Spatial and Environmental Statistics

Dale Zimmerman

Department of Statistics and Actuarial Science
University of Iowa

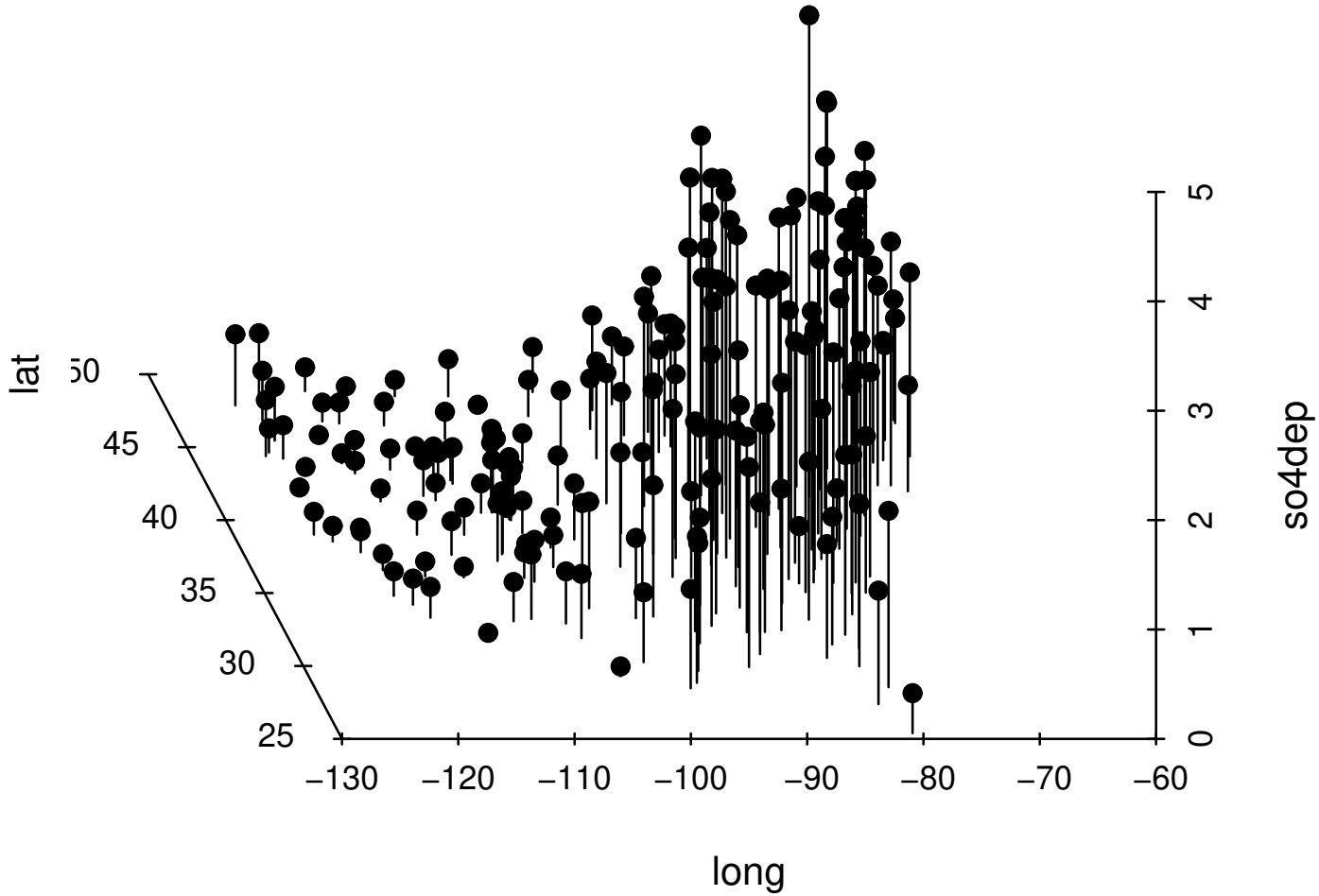
August 22, 2020

Overview

1. Spatial Datasets
2. What is Spatial (and Environmental) Statistics?
3. Three Important Types of Spatial Data

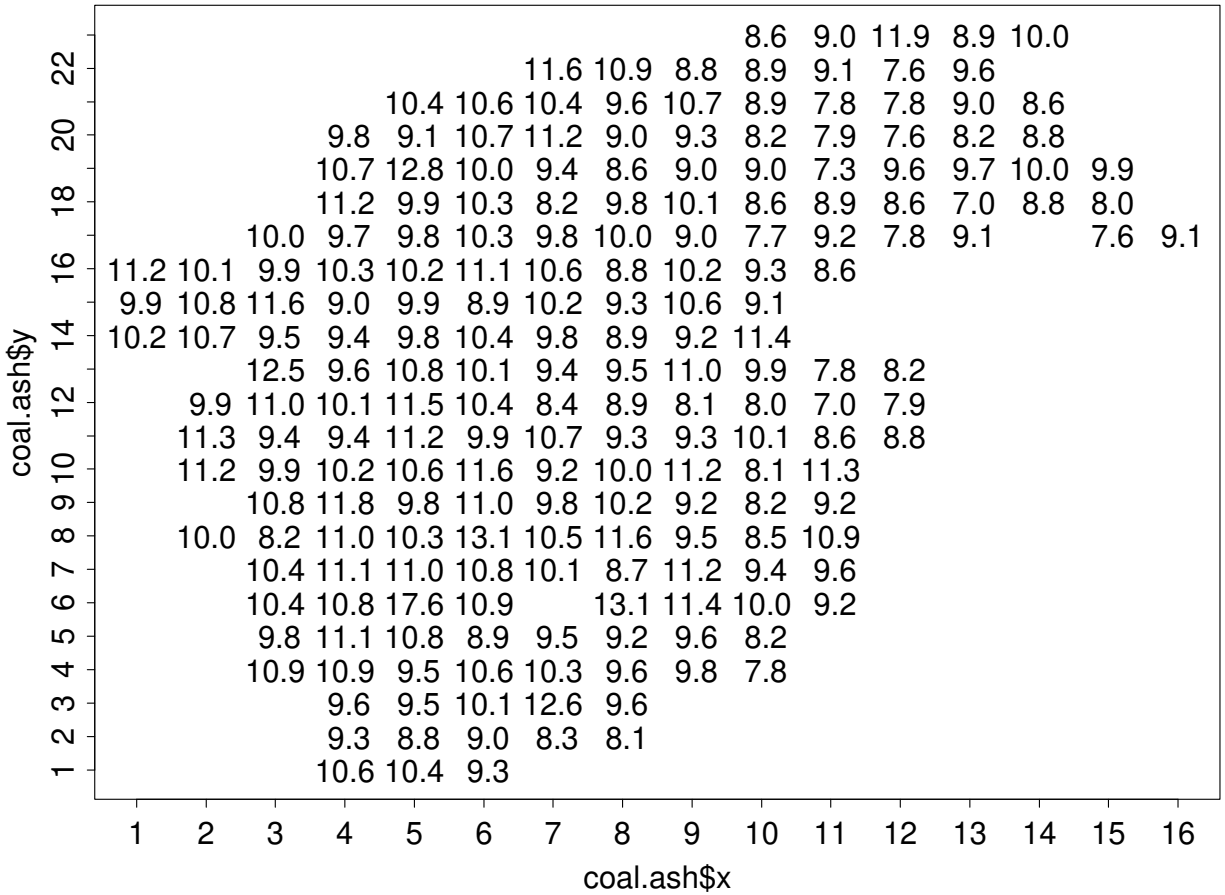
Spatial Datasets

Wet deposition of SO_4 (g/m^2) in 1987 at National Acid Deposition Program sites.



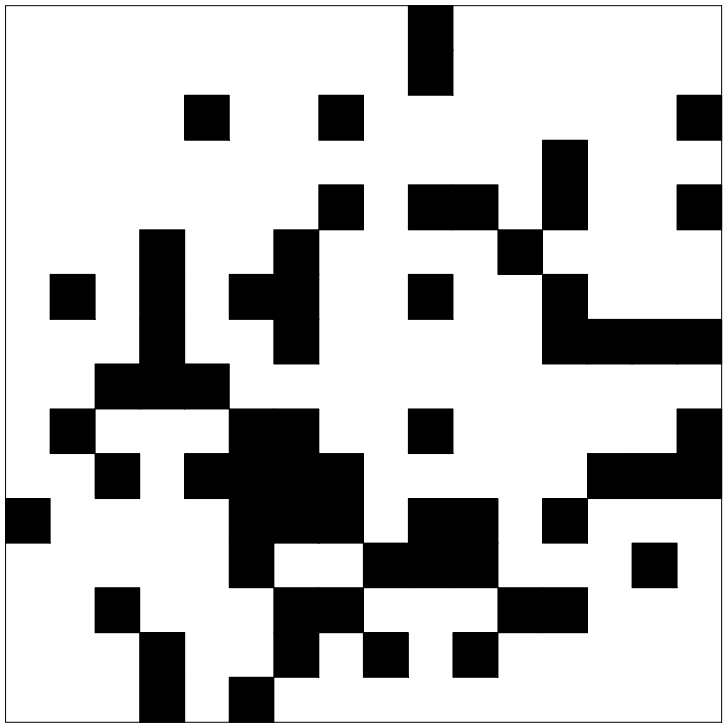
Spatial Datasets

Coal ash samples from a mine in Pennsylvania.



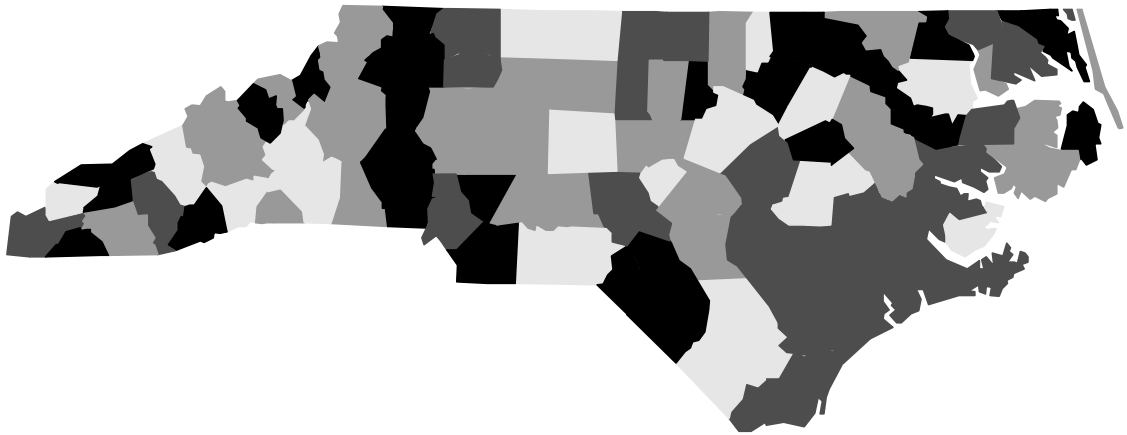
Spatial Datasets

Presence (black) or absence (white) of *Atriplex hymenelytra* on a grid of quadrats in Death Valley, CA.



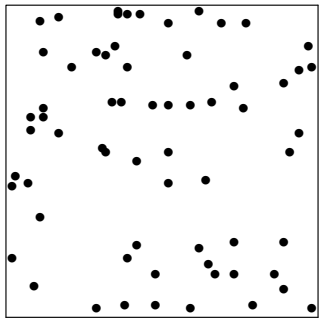
Spatial Datasets

Population-adjusted mortality rates due to SIDS in counties of North Carolina, 1974-1978.

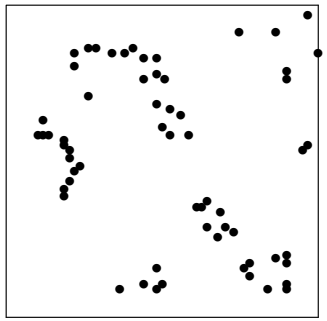


Spatial Datasets

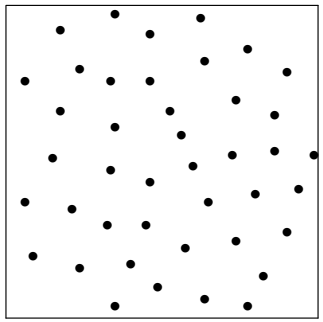
Locations of Japanese pines, redwood saplings, biological cells, and scouring rushes in various study areas.



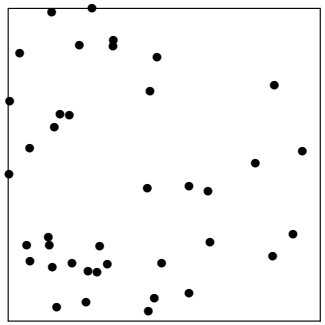
Pines



Redwoods



Cells



Rushes

What is Spatial Statistics?

- Basic ingredients:
  - Observations on one or more "response" variables are taken at multiple, identifiable sites in some spatial domain.
  - Locations of these sites are observed and are attached, as labels, to the observations.
  - An analysis of the observations is performed, in which the spatial locations of sites are taken into account.
  - Either the observations or the spatial locations (or both) are modelled as random variables, and inferences are made about these models and/or about additional unobserved variables.

- Thus, spatial statistics would include any investigation in which the data's spatial locations play a role in a probabilistic or statistical analysis (we will emphasize the statistical).

What is Spatial Statistics?

- Spatial statistics is a vast subject, in large part because spatial data are of so many different types. The response variable may be:
  - univariate or multivariate
  - categorical or continuous
  - real-valued (numerical) or not real-valued (e.g. set-valued)
  - observational or experimental

- The data locations may:
  - be points, regions, or something else
  - be regularly or irregularly spaced
  - be regularly or irregularly shaped
  - belong to a Euclidean or non-Euclidean space (e.g., river network)

What is Spatial Statistics?

- The mechanism that generates the data locations may be:
  - known or unknown
  - random or non-random
  - related or unrelated to the processes that govern the responses
- Related subjects:
  - Time series analysis
  - Reliability/survival analysis
  - Longitudinal data analysis

Three Important Types of Spatial Data

1. Geostatistical data
   - The response variable is observed over very small, non-overlapping subregions. The subregions are so small relative to the spacing between them that nothing of consequence is lost by idealizing the subregions as points.
   - Examples:
     (a) Annual acid rain deposition in U.S.
     (b) Richness of iron ore within an ore body
     (c) Maximum ozone level over a year's time at each of several monitoring sites in a large city

Three Important Types of Spatial Data

2. Areal (sometimes called lattice) data
   - The response variable exists and is observed only on a finite set of non-overlapping (usually contiguous) subregions within the study region.
   - Examples:
     (a) Presence or absence of a plant species in square quadrats over a study area
     (b) Numbers of deaths due to SIDS in the counties of North Carolina
     (c) Pixel values from remote sensing (satellites)

Three Important Types of Spatial Data

3. Spatial point patterns
   - Data are the spatial locations of point "events" within the study region. No response variable is observed at the locations.
   - Examples:
     (a) Locations of *Equisetum arvense* plants at a marsh edge — evidence of environmental gradient?
     (b) Location of lunar craters — meteor impacts or volcanism?
     (c) Locations of residences of individuals with lung cancer within 50 miles of a large incinerator — does disease risk increase with proximity to the incinerator?
   - A more general kind of spatial point pattern is a *marked* spatial point pattern, in which a nontrivial response variable (called the mark) is observed at each point. If the mark is discrete, we have a multivariate spatial point pattern.

Three Important Types of Spatial Data

The distinctions between these three types are not always clear-cut. In particular, areal data and geostatistical data have many similarities. In a sense, areal data are not as refined as geostatistical data or spatial point patterns since you can obtain areal data by various reductions (integration or counting) of the other two.

In addition to indicating some prototypes of spatial data, the examples listed above indicate the breadth of scientific disciplines which are concerned with spatial data.

Spatial Statistics — Why Bother?

There are two main reasons why we bother with spatial statistics for spatial data (instead of just using classical statistics):

1. Characterizing the spatial structure of the data may be of direct interest.

2. The spatial structure may not be of direct interest, but modeling or otherwise accounting for it may improve other inferences.

Spatial Statistics — Why Bother?

More details on the first reason:

The observations are suspected of having a coherent spatial structure, the characterization of which may be important. The kinds of spatial structure that may occur vary across types, but there are some commonalities. It has been observed over and over again in practice that observations taken at sites close together tend to be more alike than observations taken at sites far apart. In the spatial context, this is sometimes called the "First Law of Spatial Statistics."

This law can manifest through either the "large-scale" (global) structure or the "small-scale" (local) structure, or both.

Spatial Statistics — Why Bother?

- Large-scale structure
  - Mean function of geostatistical process
  - Mean vector of areal process
  - Intensity of spatial point process

- Small-scale structure
  - Variogram, covariance function of geostatistical process
  - Neighbor weights for areal process
  - Ripley's K-function, second-order intensity, nearest-neighbor functions for spatial point process

In due time we will precisely define and study all of the above.

Spatial Statistics — Why Bother?

Two important types of spatial structure are stationarity and isotropy. Formal definitions of these will be given later. For now, the following descriptions will suffice.

(a) Stationarity — the property whereby the behavior of the process is similar across all of the spatial domain under study. This implies:

- constant (no trend) large-scale structure
- small-scale structure that depends on the spatial locations only through their relative positions (displacement)

(b) Isotropy — the property whereby the process is stationary, plus the small-scale structure depends on the spatial locations only through the Euclidean distance between them.

Spatial Statistics — Why Bother?

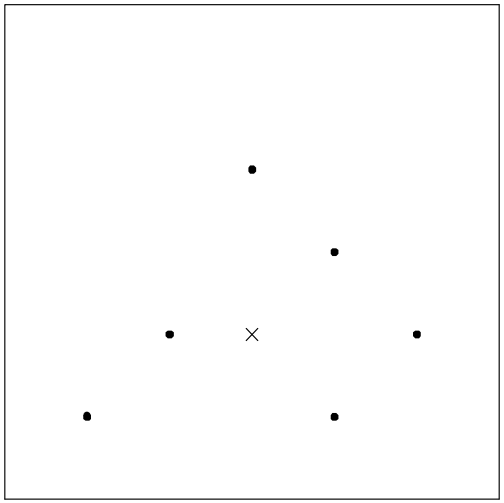
Characterization of the spatial structure is usually achieved by one or more of the following types of statistical inference:

- Testing for the existence of spatial structure
- Estimating spatial structural parameters
- Choosing between alternative structural models
- Prediction of unobserved variables using estimated structure (almost exclusively geostatistical, where it is known as kriging)

Now we elaborate on the second reason why we bother with spatial statistics, i.e., taking account of the spatial structure to improve non-spatial inferences.

Spatial Statistics — Why Bother?

Example 1 (from geostatistics). Prediction of an unobserved response at the $\times$ from 6 nearby observations.

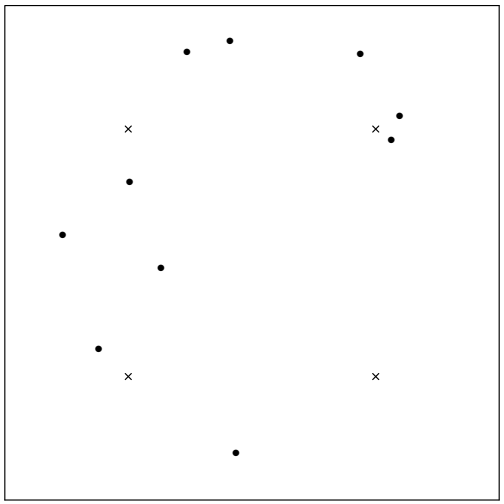


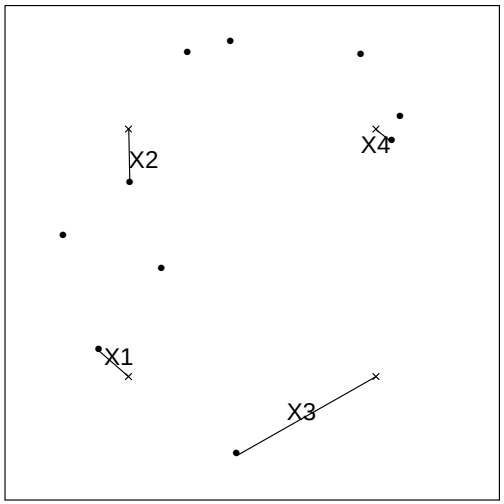
If all of the observed responses are uncorrelated with each other and with the predictand at $\times$, then the average of the observed responses is the best linear unbiased predictor. If, however, the responses are spatially correlated, then the simple average is inefficient.

Spatial Statistics — Why Bother?

Example 2 (from spatial point pattern analysis). Estimation of the number, N, of trees in a forest of area $|A|$.

One method for estimating N is based on measuring the distance, X_i, to the nearest tree from each of m fixed points.





Spatial Statistics — Why Bother?

If tree locations are *completely spatially random* (i.e., a random sample from the uniform distribution on A), then the MLE of N is

$$\hat{N} = \frac{m|A|}{\sum_{i=1}^{m} \pi X_i^2}.$$

If not, then $\hat{N}$ can be badly biased.

Spatial Statistics — Why Bother?

Example 3 (from areal data analysis). Variance of the sample mean.

Consider 16 observations taken over square subregions in a 4×4 grid, indexed by rows and columns as $Z(i, j)$:

Z(4,1)			Z(4,4)
Z(1,1)			Z(1,4)

Spatial Statistics — Why Bother?

Suppose that the observations have common mean μ and common variance 1, and

$$\text{corr}[Z(i,j), Z(k,l)] = 0.5^{|i-k|+|j-l|}.$$

Suppose we wish to estimate μ by the sample mean, $\bar{Z}$. It's tedious but mathematically easy to show that $\text{var}(\bar{Z}) \doteq 0.266$.

If there were no spatial correlation, then $\text{var}(\bar{Z}) = 1/16 = 0.0625$. Thus if we obtain a 95% (say) confidence interval for μ by acting as though there is no correlation, our interval will actually be much narrower than it should be.

Spatial Statistics — Why Bother?

Example 4 (from areal data analysis). Spatial experimental design.

Consider a field-plot experiment with 50 units, laid out in 10 linear blocks of 5 plots each. Suppose there are 5 treatments and each is to occur once in each block. Consider two designs:

- Randomized block design (RBD)
- First-order nearest-neighbor balanced design (first-order NNBD)

5	4	1	3	2
2	5	4	1	3
3	2	5	4	1
1	3	2	5	4
4	1	3	2	5
5	1	2	4	3
3	5	1	2	4
4	3	5	1	2
2	4	3	5	1
1	2	4	3	5

Spatial Statistics — Why Bother?

It turns out that if treatment-adjusted responses are independent across blocks but positively spatially correlated within blocks, then a first-order NNBD is optimal in the sense of minimizing the average variance of treatment contrasts. It can be considerably superior to all RBD's if the correlation is strong.