examples:

- Weekly concentrations of Nitric Acid in the US
- Weekly concentrations of ozone in the US
- Annual acidic deposition in the US

Types of spatial data cont.

2) Lattice data (regional data): The domain D is fixed and discrete. They are spatially aggregated over areal regions.

Examples:

Remote sensing data (pixel)

US Census bureau data (census tract)

Number of deaths, crimes reported for counties or zip codes

3) Point patterns: The domain D is random. Point patterns arise when the important variable to be analyzed is the location of events.

Examples:

Locations of rare species

Location at which weeds emerge in a garden

Locations of birds’ nests in a suitable habitat – evidence of territoriality?

Location of disease

Syllabus

Intro to statistics

- Framework of statistics
- Types of variables:

  nominal (categorical): No calculation is meaningful.

  ordinal: Order matters but not the difference between values.

  interval: Difference between two values is meaningful.

  ratio: Having the properties of an interval variable and also has an absolute 0.

Intro to statistics cont.

- Measurement of central location

Sample mean: $$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Population mean: $$\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \ldots + x_N}{N}$$

Weighted mean: $$\overline{x_w} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} = \frac{w_1 x_1 + w_2 x_2 + \ldots + w_n x_n}{w_1 + w_2 + \ldots + w_n}$$

Median: Middle value when a set of n measurements is arranged in increasing or decreasing order of magnitude.

Mode: Most frequently occurring value

Intro to statistics cont.

- Measurement of variation (spread)

Range: R = max(x_i) – min(x_i) - influenced by outliers

Mean deviation: $$MD = \frac{\sum_{i=1}^{n} | x_i - \bar{x} |}{n}$$

Sum of squares of deviation from the mean: $$SS = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Corrected sum of squares of deviation from the mean: $$SS_p = \sum_{i=1}^{N} (x_i - \mu)^2$$

Sample variance:

$$s^2 = \frac{SS}{n-1} = \frac{\sum_{i=1}^{n} x_i^{\,2} - \frac{(\sum_{i=1}^{n} x_i)^2}{n}}{n-1}$$

Population variance:

$$\sigma^2 = \frac{SS_p}{N} = \frac{\sum_{i=1}^{N} x_i^{\,2} - \frac{(\sum_{i=1}^{N} x_i)^2}{N}}{N}$$

Standard deviation: $s = \sqrt{s^2}$

Coefficient of variance: $CV = 100\frac{s}{x}$

Intro to statistics cont.

- Measure of symmetry (third moment about the mean): Skewness

$$g_1 = \frac{m_3 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^3}{s^3}$$

g_1 = 0 – symmetric,
g_1 <0 – negative skew,
g_1 > 0 – positive skew

- Measure of peakness: Kurtosis

$$g_2 = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x_i})^4}{s^4} - 3$$

g2 = 0 – Normal distribution
g2 >0 – More peaked than normal distribution
g2 <0 – Less peaked than normal distribution

- Measure of position

Standard scores: indicating how many standard deviations an observation is above or below the mean value.

$$z_i = \frac{x_i - \bar{x}}{s}$$

Percentile: Pth percentile is the value which p percent of the data fall at or below.

Deciles: 10th, 20th, …, 90th percentiles; Quartiles: 25th, 50th, 75th percentiles

Median: 50th percentiles

Intro to statistics cont.

RV: A well defined numerical description of the outcomes in the sample space of a random experiment.

Discrete RV: mean

$$\mu = E(X) = \sum x_i p(x_i)$$

variance

$$\sigma^2 = E[(X-\mu)^2] = \sum (x_i - \mu)^2 p(x_i)$$

covariance

$$\sigma^2{}_{XY} = \mathrm{cov}(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] = \sum\sum (x_i - \mu_X)(y_j - \mu_Y) p(x_i, y_j)$$

Continuous RV:

mean

$$\mu = \int x_i f(x_i) dx$$

variance

$$\sigma^2 = \int (x_i - \mu)^2 f(x_i) dx$$

covariance

$$\sigma^2{}_{XY} = \int\int (x_i - \mu_X)(y_j - \mu_Y) f(x_i, y_j) dydx$$

$$\rho_{XY} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

Intro to statistics cont.

$$E(aX \pm b) = aE(X) \pm b$$

$$E(X \pm Y) = E(X) \pm E(Y) = \mu_X \pm \mu_Y$$

$$E(XY) = E(X)E(Y) = \mu_X \mu_Y$$ If X, Y are independent

$$\sigma^2{}_{X \pm b} = \sigma^2{}_X$$

$$\sigma^2{}_{aX} = a^2 \sigma^2{}_X$$

$$\sigma^2{}_{aX \pm b} = a^2 \sigma^2{}_X$$

$$\sigma^2{}_{aX \pm bY} = a^2 \sigma^2{}_X + b^2 \sigma^2{}_Y \pm 2ab\sigma_{XY}$$

Normal distribution

$$f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \; where -\infty < x < \infty$$

Normality test

H_0: Sampled data come from normal distribution
H_a: Sampled data do not come from normal distribution

Shapiro-Wilk test is one of the most powerful normality test, especially for small samples.
Box-Whisker plot
QQ plot

Transformation

Central Limit Theorem

1. The sample mean $\overline{x}$ is a RV since its value changes from sample to sample.
2. The mean of all possible sample means = population mean: $E(\overline{X}) = \mu$
3. The spread of all possible sample means is the standard error of the mean.

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} or$$

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

4. 1) The probability distribution of all possible sample means is normal when the parent population is normally distributed or 2) is approximately normal when the sample size (n) is sufficiently large (n>30). Regardless of the distribution of the parent population, as n approaches infinity, the probability distribution of all possible sample means becomes normal.

Any problems with average and sum of iid RV suggest use of CLT.

Central Limit Theorem

Assume length of a worm (in mm) is genetically determined by 81 loci on the chromosomes. Each locus independently contributes

2 (with prob of 0.3)

0 (with prob of 0.7)

The length of the worm (mm) is the sum of these 81 contributions.

Max = 162 mm; min = 0 mm

a) What is expected length?

b) What is the probability of length between 40 mm and 50 mm?

$$S = X_1 + X_2 + \ldots + X_n$$

$$E(S) = E(X_1) + E(X_2) + \ldots + E(X_n) = n\mu$$

$$\operatorname{var}(S) = \operatorname{var}(X_1) + \operatorname{var}(X_2) + \ldots + \operatorname{var}(X_n) = n\sigma^2$$

$$sd(S) = \sqrt{n}\sigma$$

Regression

Assumptions

1) The x_i values are fixed and not random variables.
2) The y_i values for any given x_i are values of a RV (The samples of y_i for any given x_i value are randomly selected from the population of RV of Y).
3) The y_i values for any given x_i have a uniform variance (homoscedasticity), more precisely, expect constant variance in the residuals (vs. predictions or covariates).
4) The y_i values for any given x_i is independent from other y_i values in the sample, more precisely, residuals are not correlated.
5) The residuals are normally distributed. Sometimes you see "y is normally distributed".
6) Linearity between X and Y (model specification)

Examples