Spatial and Environmental Statistics

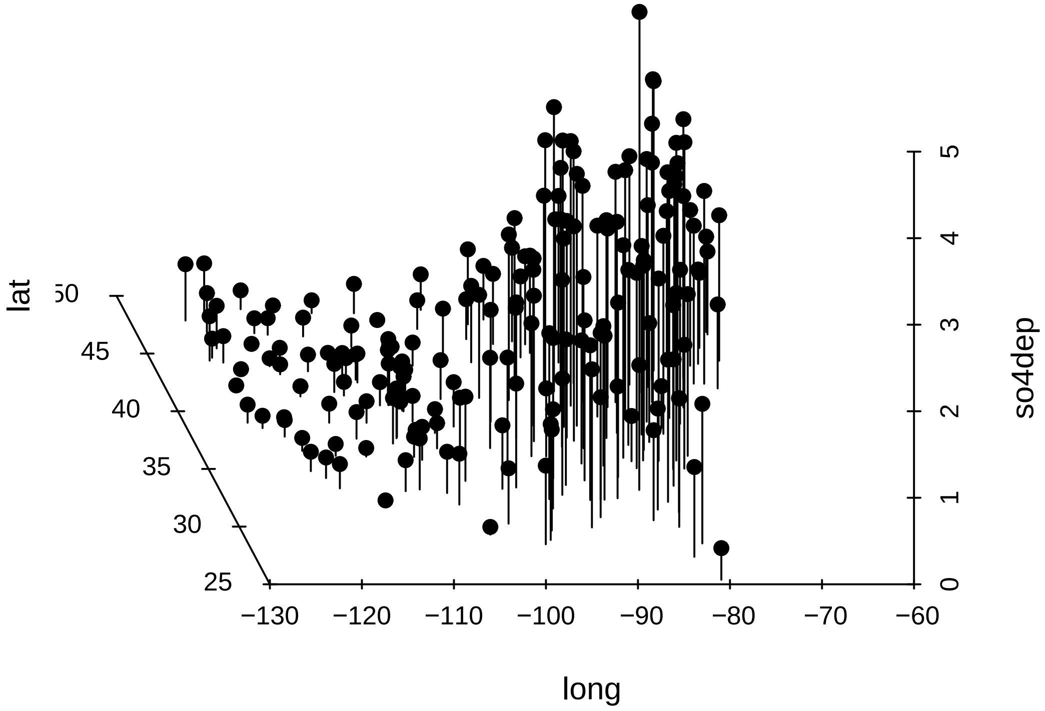
by

Dale L. Zimmerman

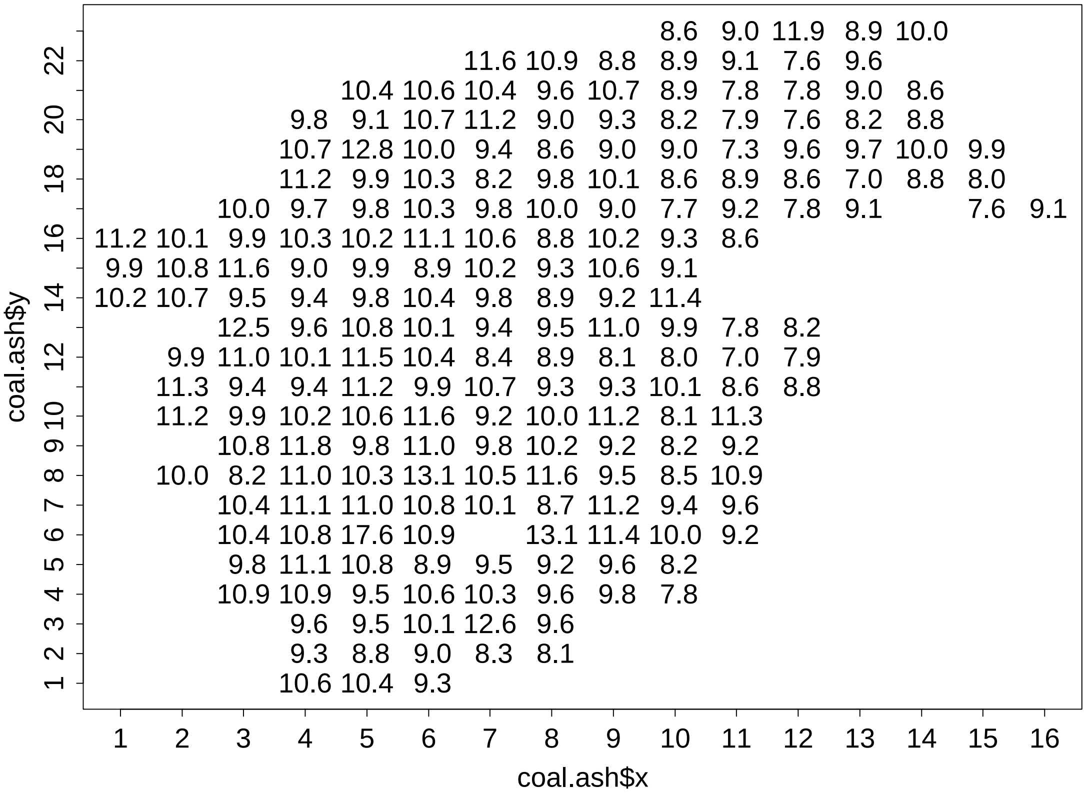
I. INTRODUCTORY CONCEPTS

A. Spatial Datasets

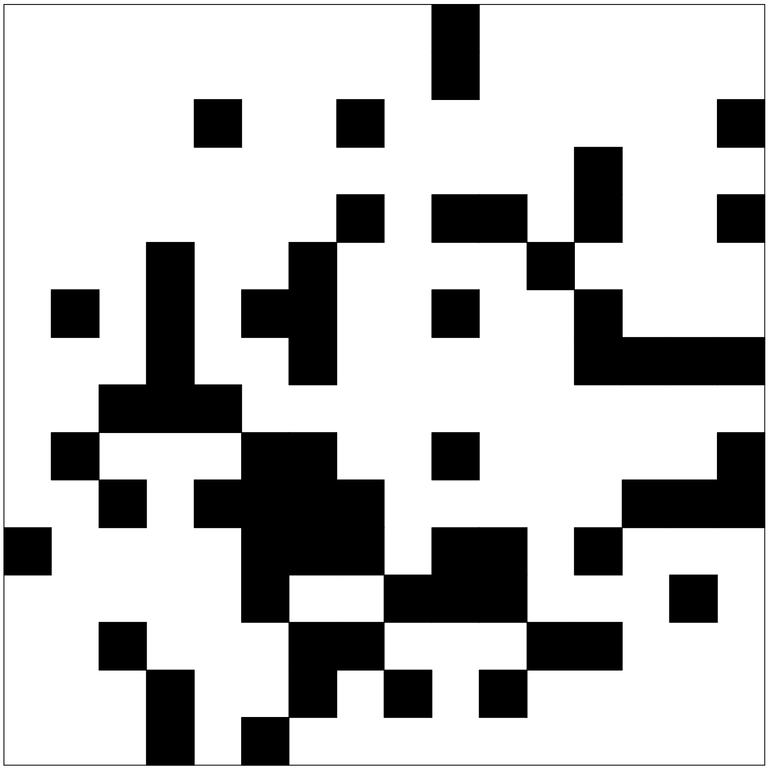
(a) Wet deposition of SO_4 (g/m^2) in 1987 at National Acid Deposition Program sites.



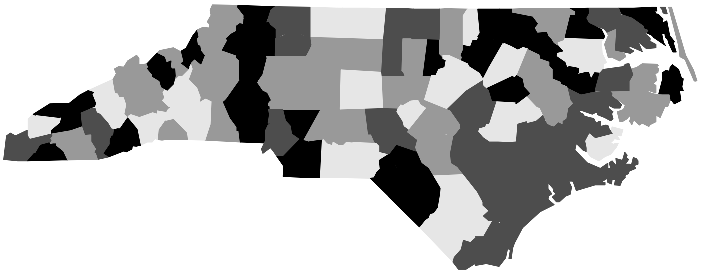
(b) Coal ash samples from a mine in Pennsylvania.



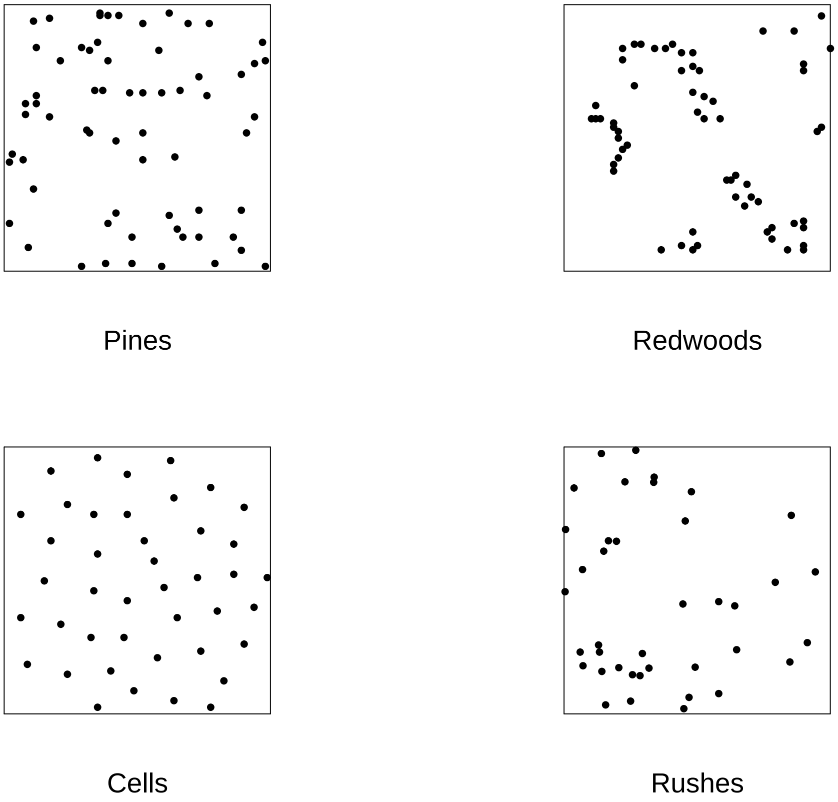
(c) Presence (black) or absence (white) of *Atriplex hymenelytra* on a grid of quadrats in Death Valley, CA.



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



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



Pines

Redwoods

Cells

Rushes

B. 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).
- 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, line segments, or curves
  - be regularly or irregularly spaced
  - be regularly or irregularly shaped
  - belong to a Euclidean or non-Euclidean space

  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

C. Three Important Types of Spatial Data

1. Geostatistical data
   - The response variable exists at every point in the study region; however, we observe the response at only a finite number of points or subregions.
   - Examples:
     (a) Annual acid rain deposition in U.S.
     (b) Richness of iron ore within an ore body
2. Lattice data
   - The response variable exists and is observed only on a finite set of points or 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)
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.

The distinctions between these three types are not always clearcut. In particular, lattice data and geostatistical data have many similarities. In a sense, lattice data are not as refined as geostatistical data or spatial point patterns since you can obtain lattice data by various reductions of the other two.

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

D. Basic Notation and Statistical Model

1. Notation

   - Space $\mathcal{S}$, which we will usually assume to be Euclidean, i.e. $\mathcal{S} = R^d$ where $d = 1$, 2, or 3. We will emphasize the case $d = 2$ but not to the complete exclusion of the other cases.
   - Arbitrary point in $\mathcal{S}$, $\mathbf{s} \in \mathcal{S}$.
   - Study region, $A \subset \mathcal{S}$.
   - Spatial locations $S_1, \ldots, S_n$, $S_i \in D \subset A$
     - observed
     - usually (but not necessarily) distinct, i.e. there is usually no replication at sites (at a single time)

   The only kinds of locations that we consider are points and regions.

   - Responses $\mathbf{Z}(S_1), \ldots, \mathbf{Z}(S_n)$. In general these are multivariate, but we will emphasize the univariate case.
   - Covariates $\mathbf{X}(S_1), \ldots, \mathbf{X}(S_n)$.

2. Model: $\{\mathbf{Z}(S), \mathbf{X}(S)\colon\ S \in D\}$.

   - This is a stochastic process, i.e. a collection of random variables, indexed by points or regions in D.
   - Either the $\mathbf{Z}$-values or S-values or both are random. The $\mathbf{X}$-values are assumed to be nonrandom; or if they are random, all inference is regarded as conditional on the observed values.

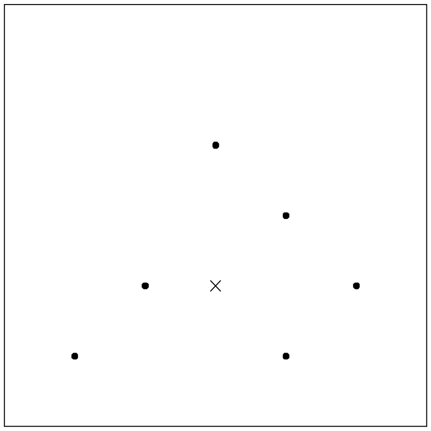
E. Spatial Statistics — Why Bother?

(a) Characterizing the spatial structure of the data may be of direct interest. More on this will follow.

(b) The spatial structure may not be of direct interest, but modelling or otherwise accounting for it may improve other inferences.

Examples:

- (Geostatistics.) Prediction of an unobserved response, $Z(\mathbf{s}_0)$, where $\mathbf{s}_0$ is a specified point site.

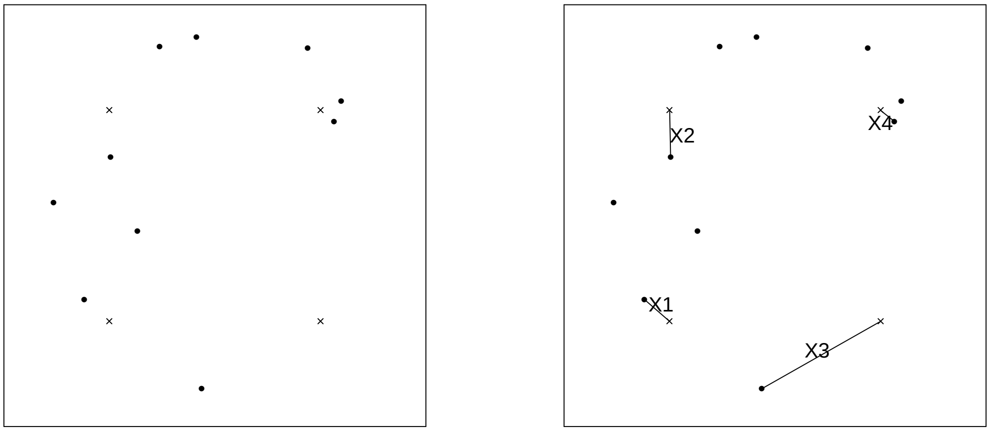


If all of the observed responses are uncorrelated with each other and with $Z(\mathbf{s}_0)$, then $\bar{Z}$ (the average of the observed responses) is the best linear unbiased predictor.

If, however, the responses are spatially correlated, then $\bar{Z}$ is inefficient.

- (Spatial point patterns.) 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.



If tree locations are *completely spatially random* (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.

- (Lattice data analysis.) Variance of the sample mean.

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

Consider 16 responses taken over sites forming a 4 × 4 square grid. Suppose the observations $Z(i,j)$ 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 $\text{var}(\bar{Z}) = 1/16 = 0.0625$.

- (Lattice 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:

  1. Randomized block design
  2. First-order nearest-neighbor balanced design

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

  If treatment-adjusted responses are independent across blocks but positively spatially correlated within blocks, then the second design is optimal (and considerably superior to the RBD) in the sense of minimizing the average variance of treatment contrasts.

F. Spatial Structure

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."

- Large-scale structure (Global)
  - Mean function of geostatistical process
  - Intensity of spatial point process
  - Mean vector of lattice data
- Small-scale structure (Local)
  - Variogram, covariance function of geostatistical process (and lattice process)
  - Ripley's K-function, second-order intensity, nearest-neighbor functions for spatial point process
  - Neighbor weights for lattice process

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.

1. Stationarity — the property whereby the behavior of the process is similar across all of A. This implies:
   - constant large-scale structure
   - small-scale structure which depends on the spatial locations only through their relative positions

2. 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.

G. Main Objectives of Spatial Statistics

1. Inference for spatial structure. Examples:
   - Testing for existence of spatial structure
   - Estimating spatial structural parameters
   - Choosing between alternative models
2. Inference for non-spatial structure. Examples:
   - Estimating treatment effects in spatial experiments
   - Effects of covariates on intensity of a spatial point process
   - Estimation of the number of plants living in a region
3. Prediction of unobserved variables (almost exclusively geostatistical, where it is known as kriging)
4. Design issues, such as where to take observations or how to arrange treatments in a spatial experiment.

H. Temporal Statistics, Spatial Statistics, and Spatio-Temporal Statistics

1. Temporal Statistics vs. Spatial Statistics

   - Inherent difference: Time flows in one direction only, from past to present to future. Not so in space.

   - Contrast between time series analysis and geostatistics/lattice data analysis:

     1. In time series, observations usually are regularly spaced. In geostatistics particularly, but also in lattice data analysis, irregularly spaced data are at least as common as regularly spaced data. So geostatistical and lattice models must be more flexible.
     2. In classical time series models, observations usually are assumed to be dependent but identically distributed (stationarity). In geostatistics and lattice data analysis, observations are usually assumed to be dependent and non-identically distributed; in particular, models usually include a trend.
     3. Due to the unidirectional flow of time, time series models incorporate interaction that results from regarding each observation as dependent on quantities that occurred in the "past" or "present" only. In space, interaction generally occurs in all directions, so most geostatistical/lattice models incorporate omnidirectional interaction.
     4. In time series, prediction usually consists of extrapolating to a future time point. In geostatistics, because we can "go back" in space, interpolation is as important as extrapolation (usually more so).

   - Geostatistics and lattice data analysis are most similar to that subfield of modern longitudinal data analysis which explictly models the temporal correlation among the observations.

     – This similarity is exemplified (and exploited) by SAS PROC MIXED.
     – Key difference: independent replications generally exist in the longitudinal case but not in the spatial case.

   - Spatial point pattern analysis is most similar to failure time data analysis. Some ways that spatial point patterns (SPPs) can be contrasted with failure time data analysis are as follows:

     1. A SPP is usually a window of a process which actually occurs over a larger region. (An FTD may be observed until the process "ends.")
     2. SPP analysis has edge and possibly overlap effects. (Not so for FTD.)
     3. Sometimes we don't observe the whole SPP (sparsely sampled patterns). (FTD are typically not sparsely sampled.)
     4. SPP models feature neighbor interactions prominently. (Not so prominently for FTD.)

2. Spatiotemporal Statistics

- Spatiotemporal data are observations with identifiable and observed spatial *and* temporal labels. Spatiotemporal statistics accounts for these labels in a statistical analysis.

- Types of spatio-temporal data are myriad. The many possibilities result from combining spatial data types with temporal data types and from interactions between spatial and temporal factors governing the data.

- Examples of space-time data:

  – earthquakes (locations random in time and space)

  – change in locations of trees over time (locations random in space but possibly nonrandom in time)

  – environmental monitoring of water quality (locations nonrandom in time and space).

- Possible questions of interest:

  – How are earthquakes clustered in space and time?

  – Do locations of trees at a given time influence the locations of trees at a later time? Is the spatial pattern similar over time?

  – Does a spatial trend in water quality change over time?

- Can model space-time data as:

  – a collection of spatially correlated time series, or

  – a collection of temporally correlated spatial random fields, lattice processes, or spatial point processes

- We'll focus mostly on "pure" spatial (and pure temporal) statistics, but later we will discuss spatiotemporal extensions of certain issues, topics, or methods.

II. EXPLORATORY DATA ANALYSIS FOR TIME SERIES, GEOSTATISTICAL, AND LATTICE DATA

1. Non-spatial (and Non-temporal) Summaries

(a) Numerical summaries: Mean, median, standard deviation, range, etc.

- Reduce the attribute data to a few numbers, but are not that useful here because they ignore the temporal/locational information. (Remember: such data should not be regarded as having come from a single population, as in classical statistics.)

- One relevant R function is `summary()`; for example,
  `summary(so4dep)`
  produces the following output:

```
Min.  1st Qu.  Median Mean 3rd Qu.  Max.
0.0190 0.3320 1.0440 1.2178 1.9250 4.4020
```

(b) Stem-and-leaf display

- Gives a more complete picture of the attribute data than a numerical summary such as the mean, but it has the same defect: because the temporal/locational information is ignored, it gives no indication of the data's temporal/spatial structure. Moreover, it does NOT, as in classical statistics, represent an estimate of a probability distribution from which any particular datum was drawn (unless the observations happen to be iid).

- The relevant R function is `stem()`; for example
  `stem(so4dep)`
  produces the following output:

```
 0  |  2446678999901234455667999
 2  |  1222222334455577889901133346777 8
 4  |  01556889991122377799
 6  |  112459369
 8  |  177814677
10  |  2444890257789
12  |  099924577
14  |  23468847
16  |  11244888889
18  |  0033457901233458
20  |  2260012358889
22  |  13569
24  |  4689112
26  |  34689
28  |  1149113
30  |  46
32  |  2935
34  |
36  |  5
38  |
40  |
42  |
44  |  0
```

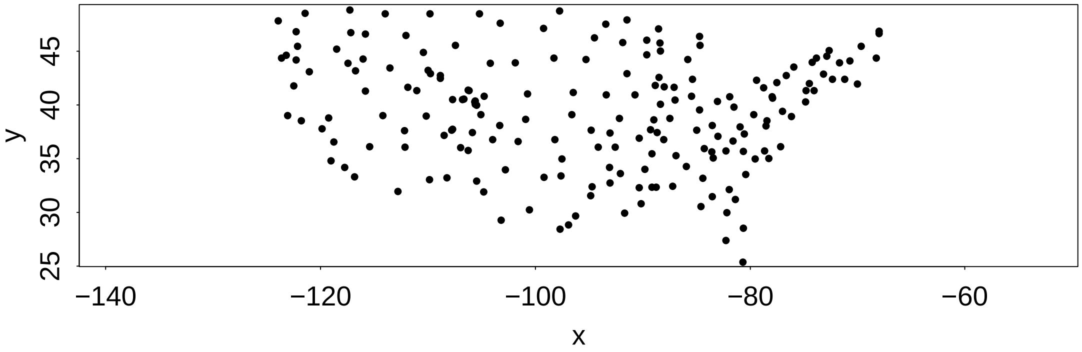
2. Maps of Data Locations and Neighbors

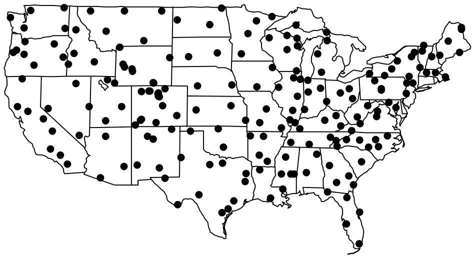
Having explored the attributes “in isolation” on the previous page, we can also explore the data locations stripped of their attribute values.

(a) Scaled map of data locations

- For the sulfate deposition data, we can create two maps by the following S+SpatialStats code:

```
scaled.plot(so4x,so4y)
usa()
points(so4x,so4y)
```

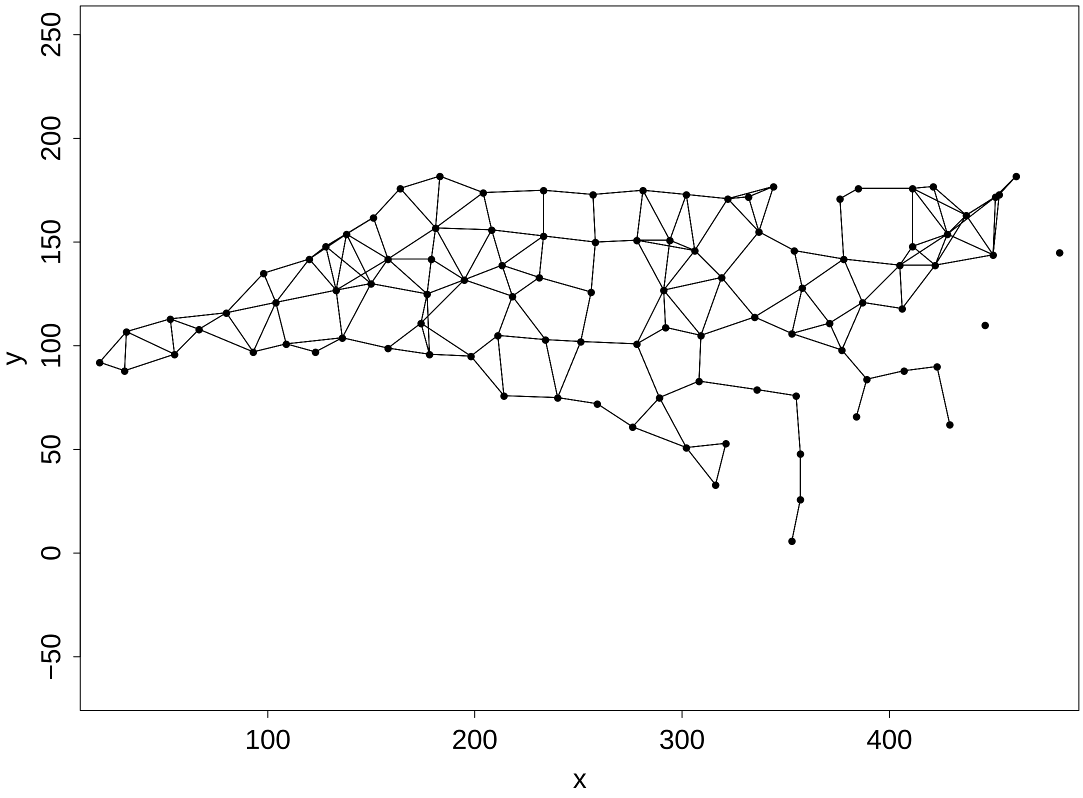




(b) Neighbor identification map

- Each data location is represented by a point, and every neighbor of that location is connected to the point by a line segment.
- For the SIDS data, we can create the plot below using the following S+SpatialStats code:

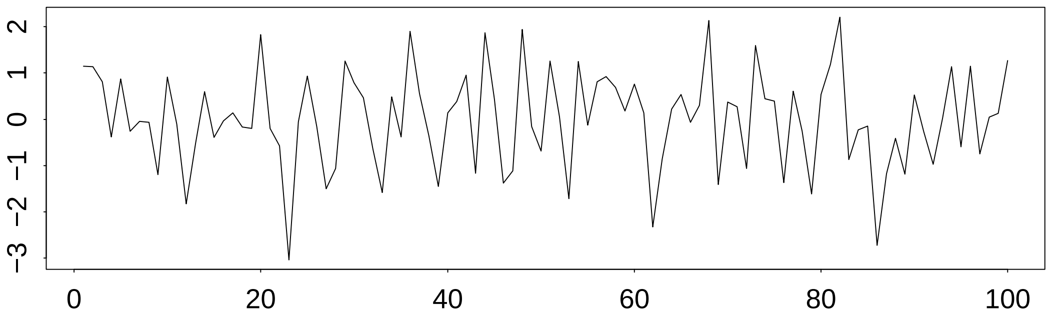
```
attach(sids)
scaled.plot(easting,northing)
segments(easting[sids.neighbor$row.id],northing[sids.neighbor$row.id],
easting[sids.neighbor$col.id],northing[sids.neighbor$col.id])
detach(sids)
```

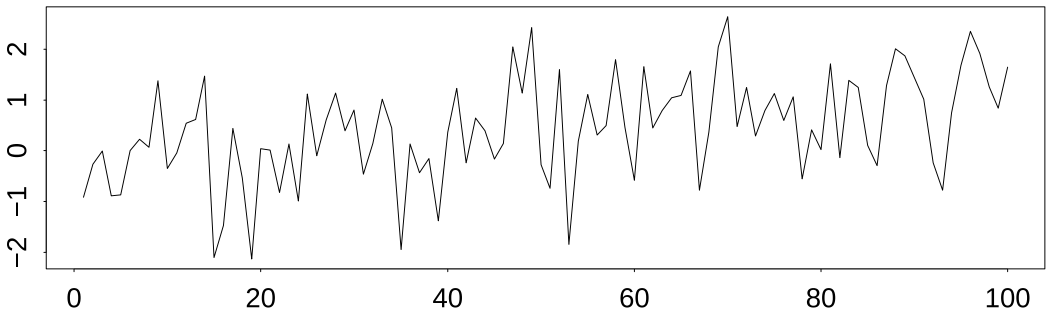


3. Methods used mainly to explore large-scale temporal and spatial variation

(a) Time series plot (for temporal data)

- Simply a plot of attribute versus time, connected in sequence by line segments
- First example below: $X(t)$'s are 100 iid N(0,1) random variables
- Second example below: $Z(t) = 0.01 * t + X^*(t)$ where $X^*(t)$'s are 100 iid N(0,1)'s





(b) 3-D Scatter plot (assuming spatial data with $d = 2$)

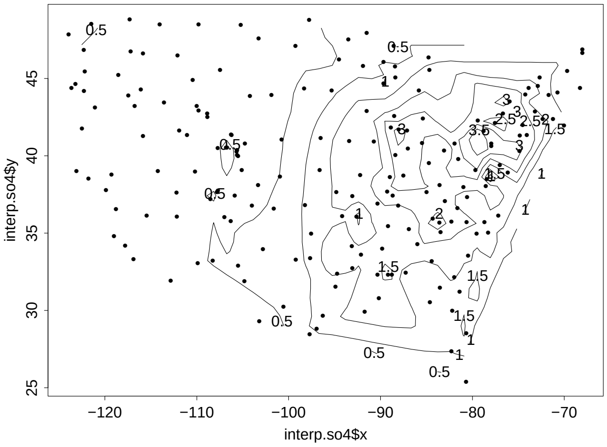
- A plot of Z_i (raw or smoothed values) versus location.
- Gives an impression of the attribute values over space, but interesting features of the data are often obscured.
- It is NOT a depiction of the attribute variable's probability distribution.
- Example (raw values) for the sulfate deposition data — See the first page of these notes.
- Another example (interpolated values) for the sulfate deposition data — using the R function `persp()` — will be seen in the in-class activity.

(c) Plots of Z_i versus each marginal coordinate.

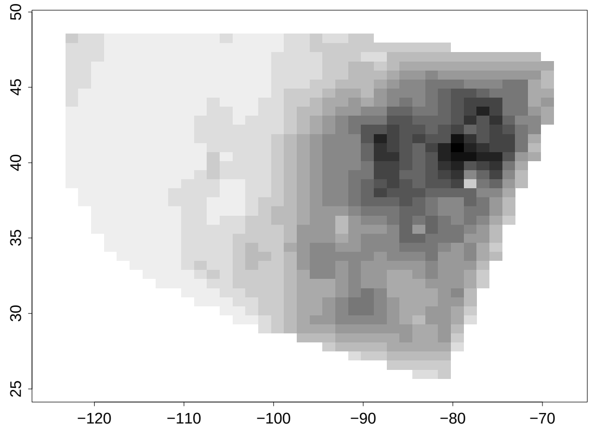
(d) Plot of row or column mean or median of Z_i versus row or column index (assuming that data locations lie on a regular grid or have been assigned to such).

(e) 2-D scatterplot of data locations with symbols indicating whether Z_i is above or below the median of the Z_i's.

(f) Contour plot of smoothed Z_i's



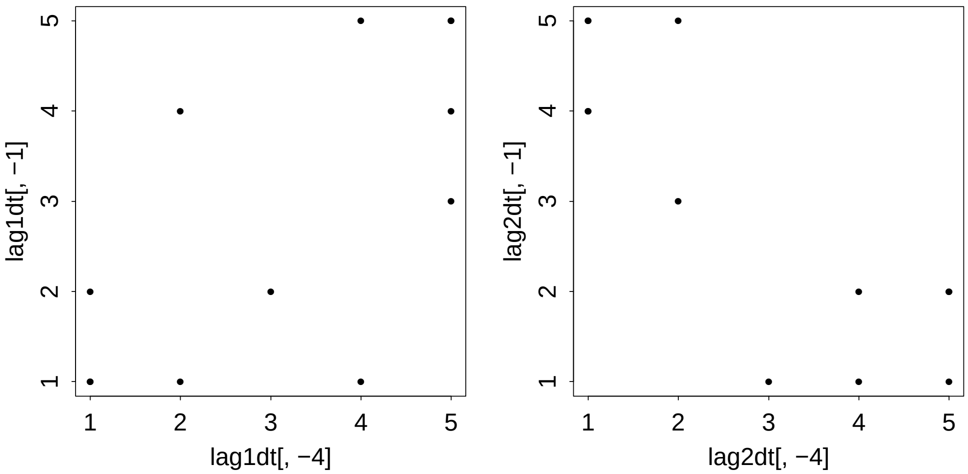
(g) Gray-scale plot of Z_i's



4. Methods used mainly to explore small-scale dependence

(a) Same-lag scatterplots ("h-scatterplots")

- For a fixed vector $\mathbf{e}$ of unit length and a fixed scalar h, plot $Z(\mathbf{s}_i + h\mathbf{e})$ versus $Z(\mathbf{s}_i)$ for all suitable i.
- Requires regular spacing between data locations.
- Positive (negative) correlation in plot indicates positive (negative) spatial dependence at that lag.
- Individually, these plots may reveal outliers.
- Comparisons among the plots may reveal the existence of nonstationarity in the mean and/or variance or the existence of anisotropy.



(b) Sample autocovariance function (acf)

- Measures similarity of attributes at sites (or times) a given "lag" apart
- Plot of $\hat{C}(\mathbf{h}_u)$ versus $\mathbf{h}_u$, where

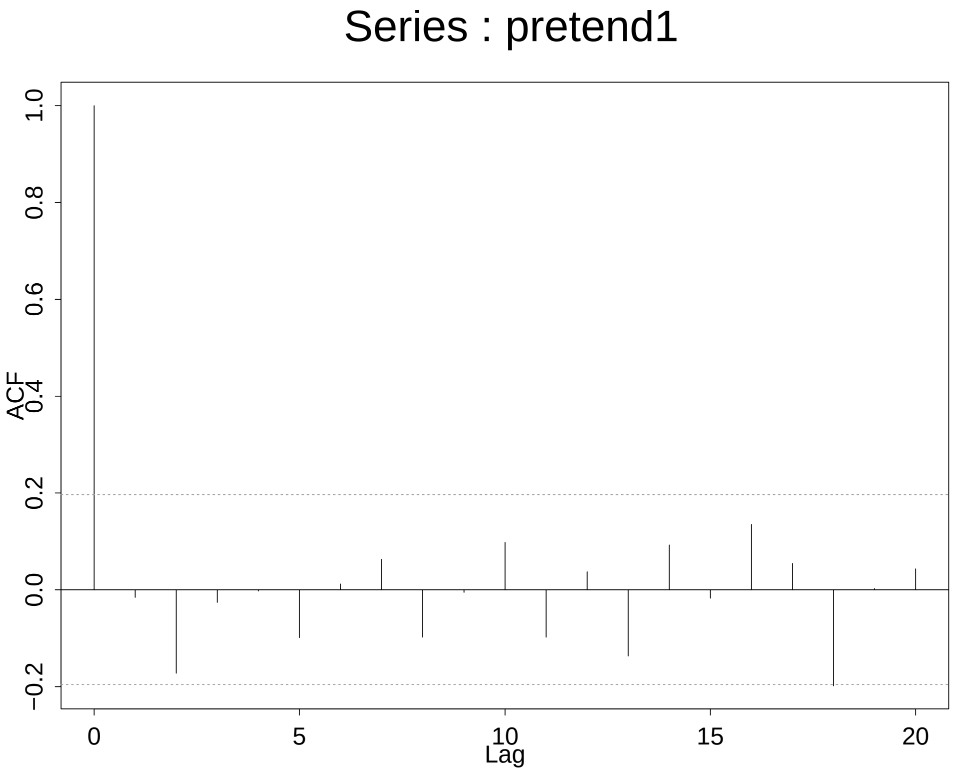
$$\hat{C}(\mathbf{h}_u) = \frac{1}{N(\mathbf{h}_u)} \sum_{\mathbf{s}_i - \mathbf{s}_j = \mathbf{h}_u} (Z(\mathbf{s}_i) - \bar{Z})(Z(\mathbf{s}_j) - \bar{Z}) \qquad (u = 1, \ldots, k).$$

  Here $\mathbf{h}_1, \ldots, \mathbf{h}_k$ are the distinct values of $\mathbf{h}$ represented in the data set, and $N(\mathbf{h}_u)$ is the number of times that lag $\mathbf{h}_u$ occurs in the data set.

- For time series data, it is more simply a plot of $\hat{C}(h)$ versus h, where

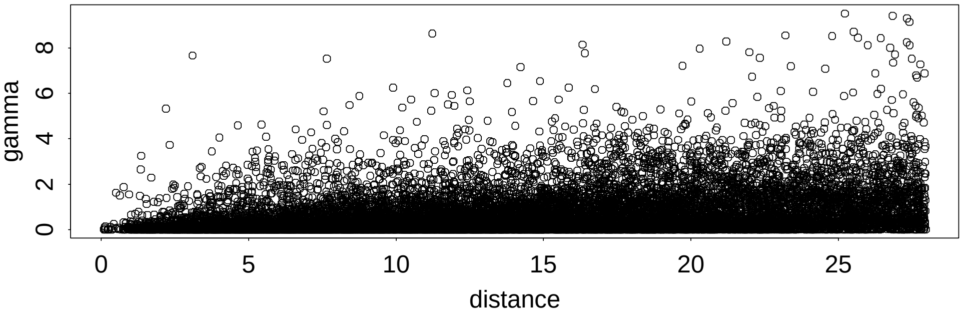
$$\hat{C}(h) = \frac{1}{N-h} \sum_{i=1}^{N-h} (Z(i) - \bar{Z})(Z(i+h) - \bar{Z}).$$

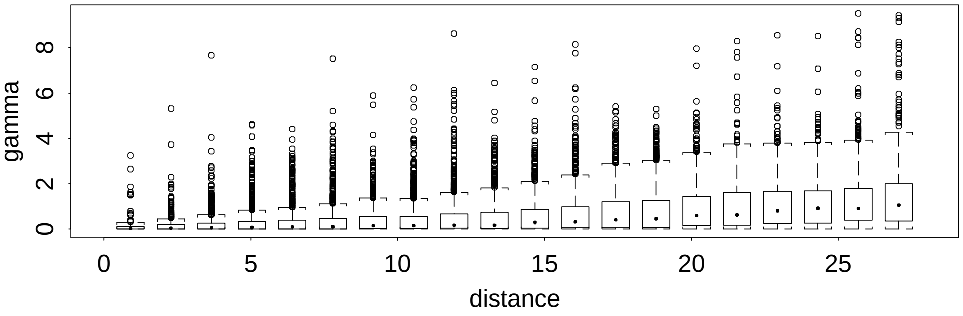
  Below is the acf for the 100 iid N(0,1) random variables in first example on page 22:



(c) Variogram cloud (and square-root-differences cloud)

- Historically has been used for spatial data much more than temporal data.
- Plot $(Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2$ versus $[(\mathbf{s}_i - \mathbf{s}_j)'(\mathbf{s}_i - \mathbf{s}_j)]^{1/2}$ for all possible pairs of observations. (For the square-root-differences cloud, use $|Z(\mathbf{s}_i) - Z(\mathbf{s}_j)|^{1/2}$ as the ordinate instead). For example, for the sulfate deposition data,





- The plot is often an unintelligible mess, hence it may often be advisable to bin the lags and plot a boxplot for each bin.
- A tendency for the cloud to "increase" as distance increases is indicative of positive spatial association.
- Note that isotropy is implicitly assumed.
- The square-root-differences are more resistant to outliers.

(d) Sample semivariogram (or variogram)

- Plot one-half the average squared difference (or, for the variogram, merely the average squared difference) of observations lagged the same distance and direction apart, versus the lag.

- We assume here that the data are regularly spaced; the more general case of irregularly spaced data will be considered later.

- Formally, we plot $\hat{\gamma}(\mathbf{h}_u)$ versus $\mathbf{h}_u$, where

$$\hat{\gamma}(\mathbf{h}_u) = \frac{1}{2N(\mathbf{h}_u)} \sum_{\mathbf{s}_i - \mathbf{s}_j = \mathbf{h}_u} \{Z(\mathbf{s}_i) - Z(\mathbf{s}_j)\}^2$$
$$(u = 1, \ldots, k),$$

  $\mathbf{h}_1, \ldots, \mathbf{h}_k$ are the distinct values of $\mathbf{h}$ represented in the data set, and $N(\mathbf{h}_u)$ is the number of times that lag $\mathbf{h}_u$ occurs in the data set.

- The variogram is a measure of dissimilarity.

- Stationarity of some kind is implicity assumed.

- Toy example:

- If $d = 2$, you can display as a 3-D plot or you can superimpose it for a few selected directions (e.g. N-S, NE-SW, E-W, and SE-NW) on the same 2-D plot.

- If isotropy is assumed, you can plot in 2-D for both cases $d = 2$ and $d = 3$.

- Will discuss in much more detail later.

(e) 3-D plot of correlation range versus spatial location, computed from a moving window.

5. Methods used mainly to explore small-scale variability

(a) 3-D plot of standard deviation versus spatial location, computed from a moving window

- May reveal nonstationarity in variability, and indicate which portion(s) of A is (are) more variable than the rest

(b) Scatterplot of standard deviation versus mean, computed from a moving window

- May also reveal nonstationarity in variability, but differently from the previous method.

(c) For areal lattice data: Scatterplot of attribute versus area (or population) of subregion

6. Methods used mainly for detecting outliers

Data often include *outliers*, i.e., attribute observations that look suspiciously atypical. In the case of geostatistical data, outliers may be of two types:

1. Distributional outliers — observations that seem unusual with respect to the data's overall distribution

2. Spatial outliers — observations that may not be unusual with respect to the data's overall distribution, but are unusual with respect to their neighbors

Methods that graph and summarize the data should ideally be relatively unaffected by outliers and should be able to highlight them. Such methods are called *resistant*.

Outlier Detection Methods:

- Stem-and-leaf display — can identify distributional outliers, but not necessarily spatial outliers
- Plots of sample means *and* medians versus row index or column index (assumes that the data lie on a rectangular grid or can be assigned to such).

- Compute, for each row and column,

  $$n^{1/2}|\bar{Z} - \mathrm{med}(Z_i)|/[\sigma(0.5708)^{1/2}]$$

  where $\bar{Z}$ is the sample mean and $\mathrm{med}(Z_i)$ is the sample median within that row or column and $\sigma = \mathrm{IQR}/1.349$. Assess as a standard normal deviate. Values of 3 or larger are of concern.
- Median polish — large residuals correspond to outliers.
- Plot of each datum versus its nearest neighbor (or versus the average of its m nearest neighbors, or versus the average of all observations within a fixed distance δ of it).

  A hypothetical example:

III. GENERAL MODEL

Popular statistical model for many kinds of data:

Datum = mean + residual

where the mean is a nonrandom quantity (a number) that may vary from datum to datum and the residual is a random variable with mean zero.

- For a random sample from a single distribution (the classical statistics paradigm), the mean would be the same for each datum, the residuals would all have the same variance and the residuals would be independent.
- For the classical regression situation, the mean is taken to be a linear function of unknown parameters and the residuals are taken to be iid, with mean zero.

Neither the random sampling nor classical regression assumptions are generally appropriate in the contexts we consider in this course; nevertheless, the basic model is still useful and takes the form

$$Z(\mathbf{s}) = m(\mathbf{s}) + \epsilon(\mathbf{s}).$$

Here:

- $\mathbf{s}$ generally represents a point in space, but it could also represent a point in time.
- $m(\mathbf{s}) \equiv E[Z(\mathbf{s})]$ is the *mean function*.
- $\{\epsilon(\mathbf{s})\colon \mathbf{s} \in D\}$ is a zero-mean stochastic process, also called a *random field*.

Further remarks:

- For a time series, this model defines a random time series plot. If D is a two-dimensional region, this model defines a random surface over D.
- This model decomposes the total variation into large-scale variation (the mean function) and small-scale variation (the residual process).

Note: The observed data represent a sparse sample from a single realization of $\{Z(\mathbf{s})\colon \mathbf{s} \in D\}$.

1. Models for the Mean Function

The principle of spatial (or temporal) continuity suggests that sites close to one another in space (or time) should have similar means, but sites far apart need not. This leads to the postulation of a continuous, relatively smooth (but otherwise unconstrained) model for the mean function.

A very useful class of mean functions are the polynomials. For example, in two dimensions, with a site's coordinates denoted as $\mathbf{s} = (x, y)$, the full first-order and second-order polynomials are as follows:

$$m(x, y) = \beta_0 + \beta_1 x + \beta_2 y$$

$$m(x, y) = \beta_0 + \beta_1 x + \beta_2 y + \beta_{11} x^2 + \beta_{12} xy + \beta_{22} y^2$$

The polynomials are a class of *linear* mean functions, i.e. functions of the form

$$m(\mathbf{s}) = \sum_{j=1}^{p} \beta_j f_j(\mathbf{s})$$

where $f_1(\mathbf{s}), \ldots, f_p(\mathbf{s})$ are known functions.

To emphasize the dependence of a mean function on an unknown vector of parameters $\boldsymbol{\beta}$, we may write it as $m(\mathbf{s}; \boldsymbol{\beta})$.

Other possibilities for mean functions:

- Nonlinear functions (smooth)
- Trigonometric functions (smooth and periodic)
- Median polish surface (continuous but less smooth)
- Nonparametric smooth functions (splines, LOESS)

2. Small-Scale Variation

To make statistical progress we must at least partially specify the behavior of $\{\epsilon(\mathbf{s})\colon \mathbf{s} \in D\}$. Since its first-order properties have been fully specified (zero mean), we focus on second-order properties.

Assume that $\{\epsilon(\mathbf{s})\colon \mathbf{s} \in D\}$ has associated with it a *covariance function*, which expresses the covariance between two values of $\epsilon(\cdot)$ as a function of the coordinates of the two corresponding sites. That is,

$$C(\mathbf{s}, \mathbf{t}) \equiv \mathrm{cov}\{\epsilon(\mathbf{s}), \epsilon(\mathbf{t})\} = \mathrm{cov}\{Z(\mathbf{s}), Z(\mathbf{t})\}.$$

This function satisfies the following two properties:

1. *Symmetry*, i.e., $C(\mathbf{s}, \mathbf{t}) = C(\mathbf{t}, \mathbf{s})$ for all $\mathbf{s}, \mathbf{t} \in D$.
2. *Nonnegative definiteness*, i.e.,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j C(\mathbf{s}_i, \mathbf{s}_j) \geq 0$$

for all n, all sequences $\{a_i\colon i = 1, \ldots, n\}$, and all sequences of spatial locations $\{\mathbf{s}_i\colon i = 1, \ldots, n\}$.

If the covariance function is well-defined, then we can also define the *correlation function*

$$\rho(\mathbf{s}, \mathbf{t}) = \mathrm{corr}\{Z(\mathbf{s}), Z(\mathbf{t})\} = \frac{C(\mathbf{s}, \mathbf{t})}{[C(\mathbf{s}, \mathbf{s})C(\mathbf{t}, \mathbf{t})]^{1/2}}.$$

Note: If $m(\mathbf{s})$ and $C(\mathbf{s}, \mathbf{t})$ were completely known, the distribution of $\{Z(\mathbf{s})\colon \mathbf{s} \in D\}$ still would not be completely specified. If, however, we also assume that $\{\epsilon(\mathbf{s})\colon \mathbf{s} \in D\}$ is a Gaussian process, then the distributions of $\{\epsilon(\mathbf{s})\colon \mathbf{s} \in D\}$ and $\{Z(\mathbf{s})\colon \mathbf{s} \in D\}$ are completely determined.

3. Stationarity

To simplify the consideration of models for the covariance function and to make statistical inference more feasible, it is helpful to assume *stationarity*, which asserts that even a single realization of $\{\epsilon(\mathbf{s})\colon \mathbf{s} \in D\}$ has a kind of replication built into it.

Two types of stationarity (for now):

1. Strict stationarity — requires that the joint probability distribution of $\{\epsilon(\mathbf{s})\colon \mathbf{s} \in D\}$ depends only on the relative positions of sites, i.e.,

$$F_{\mathbf{s}_1+\mathbf{h},\ldots,\mathbf{s}_m+\mathbf{h}}(e_1, \ldots, e_m) = F_{\mathbf{s}_1,\ldots,\mathbf{s}_m}(e_1, \ldots, e_m)$$

   for all m; $\mathbf{s}_1, \ldots, \mathbf{s}_m$; $\mathbf{h}$; and all $e_1, \ldots, e_m$. This implies, for example, that $P[\epsilon(\mathbf{s}_1+\mathbf{h}) \leq e_1] = P[\epsilon(\mathbf{s}_1) \leq e_1]$ for all $\mathbf{s}_1$, $\mathbf{h}$, and e_1.

2. Second-order stationarity — requires that:

   - the mean is constant (which has already been assumed);
   - the way that two values of $\epsilon(\cdot)$ co-vary is consistent for values at sites having the same relative positions. That is, the covariance between variates at two sites depends on only the sites' relative positions. This can be expressed in either of the following two ways:

     (a) $C(\mathbf{s}, \mathbf{t}) = C(\mathbf{s} + \mathbf{h}, \mathbf{t} + \mathbf{h})$ for "all" $\mathbf{h}$.

     (b) $C(\mathbf{s}, \mathbf{t}) = C(\mathbf{h})$, where $\mathbf{h} = \mathbf{s} - \mathbf{t}$, for all $\mathbf{s}, \mathbf{t} \in D$.

In practice, an assumption of second-order stationarity is often sufficient for inference purposes, and so it will be one of our basic assumptions.

IV. MODELS FOR TIME SERIES DATA

We now consider models for data observed at equally spaced time points $t = 1, 2, \ldots, n$. Assume initially that the mean is zero at all times, i.e. $m(t) \equiv 0$.

Important models (for the covariance structure):

1. Moving average model of order 1, MA(1)

   $$Z(t) = a_t - \theta_1 a_{t-1},$$

   where the a_t's are iid $\mathrm{N}(0, \sigma_a^2)$ and θ_1 is an unconstrained parameter.

   Observe that

   $$\begin{aligned} \mathrm{var}(Z(t)) &= \\ \mathrm{cov}(Z(t), Z(t+1)) &= \\ \mathrm{cov}(Z(t), Z(t+h)) &= \end{aligned}$$

   Thus,

   $$\rho(h) =$$

2. Extension of MA(1) to MA(q)

   $$Z(t) = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q}$$

   where the a_t's are iid $\mathrm{N}(0, \sigma_a^2)$ and $\theta_1, \ldots, \theta_q$ are unconstrained parameters.

   This is also stationary. Expressions for $C(h)$ and $\rho(h)$ can be given, but they're more "messy" than in the first-order case.

3. Autoregressive model of order 1, AR(1)

   $$Z(t) = \phi_1 Z(t-1) + e_t,$$

   where the e_t's are iid $\mathrm{N}(0, \sigma_e^2)$ and ϕ_1 is a parameter satisfying $|\phi_1| < 1$, and we "start up" the process by defining $Z(0) \sim N(0, \frac{\sigma_e^2}{1-\phi_1^2})$, independently of the e_t's.

   With some effort it can be shown that

   $$\begin{aligned} \mathrm{var}(Z(t)) &= \\ \mathrm{cov}(Z(t), Z(t+1)) &= \\ \mathrm{cov}(Z(t), Z(t+h)) &= \\ \rho(h) &= \end{aligned}$$

4. Extension of AR(1) to AR(p)

$$Z(t) = \phi_1 Z(t-1) + \phi_2 Z(t-2) + \cdots + \phi_p Z(t-p) + e_t,$$

   where the e_t's are iid N$(0, \sigma_e^2)$ and $\phi_1, \ldots, \phi_p$ are parameters satisfying certain complicated constraints.

   This is also stationary, provide that appropriate "start-ups" are specified. Expressions for $C(h)$ and $\rho(h)$ can be given, but they're more "messy" than in the first-order case.

5. ARMA models, e.g. ARMA(1,1)

$$Z(t) = \phi_1 Z(t-1) + a_t - \theta_1 a_{t-1}$$

6. Random walk

$$Z(t) = Z(t-1) + a_t,$$

   where the a_t's are iid N$(0, \sigma_a^2)$ and $Z(0) \equiv 0$.

   Note: Unlike the previous examples, a random walk is nonstationary (see problem in Homework 2).

To allow for nonzero mean or nonconstant mean in these models, we simply subtract the mean function from each $Z(t)$. For example, the model equations for the MA(1) and AR(1) with nonconstant means are respectively as follows:

$$\begin{aligned} Z(t) - m(t) &= a_t - \theta_1 a_{t-1}, \\ Z(t) - m(t) &= \phi_1[Z(t-1) - m(t-1)] + e_t \end{aligned}$$

For example, we might take $m(t) = \beta_0$ (constant mean) or $m(t) = \beta_0 + \beta_1 t$ (linear trend).

V. MODELS FOR GEOSTATISTICAL DATA

We now consider models for data observed at possibly irregularly spaced points within a two-dimensional (or even higher-dimensional) region where the process is also defined.

1. Isotropy

We have already defined the notions of a mean function, a covariance function, and stationarity. In two and more dimensions, we also have the notion of *isotropy*, which further simplifies modeling and inference.

Isotropy of $\{\epsilon(\mathbf{s})\colon \mathbf{s} \in D\}$ requires that the covariance between any two values of $\epsilon(\cdot)$ depends only on the Euclidean distance between their corresponding locations, i.e.,

$$C(\mathbf{h}) = C(\|\mathbf{h}\|) = C([\mathbf{h}'\mathbf{h}]^{1/2}) \text{ for "all" } \mathbf{h}.$$

Under isotropy the *equicorrelation contours*, i.e. the locations of all the variates that are equally correlated with any given variate, lie on the perimeter of a circle (in 2-D space) centered at the given variate's location.

Types of anisotropy:

1. *Geometric anisotropy.* A covariance function is geometrically anisotropic if a positive definite matrix $\mathbf{A}$ exists such that

$$C(\mathbf{h}) = C([\mathbf{h}'\mathbf{A}\mathbf{h}]^{1/2}) \text{ for "all" } \mathbf{h}.$$

   - The equicorrelation contours are ellipses (in 2-D).
   - Isotropy can be regarded as the special case in which $\mathbf{A} = \mathbf{I}$.

2. *Zonal anisotropy* — Any kind of anisotropy that is not geometric.

2. Models for the Covariance Function

As for the mean function, the principle of spatial continuity suggests that we consider relatively smooth functions as models for covariance functions. However, there are two important differences in modeling covariance functions:

1. Not every function satisfies the inherent mathematical requirements of a covariance function.

2. Among those functions that do satisfy the inherent mathematical requirements of a covariance function, they may do so only for certain parameter values. That is, the parameters may be *constrained.*

The mathematical requirements of a second-order stationary covariance function are as follows:

1. Evenness, i.e.
$$C(\mathbf{h}) = C(-\mathbf{h}) \text{ for ``all'' } \mathbf{h}.$$

2. Nonnegative definiteness, i.e.
$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j C(\mathbf{s}_i - \mathbf{s}_j) \geq 0$$
for all n, all sequences $\{a_i\colon\ i = 1, \ldots, n\}$, and all sequences of spatial locations $\{\mathbf{s}_i\colon$ $i = 1, \ldots, n\}$.

Bochner's Theorem from analysis tells us, in effect, that any real-valued characteristic function of a d-dimensional random vector $\mathbf{X}$ is even and nonnegative definite, and thus any real-valued d-dimensional characteristic function could serve as a *valid* covariance function in R^d.

The properties of evenness and nonnegative definiteness imply the following facts:

1. $C(\mathbf{0}) \geq 0$

2. $|C(\mathbf{h})| \leq C(\mathbf{0})$ for all $\mathbf{h}$

Furthermore, the principle of spatial continuity and some other related notions suggest that we mainly (though not exclusively) consider models for which:

1. $C(\mathbf{h})$ decreases as the inter-site distance, $\|\mathbf{h}\| \equiv (\mathbf{h}'\mathbf{h})^{1/2}$, increases in any given direction.

2. $C(\mathbf{h}) \to 0$ as $\|\mathbf{h}\|$ increases.

3. $C(\mathbf{h}) \geq 0$ for all $\mathbf{h}$.

Examples of isotropic covariance function models (letting $r = \|\mathbf{h}\|$):

- *Triangular (tent, piecewise linear) model* (valid in R^1 only)

$$C(r;\boldsymbol{\theta}) = \begin{cases} \theta_1(1 - r/\theta_2) & \text{for } 0 \le r \le \theta_2 \\ 0 & \text{for } r > \theta_2 \end{cases} \qquad (\theta_1 \ge 0, \theta_2 \ge 0)$$

- *Spherical model*

$$C(r;\boldsymbol{\theta}) = \begin{cases} \theta_1\left(1 - \frac{3r}{2\theta_2} + \frac{r^3}{2\theta_2^3}\right) & \text{for } 0 \le r \le \theta_2 \\ 0 & \text{for } r > \theta_2 \end{cases} \qquad (\theta_1 \ge 0, \theta_2 \ge 0)$$

- *Exponential model*

$$C(r;\boldsymbol{\theta}) = \theta_1 \exp(-r/\theta_2) \qquad (\theta_1 \ge 0, \theta_2 \ge 0)$$

- *Gaussian model*

$$C(r;\boldsymbol{\theta}) = \theta_1 \exp(-r^2/\theta_2) \qquad (\theta_1 \ge 0, \theta_2 \ge 0)$$

- *Rational quadratic model*

$$C(r;\boldsymbol{\theta}) = \theta_1(\theta_2 - \frac{r^2}{1+r^2/\theta_2}) \qquad (\theta_1 \geq 0, \theta_2 \geq 0)$$

- *Cosine model*

$$C(r;\boldsymbol{\theta}) = \theta_1 \cos(r/\theta_2) \qquad (\theta_1 \geq 0, \theta_2 \geq 0)$$

- *Wave or hole-effect model*

$$C(r;\boldsymbol{\theta}) = \theta_1\theta_2\frac{\sin(r/\theta_2)}{r} \qquad (\theta_1 \geq 0, \theta_2 \geq 0)$$

- *Matern class of models*

$$C(r;\boldsymbol{\theta}) = \theta_1\frac{1}{2^{\theta_3-1}\Gamma(\theta_3)}(\frac{2r\sqrt{\theta_3}}{\theta_2})^{\theta_3}K_{\theta_3}(\frac{2r\sqrt{\theta_3}}{\theta_2}) \qquad (\theta_1 \geq 0, \theta_2 \geq 0, \theta_3 > 0)$$

- Compare to the "no-correlation" or nugget effect model:

$$C(r;\boldsymbol{\theta}) = \begin{cases} \theta_1 & \text{for } r = 0 \\ 0 & \text{for } r > 0 \end{cases} \qquad (\theta_1 \geq 0)$$

Some notes about these covariance models:

- The triangular, exponential, spherical, Gaussian, rational quadratic, and Matern models all decrease monotonically to 0 as distance increases; thus they conform to the principle of spatial continuity.
- The wave, or hole-effect model, also tends to 0 as distance increases, but not monotonically. It would be appropriate if there is some kind of periodic co-variation in the data that damps out over space.
- The cosine model does not tend to 0 as distance increases. It would be appropriate only if there is periodic co-variation that does not damp out over space.
- Exponential model = Matern model with $\theta_3 = \frac{1}{2}$; Gaussian model = Matern model as $\theta_3 \to \infty$.

Attributes of these covariance models:

- *Scale parameter, or variance.* For most of these models, $C(0) = \theta_1$. Thus for these models, θ_1 is the constant variance of the random field. For the other models, θ_1 affects the variance.

- *Correlation scale parameter.* For each of these models, θ_2 is a parameter that controls how fast the covariance changes.

- *Range, or effective range.*

  - The range of an isotropic covariance function, if one exists, is defined as the distance beyond which the covariance function is equal to 0. Of the models listed, only the triangular and spherical models have a range (which is equal to θ_2 for them).
  - For isotropic models that do not have a range, the effective range, if one exists, is defined as the distance beyond which the covariance function does not exceed $0.05\times$variance. The exponential, Gaussian, rational quadratic, Matern, and wave models all have effective ranges; the cosine model, however, does not.

- *Continuity at 0.* Observe that all of these covariance functions are continuous at 0, except for the nugget effect model. Continuity of a covariance function at 0 has implications for the behavior of the random field, as we will see shortly.

- *Shape parameter.* For the Matern model, θ_3 controls the "shape" of the function near 0.

From basic covariance models such as the ones just listed, we can construct more complicated models using the following rules:

- A valid isotropic covariance function in R^{d_1} is a valid isotropic covariance function in R^{d_2} if $d_1 > d_2$. The converse, however, is not true; a counterexample is the triangular model, which is valid in R^1 but not in higher dimensions. With the exception of the triangular model, all the models we've listed are valid in R^2 and R^3.

- If $C_1(\cdot)$ and $C_2(\cdot)$ are valid covariance functions in R^d, then so is $C(\cdot) \equiv C_1(\cdot) + C_2(\cdot)$.

  An important special case of this construction occurs when $C_1(\cdot)$ is the "no-correlation" model, in which case we can write the sum as

$$C(r; \boldsymbol{\theta}) = \begin{cases} \theta_0 + C_2(0; \boldsymbol{\theta}) & \text{for } r = 0 \\ C_2(r; \boldsymbol{\theta}) & \text{for } r > 0. \end{cases}$$

  θ_0 is called the *nugget effect.* A covariance function that has a nonzero nugget effect is discontinuous at zero.

- If $C_0(\cdot)$ is a valid covariance function in R^d and $b > 0$, then $C(\cdot) \equiv b \cdot C_0(\cdot)$ is a valid covariance function in R^d.

- If $C_1(\cdot)$ and $C_2(\cdot)$ are valid covariance functions in R^{d_1} and R^{d_2}, respectively, then $C(\cdot) \equiv C_1(\cdot) \times C_2(\cdot)$ is a valid covariance function in $R^{d_1+d_2}$.

3. Intrinsic Stationarity and the Semivariogram

Traditionally, geostatistical practitioners have adopted a slightly more general kind of stationarity assumption on $\{\epsilon(\mathbf{s})\colon \mathbf{s} \in D\}$ than second-order stationarity, and they have modelled the small-scale spatial variation through a function somewhat different than the covariance function.

The more general stationarity assumption is called *intrinsic stationarity*, which specifies that:

1. The mean is constant;
2. $\frac{1}{2}\text{var}[\epsilon(\mathbf{s}) - \epsilon(\mathbf{t})]$ $(= \frac{1}{2}\text{var}[Z(\mathbf{s}) - Z(\mathbf{t})])$ depends only on the lag $\mathbf{s} - \mathbf{t}$, i.e.,

$$\frac{1}{2}\text{var}[\epsilon(\mathbf{s}) - \epsilon(\mathbf{t})] = \gamma(\mathbf{s} - \mathbf{t}), \text{ for all } \mathbf{s}, \mathbf{t} \in D.$$

The function $\gamma(\cdot)$ defined by the second condition above is called the *semivariogram*.

A “typical” isotropic semivariogram:

Remarks:

- The semivariogram can be defined by the same expression for a non-intrinsically stationary process, but in that case it must be represented as $\gamma(\mathbf{s},\mathbf{t})$.
- The semivariogram can also be expressed as follows:

$$\gamma(\mathbf{h}) = \frac{1}{2}E\{[\epsilon(\mathbf{s}) - \epsilon(\mathbf{t})]^2\} = \frac{1}{2}E\{[Z(\mathbf{s}) - Z(\mathbf{t})]^2\}, \text{ where } \mathbf{h} = \mathbf{s} - \mathbf{t}.$$

- A second-order stationary random process with covariance function $C(\cdot)$ is intrinsically stationary, with semivariogram

$$\gamma(\mathbf{h}) = C(\mathbf{0}) - C(\mathbf{h}),$$

  but the converse is not true in general. That is, intrinsic stationarity is more general than second-order stationary. Proof:

- A *valid* semivariogram model must satisfy four mathematical properties:

  1. Vanishes at $\mathbf{0}$, i.e., $\gamma(\mathbf{0}) = 0$.
  2. Evenness, i.e., $\gamma(-\mathbf{h}) = \gamma(\mathbf{h})$.
  3. Conditional nonpositive definiteness, i.e.,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \gamma(\mathbf{s}_i - \mathbf{s}_j) \leq 0$$

     for all n, all $\mathbf{s}_1, \ldots, \mathbf{s}_n$, and all $a_1, \ldots, a_n$ such that $\sum_{i=1}^{n} a_i = 0$.
  4. $\lim_{\|\mathbf{h}\|\to\infty}\{\gamma(\mathbf{h})/\|\mathbf{h}\|^2\} = 0$.

  A function that satisfies these four properties is called a *valid* semivariogram model. Validity of a semivariogram in R^{d_1} implies validity in R^{d_2} if $d_1 > d_2$, but the converse is not true.

The same issues we noted for covariance models suggest that we mainly (though not exclusively) consider semivariogram models that increase as distance increases.

Semivariogram attributes (these don't always exist):

- Sill $= \lim_{\|\mathbf{h}\|\to\infty} \gamma(\mathbf{h})$
- Range or effective range
- Nugget effect $= \lim_{\|\mathbf{h}\|\to 0} \gamma(\mathbf{h})$
- Slope

Examples of valid models for isotropic semivariograms:

- *Triangular* (valid in R^1 only).

$$\gamma(r;\boldsymbol{\theta}) = \begin{cases} \theta_1 r/\theta_2 & \text{for } 0 \le r \le \theta_2 \\ \theta_1 & \text{for } r > \theta_2 \end{cases} \qquad (\theta_1 \ge 0, \theta_2 \ge 0)$$

- *Spherical*

$$\gamma(r;\boldsymbol{\theta}) = \begin{cases} \theta_1 \left(\frac{3r}{2\theta_2} - \frac{r^3}{2\theta_2^3}\right) & \text{for } 0 < r \le \theta_2 \\ \theta_1 & \text{for } r > \theta_2 \end{cases} \qquad (\theta_1 \ge 0, \theta_2 \ge 0)$$

- *Exponential*

$$\gamma(r;\boldsymbol{\theta}) = \theta_1\{1 - \exp(-r/\theta_2)\} \qquad (\theta_1 \ge 0, \theta_2 \ge 0)$$

- *Gaussian*

$$\gamma(r;\boldsymbol{\theta}) = \theta_1\{1 - \exp(-r^2/\theta_2)\} \qquad (\theta_1 \ge 0, \theta_2 \ge 0)$$

- *Cosine*

$$\gamma(r;\boldsymbol{\theta}) = \theta_1\{1 - \cos(r/\theta_2)\} \qquad (\theta_1 \ge 0, \theta_2 \ge 0)$$

- *Linear*

$$\gamma(r;\boldsymbol{\theta}) = \theta_1 r \qquad\qquad (\theta_1 \geq 0)$$

- *Power*

$$\gamma(r;\boldsymbol{\theta}) = \theta_1 r^{\theta_2} \qquad\qquad (\theta_1 \geq 0, 0 \leq \theta_2 < 2)$$

- *Logarithmic*

$$\gamma(r;\boldsymbol{\theta}) = \theta_1 \log r \qquad\qquad (\theta_1 \geq 0)$$

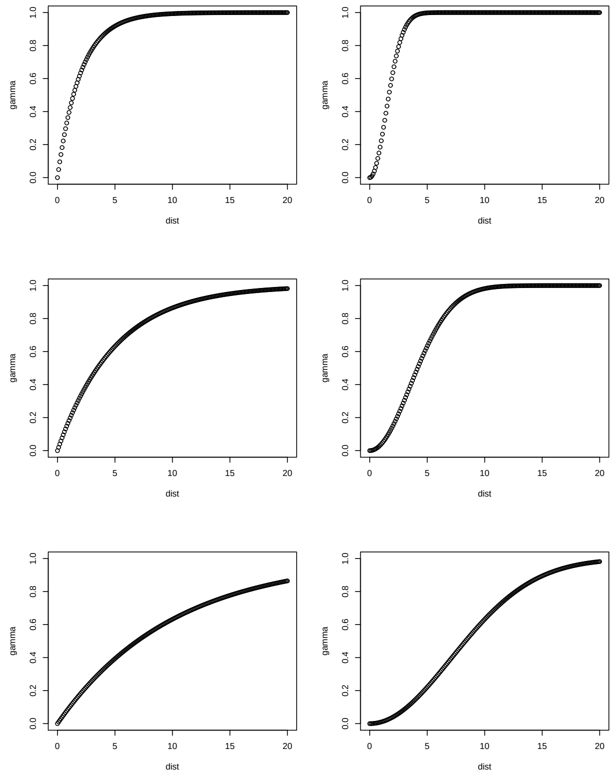
- *No-correlation (pure nugget) model*

$$\gamma(r;\boldsymbol{\theta}) = \begin{cases} 0 & \text{for } r = 0 \\ \theta_1 & \text{for } r > 0 \end{cases} \qquad\qquad (\theta_1 \geq 0)$$

For second-order stationary processes, the spatial dependence can be described by either the covariance function or the semivariogram. Each has its advantages.

On this page and the one that follows, we demonstrate that the "range parameter" of the exponential and Gaussian semivariograms in the `gstat` package of R does not coincide with the range, or the effective range, as we've defined it here; and we display what the "partial sill parameter" is.

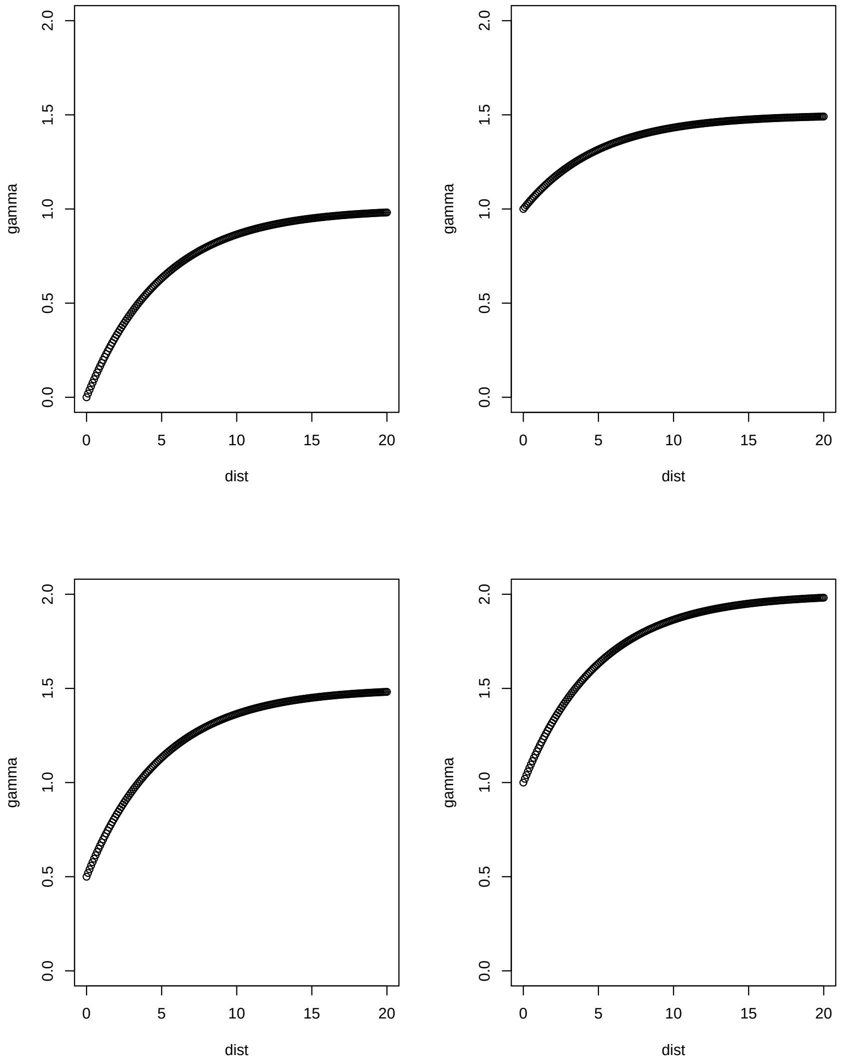
The exponential (left column) and Gaussian (right column) semivariograms displayed below all have a sill of 1.0 and a nugget of 0. The first, second, and third rows of plots have *range parameters* 2, 5, and 10, respectively. In fact, for the exponential model the range parameter, a, is related to the *practical range* (PR) as follows: $PR = a \times \ln 20 \doteq 3a$. This is obtained by putting



$$1 - \exp(-r/\theta_1) = 0.95,$$

replacing r and θ_1 with PR and a, respectively, and solving for PR in terms of a. The relationship between a and PR for the Gaussian model is similarly found to be as follows: $PR \doteq \sqrt{3}a$.

The exponential semivariograms displayed below all have a range parameter of 5. Upper left plot: partial sill=1.0, nugget=0, sill=1.0. Lower left plot: partial sill =1.0, nugget=0.5, sill=1.5. Upper right plot: partial sill=0.5, nugget=1.0, sill=1.5. Lower right plot: partial sill=1.0, nugget=1.0, sill=2.0.



4. Modeling Anisotropy

(a) Range anisotropy

- Kind of anisotropy seen most often in practice.
- *Geometric anisotropy* is the easiest to model. Any valid isotropic model can be generalized to make it geometrically anisotropic. This is done by replacing $\|\mathbf{h}\|$ in the isotropic model by $(\mathbf{h}'\mathbf{A}\mathbf{h})^{1/2}$, where $\mathbf{A}$ is a $d \times d$ positive definite matrix.
- Generally we don't know the true value of $\mathbf{A}$, so the sensible thing to do is to regard its elements as unknown parameters.
- For example, a geometrically anisotropic exponential covariance function in R^2 is

$$C(\mathbf{h};\theta) = \theta_1 \exp[-\theta_2(h_1^2 + 2\theta_3 h_1 h_2 + \theta_4 h_2^2)^{1/2})]$$

  - Note: $\theta_3 = 0$, $\theta_4 = 1 \Leftrightarrow$ isotropy
  - Also note: $\theta_3 = 0$, $\theta_4 = 4 \Leftrightarrow$ spatial correlation is twice as persistent in the E-W direction as in the N-S direction; and correlation strength in all other directions is intermediate between these two. [E-W range = 2 × N-S range]
- Non-geometric range anisotropy is possible, but seems to occur rarely.

(b) Sill anisotropy

It can be shown that if the sill exists but is direction-dependent, then either:

- a second-order stationary model is appropriate but the spatial correlation does not vanish in every direction as inter-site distance increases;
- the assumption of second-order stationarity is violated; or
- there are measurement errors which are correlated or do not all have mean zero.

Either of the first two possibilities implies that standard estimation methods (yet to be described) are ill-advised.

(c) Nugget anisotropy

- Can be caused by correlated measurement errors.
- Typically occurs in one direction only, which is not difficult to model.

(d) Slope anisotropy

Can be dealt with in a similar fashion as geometric range anisotropy.

5. The Classical Geostatistical Model, in Summary

The general model we have constructed, which will form the basis for our analysis of geostatistical data, is as follows:

$$Z(\mathbf{s}) = m(\mathbf{s}; \boldsymbol{\beta}) + \epsilon(\mathbf{s})$$

where:

- $m(\cdot; \boldsymbol{\beta})$ is a specified family of continuous functions
- $\boldsymbol{\beta}$ is a vector of unknown, unconstrained parameters
- $\{\epsilon(\mathbf{s})\colon \mathbf{s} \in D\}$ is a second-order stationary or intrinsically stationary process with mean zero and valid covariance function $C(\cdot; \boldsymbol{\theta})$ or semivariogram $\gamma(\cdot; \boldsymbol{\theta})$
- $\boldsymbol{\theta}$ is a vector of unknown parameters constrained to lie in a parameter space Θ.

If the mean function is linear, then the model for the observed data can be written in vector/matrix form as follows:

$$\mathbf{Z} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where:

- $\mathbf{X} = \begin{pmatrix} f_1(\mathbf{s}_1) & f_2(\mathbf{s}_1) & \ldots & f_p(\mathbf{s}_1) \\ f_1(\mathbf{s}_2) & f_2(\mathbf{s}_2) & \ldots & f_p(\mathbf{s}_2) \\ \vdots & \vdots & & \vdots \\ f_1(\mathbf{s}_n) & f_2(\mathbf{s}_n) & \ldots & f_p(\mathbf{s}_n) \end{pmatrix}$
- $\mathrm{var}(\mathbf{Z}) = \mathrm{var}(\boldsymbol{\epsilon}) = \mathbf{V} = \mathbf{V}(\boldsymbol{\theta})$.

VI. INFERENCE FOR GEOSTATISTICAL DATA

1. Overview of the Geostatistical Method

1. Using exploratory techniques, prior knowledge, etc., posit a model for $\{Z(\mathbf{s})\colon \mathbf{s} \in D\}$ of the form we described in the previous unit.

2. Estimate $\boldsymbol{\beta}$ in $m(\mathbf{s}; \boldsymbol{\beta})$, e.g. by ordinary least squares, median polish, or some other method that does not require knowledge of the second-order dependence structure.

3. Using fitted residuals (if necessary) from the previous step, estimate $\gamma(\mathbf{h})$ or $C(\mathbf{h})$ nonparametrically and plot it (in several directions).

4. Select a valid semivariogram model $\gamma(\mathbf{h}; \boldsymbol{\theta})$ or covariance function $C(\mathbf{h}; \boldsymbol{\theta})$ that is compatible with the plot from the previous step.

5. Fit the chosen model to the estimated semivariogram or covariance function to estimate the model's parameters.

6. Using the fitted semivariogram or covariance function, re-estimate $\boldsymbol{\beta}$ by estimated generalized least squares (or by some other method that accounts for correlation among observations).

7. Repeat Steps 3-6, if desired.

8. "Krige" (i.e. predict) unobserved values at sites (or over regions) of your choosing and estimate the corresponding variances of prediction error.

9. Determine optimal locations to take additional observations, and repeat Steps 1-8, if desired.

2. Estimating the Mean Function by Ordinary Least Squares Methods

If the mean function $m(\mathbf{s}; \boldsymbol{\beta})$ is a linear (or nonlinear) function of the elements of $\boldsymbol{\beta}$, then linear (or nonlinear) least squares can be used to fit the mean function to the data.

At this point we will briefly review the most important aspects of multiple linear regression analysis for us, starting with notation:

- Observations are available on $p+1$ scalar variables Z, x_1, x_2, $\ldots$, x_p from each of n items or individuals.
- Denote these observations as

$$\begin{gathered} Z_1,\, x_{11},\, x_{12},\, \ldots,\, x_{1p} \\ Z_2,\, x_{21},\, x_{22},\, \ldots,\, x_{2p} \\ \vdots \\ Z_n,\, x_{n1},\, x_{n2},\, \ldots,\, x_{np} \end{gathered}$$

- We wish to study how the observed values of Z (known as the dependent or response variable) might be explained by the observed values of the remaining variables (called the explanatory variables).
- A *linear model* specifies that the relationship between Z and the explanatory variables is of the form

$$Z_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ip}\beta_p + \epsilon_i, \qquad i = 1, \ldots, n \tag{1}$$

  where $\beta_1, \ldots, \beta_p$ are fixed, unknown parameters and ϵ_i is an unobservable random disturbance or residual.
- The model is said to be linear because apart from the residual ϵ_i, Z_i is a linear combination of the unknown parameters, i.e. $Z_i - \epsilon_i = \sum_{j=1}^{p} x_{ij}\beta_j$.
- In vector/matrix notation, the linear model can be rewritten as

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

  where

$$\mathbf{Z} = \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ & \vdots & & \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

Classical Assumptions:

- The x_{ij}'s are nonrandom (thus so is $\mathbf{X}$). We will also assume that they are linearly independent (thus $\mathbf{X}$ has full rank).
- The ϵ_i's are uncorrelated and have common mean 0 and common unknown positive variance σ^2 (thus $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\text{var}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$).
- The ϵ_i's are jointly normally distributed (thus $\boldsymbol{\epsilon} \sim \text{N}(\mathbf{0}, \sigma^2\mathbf{I})$).

Important Inference Problems:

- Estimation (point and interval) of the elements of $\boldsymbol{\beta}$ and of other linear functions of the form $\mathbf{c}'\boldsymbol{\beta}$.
- Estimation of σ^2.
- Hypothesis testing on the elements of $\boldsymbol{\beta}$ and on functions $\mathbf{c}'\boldsymbol{\beta}$.
- Prediction (point and interval) of "new" Z-observations corresponding to specified values of the x_{ij}'s.

Ordinary Least Squares Estimation (OLSE):

- OLSE of $\boldsymbol{\beta}$ is the value of $\boldsymbol{\beta}$ that minimizes the residual sum of squares criterion,
$$RSS(\boldsymbol{\beta}) = \sum_{i=1}^{n}(Z_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 = (\mathbf{Z} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta}).$$
- Equivalently, $\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}$.
- OLSE is also the best (minimum variance) unbiased estimator and the maximum likelihood estimator (MLE).
- $\text{var}(\hat{\boldsymbol{\beta}}_{OLS}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$
- Fitted regression equation is
$$\hat{Z} = x_1\hat{\beta}_{1,OLS} + x_2\hat{\beta}_{2,OLS} + \cdots + x_p\hat{\beta}_{p,OLS}$$
- Standard estimator of σ^2 is $\hat{\sigma}^2_{OLS} = RSS(\hat{\boldsymbol{\beta}}_{OLS})/(n-p)$.
- To test H_0: $\mathbf{C}\boldsymbol{\beta} = \mathbf{d}$ vs. H_A: $\mathbf{C}\boldsymbol{\beta} \neq \mathbf{d}$ at α level of significance, compare
$$\frac{(\mathbf{C}\hat{\boldsymbol{\beta}}_{OLS} - \mathbf{d})'[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}']^{-1}(\mathbf{C}\hat{\boldsymbol{\beta}}_{OLS} - \mathbf{d})}{\hat{\sigma}^2_{OLS} \cdot (\#\text{ rows of }\mathbf{C})}$$
to upper α percentage point of $F(\#\text{ rows of }\mathbf{C}, n-p)$ distribution.

Adaptation of Classical Regression Model to Spatial Data:

- Suppose that all of the variables are observed at distinct, known locations.
- Suppose that the locations are in 2-D space. Then denote a generic location by $\mathbf{s} = (u, v)$ and attach this as a label to each variable, i.e. $Z_i \to Z(\mathbf{s}_i)$ and $x_{ij} \to x_j(\mathbf{s}_i)$.
- The model can be written as

$$Z(\mathbf{s}_i) = x_1(\mathbf{s}_i)\beta_1 + x_2(\mathbf{s}_i)\beta_2 + \cdots + x_p(\mathbf{s}_i)\beta_p + \epsilon(\mathbf{s}_i).$$

- Some of the explanatory variables may be explicit functions of geographic coordinates, while others may not be.
- Example 1: Snow water equivalent (SWE) in western U.S.

$$\begin{aligned}
Z(\mathbf{s}) &= \text{April 1 SWE at } \mathbf{s} \\
x_1(\mathbf{s}) &= u \quad \text{(longitude)} \\
x_2(\mathbf{s}) &= v \quad \text{(latitute)} \\
x_3(\mathbf{s}) &= \text{elevation at } \mathbf{s} \\
x_4(\mathbf{s}) &= \text{slope at } \mathbf{s} \\
x_5(\mathbf{s}) &= \text{aspect at } \mathbf{s} \\
x_6(\mathbf{s}) &= \text{average winter wind speed at } \mathbf{s} \\
x_7(\mathbf{s}) &= \text{1 if } \mathbf{s} \text{ is on windward side of mountain range, 0 otherwise}
\end{aligned}$$

Trend Surface Analysis:

- If all of the explanatory variables in the classical linear regression model are explicit functions of geographic coordinates, then we have a *trend surface model.* Trend surface analysis is merely an ordinary least squares regression analysis of such a model.
- In trend surface analysis, the geographic coordinates (and functions thereof) may be serving as proxies for explanatory variables that we did not or could not observe.
- The most commonly-used family of trend surface models is the family of polynomial functions of geographic coordinates.

  First-order polynomial (planar) trend surface model:

$$Z(\mathbf{s}) = Z(u, v) = \beta_1 + \beta_2 u + \beta_3 v + \epsilon(u, v)$$

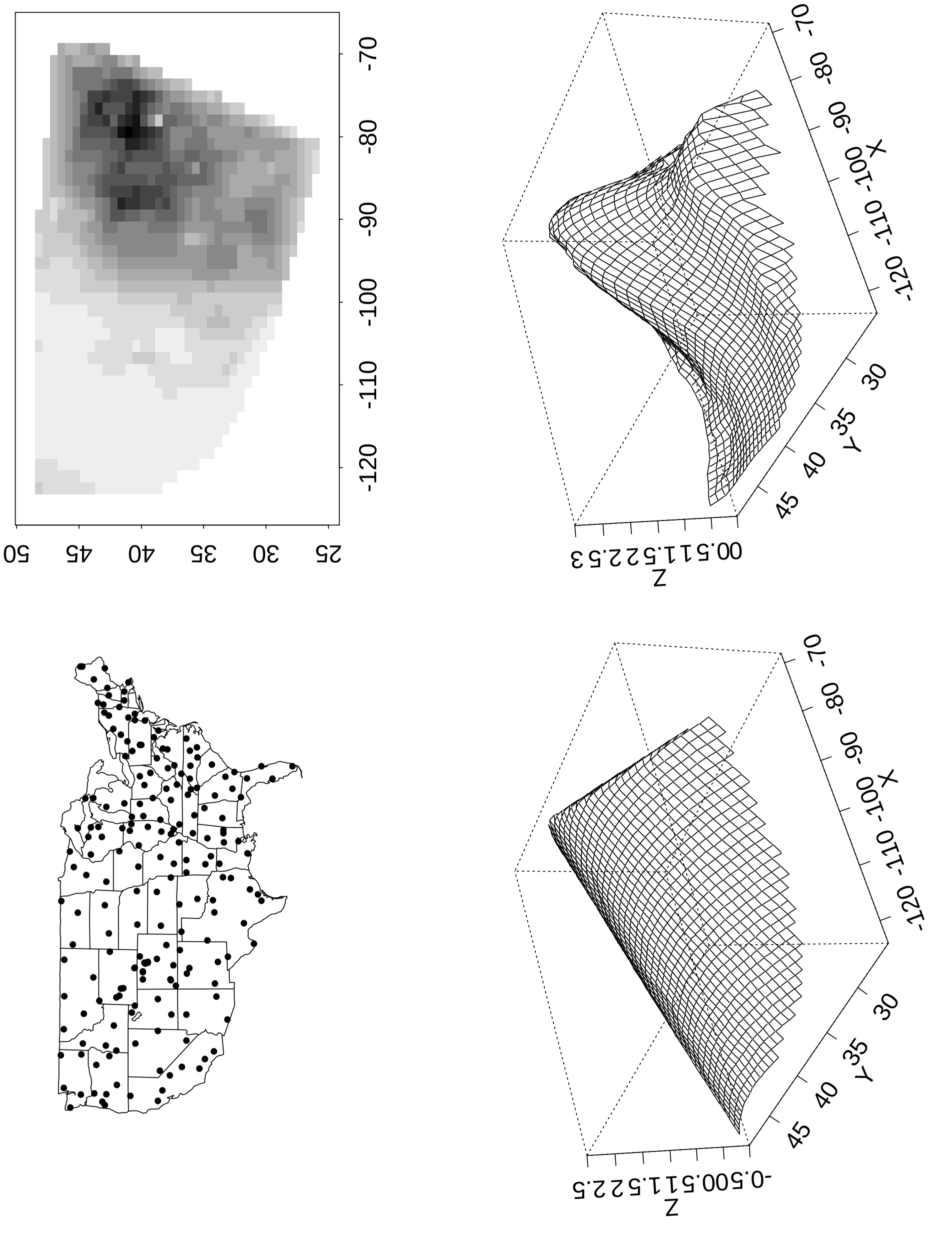
  Second-order polynomial (quadratic) trend surface model:

$$Z(\mathbf{s}) = \beta_1 + \beta_2 u + \beta_3 v + \beta_4 u^2 + \beta_5 uv + \beta_6 v^2 + \epsilon(u, v)$$

  Third-order polynomial (cubic) trend surface model:

$$Z(\mathbf{s}) = \beta_1 + \beta_2 u + \beta_3 v + \beta_4 u^2 + \beta_5 uv + \beta_6 v^2 + \beta_7 u^3 + \beta_8 u^2 v + \beta_9 uv^2 + \beta_{10} v^3 + \epsilon(u, v)$$

- Example of second-order and fifth-order polynomial surfaces fitted to wet sulfate deposition data (annual totals) from NADP/NTN network in 1987:



Trend surface analysis is quite easy to implement, since the necessary computing software (PROC REG in SAS, lm in R) is widely available. It does have some drawbacks/limitations/problems, however:

- the fitting procedure is very sensitive to outliers;
- the regression variables (which in this case are the data's spatial coordinates or functions thereof) tend to be highly correlated (multicollinearity), which causes the fitting procedure to be numerically unstable;
- the fitted surface in portions of the study region where there are no data tends to be distorted so as to better fit the observed data.

Often these drawbacks can be finessed, but only at the expense of making the fitting procedure more complicated.

For spatial data observed on a rectangular grid, there is another potentially useful family of trend surface models: row-column effects models

$$Z(\mathbf{s}) = Z(u, v) = \beta_1 + \beta_{1+u} + \beta_{1+u+v} + \epsilon(u, v)$$

where u is the column number and v is the row number.

Features/drawbacks:

- Eliminates distortion and multicollinearity problem.
- Still sensitive to outliers.
- Implies row-column additivity (no interaction).
- Restricted to data whose locations lie on a rectangular grid.

Methodologies have developed to deal with outliers (e.g. median polish fitting) and lack of flexibility (nonparametric regression).

3. Estimating the Mean Function by Median Polish

- Assume that the data lie on the nodes of a $p \times q$ rectangular grid $\{(x_l, y_k)\colon\ k = 1, \ldots, p;\ l = 1, \ldots, q\}$ (or have been assigned to such). Regard the grid nodes as cells in a 2-way table.
- Median polish takes the model for the mean function to be

$$m(x_l, y_k; \boldsymbol{\beta}) = a + r_k + c_l.$$

- The model is fit by operating iteratively on the data in this table, alternately subtracting row medians and column medians, and accumulating these medians in an extra column and row of cells.
- Repeat this procedure until another iteration produces virtually no change.
- Final entries in the extra cells are the median polish estimates of row effects $r_1, \ldots, r_p$, column effects $c_1, \ldots, c_q$, and an overall effect a.
- Final entries in the body of the table are residuals $\hat{\epsilon}_{kl}$ such that

$$Z(x_l, y_k) = \hat{a} + \hat{r}_k + \hat{c}_l + \hat{\epsilon}_{kl}.$$

- The relevant R function is `medpolish()`.

Toy example:

Completing a median polish surface over the study area:

- The fitted values at data locations are

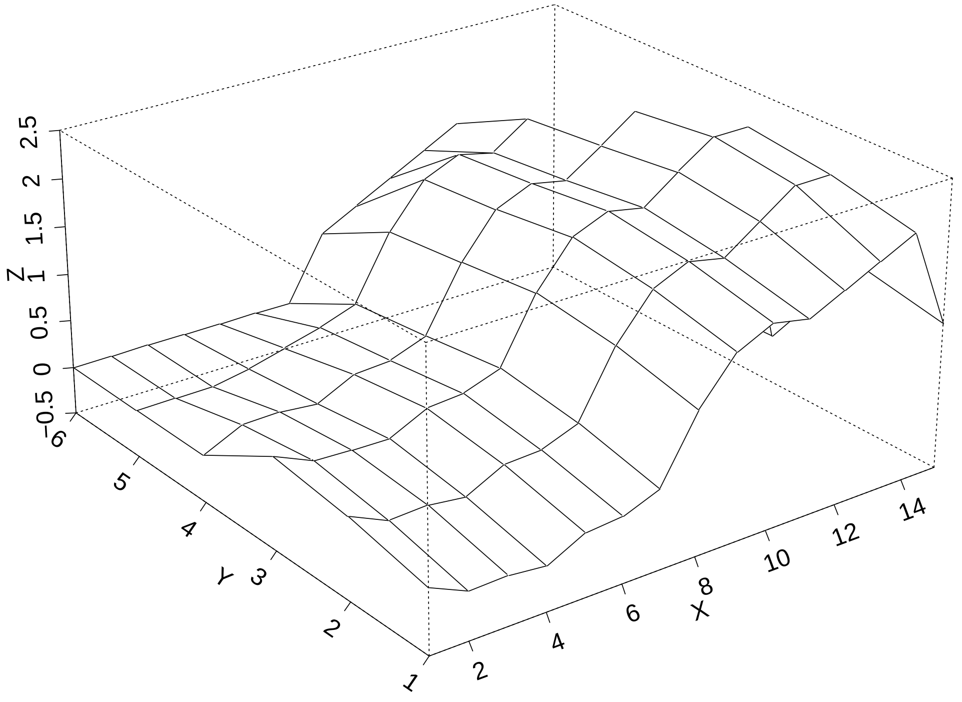
$$m(x_l, y_k; \hat{\boldsymbol{\beta}}) = \hat{a} + \hat{r}_k + \hat{c}_l.$$

- Construction of a continuous surface over all of A from the mean/median polish fits at data locations can be done by planar interpolation. For $\mathbf{s} = (x, y)'$ in the rectangular region between the four nodes $(x_l, y_k)', (x_{l+1}, y_k)', (x_l, y_{k+1})'$, $(x_{l+1}, y_{k+1})'$, where $x_l < x_{l+1}$ and $y_k < y_{k+1}$, the fit is given by the planar interpolant

$$\hat{a} + \hat{r}_k + \left(\frac{y - y_k}{y_{k+1} - y_k}\right)(\hat{r}_{k+1} - \hat{r}_k) + \hat{c}_l + \left(\frac{x - x_l}{x_{l+1} - x_l}\right)(\hat{c}_{l+1} - \hat{c}_l).$$

  Similar formulas are applicable for extrapolation when $x < x_1, x > x_q, y < y_1$, or $y > y_p$.

Example: Sulfate deposition data. (Since data locations are irregularly spaced, rows and columns of a two-way table were formed by superimposing a 6 × 15 square grid of spacing 4 deg latitude and longitude.)



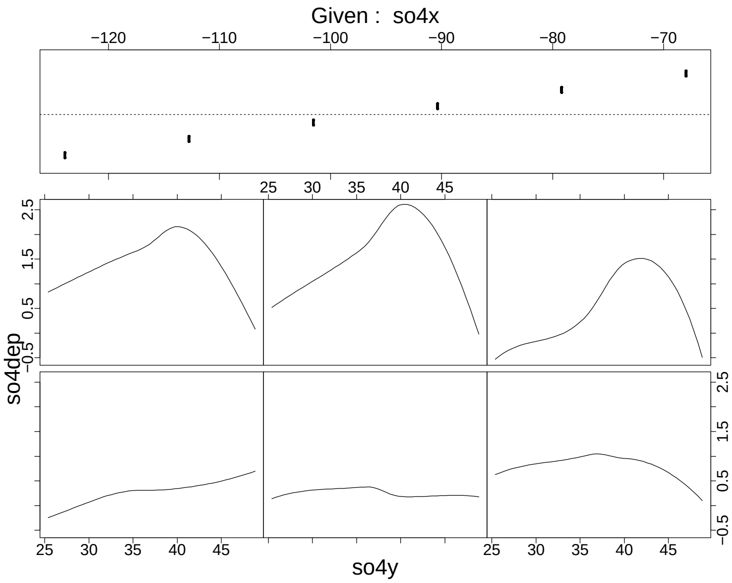
Median polish avoids some of the pitfalls of least squares:

- more resistant to outliers
- residuals from a median polish are less biased than those from OLS

However, the assumed row-column additivity and the blurring associated with irregularly spaced data are drawbacks.

4. Estimating the Mean Function by Locally Weighted Least Squares (lowess)

- Only assumes that the mean function is smooth.
- Estimates this smooth trend in a moving fashion by fitting a site-specific first-order or second-order polynomial to only the most proximate data to a site. For example, for the sulfate deposition data, we might use only data within a 100-mile radius of any given site.
- Fits the trend using weighted least squares, with weights inversely related to distance from the site.
- The relevant R function is `loess()`.



Note: None of the fitting methods in this unit have accounted for spatial correlation among the responses. But they are the necessary precursor for estimating spatial correlation in the next step of the geostatistical method.

5. Nonparametric Semivariogram/Covariance Function Estimation

The raw ingredients for semivariogram estimation are either:

- the observations $\{Z_1, \ldots, Z_n\}$, if the mean function is taken to be constant;
- the residuals
$$\hat{\epsilon}(\mathbf{s}_i) = Z(\mathbf{s}_i) - m(\mathbf{s}_i; \hat{\boldsymbol{\beta}}), \ (i = 1, \ldots, n)$$
from a fitted mean function at the data locations, otherwise.

Assume for simplicity that the study region is in 2-D space and that there are no replicate measurements at data locations. Also assume, initially, that the data locations lie on a regular rectangular grid.

The basic idea is to estimate $\gamma(\mathbf{h})$ by one-half the average squared difference of responses or residuals whose data locations are lagged by $\mathbf{h}$.

The *sample semivariogram* (also empirical or estimated semivariogram):

$$\begin{aligned} \hat{\gamma}(\mathbf{h}_u) &= \frac{1}{2N(\mathbf{h}_u)} \sum_{\mathbf{s}_i - \mathbf{s}_j = \mathbf{h}_u} \{\hat{\epsilon}(\mathbf{s}_i) - \hat{\epsilon}(\mathbf{s}_j)\}^2 \\ & \quad (u = 1, \ldots, k). \end{aligned}$$

- Here $\mathbf{h}_1, \ldots, \mathbf{h}_k$ are the distinct values of $\mathbf{h}$ represented in the data set.
- Attention can be restricted to lags with displacement angles in the interval $[0, \pi)$ since $\gamma(\mathbf{h})$ is an even function.
- $N(\mathbf{h}_u)$ is the number of times that lag $\mathbf{h}_u$ occurs in the data set. (We don't double-count.)
- This is a method-of-moments type estimator.
- The estimator is unbiased if the observations themselves are intrinsically stationary, and approximately unbiased under the model in which the mean is nonconstant.

When data locations are irregularly spaced, we partition the *lag space* $H = \{\mathbf{s} - \mathbf{t}\colon\ \mathbf{s}, \mathbf{t} \in D\}$ into lag classes or "windows" $H_1, \ldots, H_k$, say, and assign each lag in the data set to one of these classes. Then we use a similar estimator:

$$\begin{aligned} \hat{\gamma}(\mathbf{h}_u) &= \frac{1}{2N(H_u)} \sum_{\mathbf{s}_i - \mathbf{s}_j \in H_u} \{\hat{\epsilon}(\mathbf{s}_i) - \hat{\epsilon}(\mathbf{s}_j)\}^2 \\ & \quad (u = 1, \ldots, k). \end{aligned}$$

- Here $\mathbf{h}_u$ is a representative lag for the whole lag class H_u; typically $\mathbf{h}_u$ is taken to be the centroid of H_u or the average of all the realized lags in the lag class.

- $N(H_u)$ is the class frequency of H_u.

- The estimator is approximately unbiased; it is not exactly unbiased even in the case where intrinsically stationary observations themselves are used because the grouping of lags into classes causes a blurring effect.

- Two main types of partitions:

  1. "Polar" partitioning, i.e., angle and distance classes

  2. Rectangular partitioning

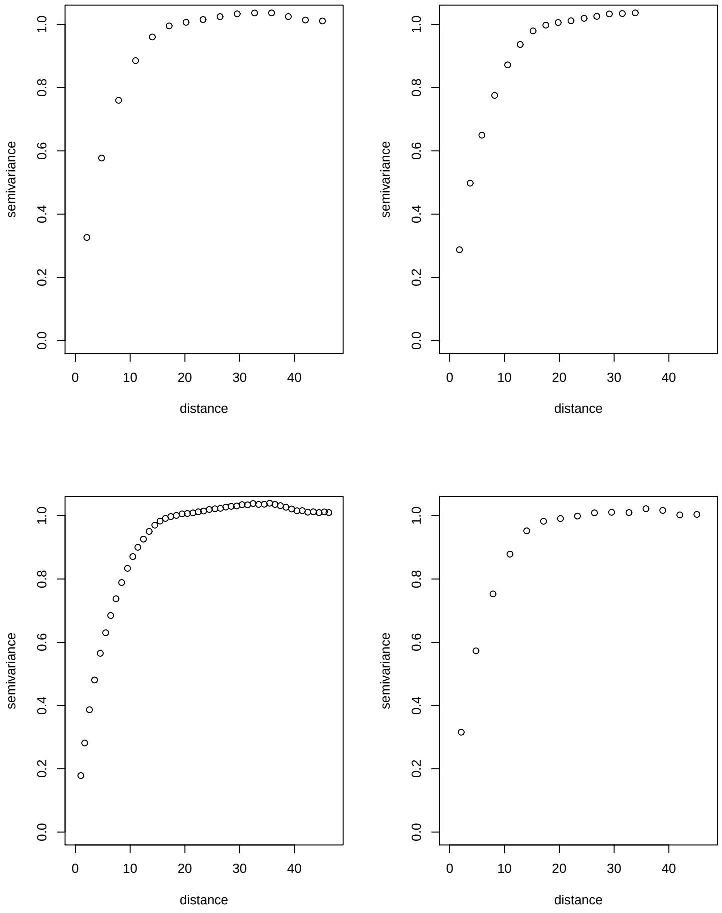
Remarks:

- Generally we construct a plot of these estimates corresponding to each of several directions.
- The polar partition more naturally allows for the construction of directional semivariograms.
- Note that we only have estimates of $\gamma(\mathbf{h})$ for a finite number of lags.
- How many lag classes (i.e. how fine a partition) should we use? A rule of thumb is to require $N(\mathbf{h}_u) \geq 25$ and $\mathbf{h}_u$ to be less than half the maximum lag represented in the dataset. But there is no harm in trying several different partitions.
- An alternative and more robust (less sensitive to outliers) estimator, proposed by Cressie and Hawkins (1980, *Journal of the International Association for Mathematical Geology*), is

$$\begin{aligned} \bar{\gamma}(\mathbf{h}_u) &= \frac{\{\frac{1}{N(H_u)}\sum_{\mathbf{s}_i-\mathbf{s}_j\in H_u}|\hat{\epsilon}(\mathbf{s}_i)-\hat{\epsilon}(\mathbf{s}_j)|^{1/2}\}^4}{.914+[.988/N(H_u)]} \\ & (u=1,\ldots,k). \end{aligned}$$

- Both $\hat{\gamma}(\mathbf{h})$ and $\bar{\gamma}(\mathbf{h})$ can be computed in R using the `variogram` function in the `gstat` package.

Example: Sample semivariogram, assuming isotropy, for a dataset simulated on a 100×100 square grid, from a Gaussian random field with isotropic exponential semivariogram, having `gstat variogram's` range argument equal to 5. Upper left plot uses default arguments; upper right plot uses `cutoff=35`; lower left plot uses `width=1`; and lower right plot is the Cressie-Hawkins robust sample semivariogram.



For the covariance function, the classical estimator is

$$\hat{C}(\mathbf{h}_u) = \frac{1}{N(H_u)} \sum_{\mathbf{s}_i - \mathbf{s}_j \in H_u} (Z(\mathbf{s}_i) - \bar{Z})(Z(\mathbf{s}_j) - \bar{Z}).$$

Remarks:

- This estimator is meaningful only if the process is second-order stationary; otherwise it's estimating something that doesn't exist.
- This estimator is the spatial generalization of the sample autocovariance function used by time series analysts.

Comparison with semivariogram estimation:

- $\hat{\gamma}(\mathbf{h}) \neq \hat{C}(\mathbf{0}) - \hat{C}(\mathbf{h})$, but the difference is usually small for large n.
- If the estimates are based on the observations themselves, then $\hat{C}(\mathbf{h})$ is biased even for regularly-spaced data whereas $\hat{\gamma}(\mathbf{h})$ is unbiased.
- If the estimates are based on residuals from a fitted mean function, then $\hat{\gamma}(\mathbf{h})$ is not as biased as $\hat{C}(\mathbf{h})$.
- If there is trend in the data that is not removed, $\hat{\gamma}(\mathbf{h})$ is not as badly biased as $\hat{C}(\mathbf{h})$.

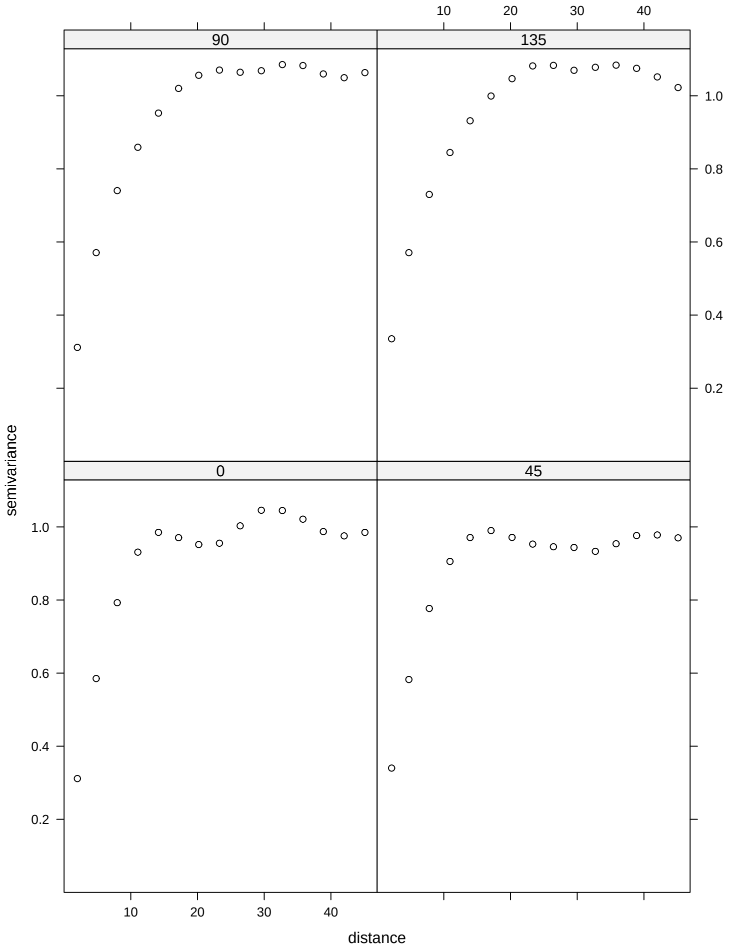
6. Checking for Isotropy

Prior to fitting a parametric model to the sample semivariogram, we may wish to determine how the semivariogram depends on the relative orientation of data locations. That is, we may want to investigate whether isotropy would be a reasonable assumption for the data.

Methods:

1. Visual comparison of directional sample semivariograms
   - At least three (preferably more) directions are needed to distinguish geometric anisotropy from isotropy.
2. Rose diagram
   - Consists of smoothing the directional sample semivariograms, then in the lag space connecting, with a smooth curve, those lag vectors $\mathbf{h}_u$ for which these smoothed semivariograms are roughly equal.
   - In effect, this plots estimated isocorrelation contours (in the case of a second-order stationary process).
   - Circular curves $\Rightarrow$ isotropy; elliptical curves $\Rightarrow$ geometric anisotropy.
3. Formal tests — we won't consider this.

Directional sample semivariograms (in the four directions N-S, NE-SW, E-W, NW-SE) for the same dataset whose isotropic sample semivariogram is displayed on page 65:



These plots were created using the `alpha=c(0,45,90,135)` argument in the `variogram` function of `gstat`. It is possible to use angle classes different from these.

7. Semivariogram Model Fitting

The next step of the geostatistical method is to select and fit a parametric family of models to the sample semivariogram. Why aren't we satisfied with the sample semivariogram itself?

- The sample semivariogram may be quite bumpy. A smoothed version may be helpful for understanding the nature of the spatial dependence.
- The sample semivariogram may violate the required property of conditional nonnegative definiteness.
- For various purposes (e.g. kriging) we may require an estimate of the semivariogram at a lag not represented in the data.

Let $\gamma(\mathbf{h}; \boldsymbol{\theta})$ denote the parametric model to be fit to the sample semivariogram and let Θ denote the parameter space for $\boldsymbol{\theta}$.

Fitting methods:

(a) By eye

(b) Ordinary nonlinear least squares, i.e. minimize

$$RSS(\boldsymbol{\theta}) = \sum_{u \in U} [\hat{\gamma}(\mathbf{h}_u) - \gamma(\mathbf{h}_u; \boldsymbol{\theta})]^2.$$

Here U is a specified subset of lag classes believed to yield reliable estimates of $\gamma(\mathbf{h})$. Generally U is taken to be of the form $U = \{u : N(\mathbf{h}_u) \geq G_1, \|\mathbf{h}_u\| \leq G_2\}$; one rule-of-thumb is to take $G_1 = 30$ and $G_2 =$ half the largest lag in the data. There may be good reasons for using an even smaller maximum lag than this.

(c) Weighted nonlinear least squares (Cressie, 1985), i.e. minimize

$$WRSS(\boldsymbol{\theta}) = \sum_{u \in U} \frac{N(\mathbf{h}_u)}{[\gamma(\mathbf{h}_u; \boldsymbol{\theta})]^2} [\hat{\gamma}(\mathbf{h}_u) - \gamma(\mathbf{h}_u; \boldsymbol{\theta})]^2.$$

The weights, $N(\mathbf{h}_u)/[\gamma(\mathbf{h}_u; \boldsymbol{\theta})]^2$, are small if either $N(\mathbf{h}_u)$ is small or $\gamma(\mathbf{h}_u; \boldsymbol{\theta})$ is large. Thus, nonparametric estimates at large lags tend to receive relatively less weight. The relevant R `gstat` function is `fit.variogram()`.

Actually, the default implementation of `fit.variogram` uses weights $N(\mathbf{h}_u)/\|\gamma(\mathbf{h}_u; \boldsymbol{\theta})\|^2$. If you want to use the weights proposed by Cressie (1985), you must supply `fit.variogram` with the argument `fit.method=2`. Incidentally, the argument `fit.method=6` implements ordinary nonlinear least squares.

(d) Generalized nonlinear least squares

- For two lags $\mathbf{h}_1$ and $\mathbf{h}_2$, $\hat{\gamma}(\mathbf{h}_1)$ and $\hat{\gamma}(\mathbf{h}_2)$ generally are dependent for two reasons:
  - They may be functions of some of the same observations;
  - Even if they have no observations in common, observations in the one estimate are generally spatially correlated with observations in the other.
- Consequently we may want to consider a generalized nonlinear least squares approach, in which we minimize

$$GRSS(\boldsymbol{\theta}) = [\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}(\boldsymbol{\theta})]'[\mathrm{var}(\hat{\boldsymbol{\gamma}})]^{-1}[\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}(\boldsymbol{\theta})].$$

- Here $\hat{\boldsymbol{\gamma}}$ is the vector of nonparametric semivariogram estimates and $\boldsymbol{\gamma}(\boldsymbol{\theta})$ is the corresponding vector of modeled semivariogram values.
- Derivation and calculation of $\mathrm{var}(\hat{\boldsymbol{\gamma}})$ can be a challenge.
- The `fit.variogram.gls` function in `gstat` performs generalized nonlinear least squares fitting to the variogram cloud (not the sample semivariogram).

(d) Maximum likelihood (and restricted maximum likelihood)

- Applicable to processes with second-order stationary errors only.
- Estimates $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ simultaneously.
- Let $\mathbf{V} = \mathbf{V}(\boldsymbol{\theta})$ denote the covariance matrix of $\mathbf{Z} = (Z_1, \ldots, Z_n)'$ and let $\mathbf{X}$ denote the model matrix for the model $\mathbf{Z} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$.
- Assuming normality, the log-likelihood function is (apart from an additive constant which does not depend on $\boldsymbol{\beta}$ or $\boldsymbol{\theta}$)

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}; \mathbf{Z}) = -\frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta}).$$

- A MLE is a value $(\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\theta}}')'$ (where $\boldsymbol{\theta} \in \Theta$) that maximizes $L(\boldsymbol{\beta}, \boldsymbol{\theta})$.
- Generally a MLE must be found by numerical optimization routines (e.g. Newton-Raphson, method of scoring, method of steepest ascent, grid search). Thus we have to be concerned with things such as starting values, respecting constraints on the parameters, and convergence criteria.
- A restricted MLE (REML estimator) is defined as a value $\boldsymbol{\theta} \in \Theta$ that maximizes the log-likelihood function associated with $n-$rank$(\mathbf{X})$ linearly independent error contrasts. It's known to be less biased than MLE's, which is important if rank$(\mathbf{X})$ is appreciable relative to n.
- MLEs and REMLEs can be obtained using the `likfit` function of R's `geoR` package. However, SAS PROC MIXED seems a little faster and more stable, so we will use it for this purpose.

Comparison of Fitting Methods:

The relatively easy-to-compute weighted least squares estimator performs almost as well as the more complicated ML and REML estimators, and has been the estimator of choice for most practitioners. Some statisticians still prefer the less *ad hoc* ML and REML estimators, however.

8. Model Selection Procedures

- Visual inspection of semivariogram plot
- Minimized weighted (or generalized) residual sum of squares function, $WRSS(\hat{\boldsymbol{\theta}})$ (or $GRSS(\hat{\boldsymbol{\theta}})$).
- Maximized log-likelihood (or restricted log-likelihood) function, $L(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$.
- Penalized likelihood criteria, e.g. Akaike's Information Criterion

$$AIC = L(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}) - \#\text{ of estimated parameters.}$$

Results for same data used on previous page: $WRSS(\hat{\boldsymbol{\theta}})$ for exponential = 72.10, $WRSS(\hat{\boldsymbol{\theta}})$ for spherical = 60.02!

Real Example: Fitted Semivariogram Model for Sulfate Deposition Data

- First, a fifth-order polynomial surface was fit to the data by ordinary least squares, and the residuals from this fit were used for further analysis.
- Based on a preliminary assessment of isotropy, the scale in the E-W direction was halved and isotropic models were fit by WLS.
- A visual examination of the sample semivariogram suggested the use of a spherical or exponential model with a nugget effect.
- Both models were fit by WLS; $WRSS(\hat{\boldsymbol{\theta}})$ was 29% smaller for the spherical model.
- The fitted spherical semivariogram was

$$\gamma(\|\mathbf{h}\|) = \begin{cases} 0, & \text{if } \|\mathbf{h}\| = 0 \\ .0802 + .1263\left\{\frac{3}{2}\left(\frac{\|\mathbf{h}\|}{508.0}\right) - \frac{1}{2}\left(\frac{\|\mathbf{h}\|}{508.0}\right)^3\right\}, & \text{if } 0 < \|\mathbf{h}\| \leq 508.0 \\ .2065, & \text{if } \|\mathbf{h}\| > 508.0 \end{cases}$$

9. Spatial Regression

Recall our general geostatistical model:

$$Z(\mathbf{s}) = m(\mathbf{s};\boldsymbol{\beta}) + \epsilon(\mathbf{s}).$$

Trend surface, median polish, and nonparametric regression models all account for spatial location only in the large-scale structure; they ignore small-scale structure through the assumption that the model residuals are uncorrelated (and homoscedastic). Now we consider models for which the residuals are spatially correlated.

For convenience assume that $m(\mathbf{s};\boldsymbol{\beta})$ is linear in the elements of $\boldsymbol{\beta}$, and assume that $\{\epsilon(\mathbf{s}) : \mathbf{s} \in D\}$ is second-order stationary.

Let $\mathbf{V} = \mathbf{V}(\boldsymbol{\theta})$ be the $n \times n$ matrix whose (i, j)th element is $C(\mathbf{s}_i - \mathbf{s}_j;\boldsymbol{\theta})$. Then the model for all the observations can be written as

$$\mathbf{Z} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad E(\boldsymbol{\epsilon}) = \mathbf{0}, \qquad \mathrm{var}(\boldsymbol{\epsilon}) = \mathbf{V} = \sigma^2\mathbf{R}$$

where $\mathbf{V} = \mathbf{V}(\boldsymbol{\theta})$ is a positive definite (hence invertible) covariance matrix, σ^2 is the constant variance, and $\mathbf{R}$ is a correlation matrix.

Toy example:

Assume constant mean β and isotropic spherical covariance function with range 4, sill 5, and nugget 0:

$$C(r;\boldsymbol{\theta}) = \begin{cases} 5\left(1 - \frac{3r}{8} + \frac{r^3}{128}\right) & \text{for } 0 < r \leq 4 \\ 0 & \text{for } r > 4 \end{cases}$$

Then $\mathbf{X}$ and $\mathbf{V}$ are

$$\mathbf{X} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \qquad \mathbf{V} = 5\begin{pmatrix} 1.000 & .633 & .492 & .249 & 0 \\ & 1.000 & .633 & .492 & 0 \\ & & 1.000 & .633 & .061 \\ & \text{symm} & & 1.000 & .014 \\ & & & & 1.000 \end{pmatrix}$$

Finally suppose that the vector of responses is $\mathbf{Z} = (1, 0, 2, 1, 6)'$.

(a) Generalized least squares (GLS) estimation

- GLSE of $\boldsymbol{\beta}$ is the value of $\boldsymbol{\beta}$ that minimizes the generalized residual sum of squares criterion,
$$GRSS(\boldsymbol{\beta}) = (\mathbf{Z} - \mathbf{X}\boldsymbol{\beta})'\mathbf{R}^{-1}(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta}).$$
- Equivalently, $\hat{\boldsymbol{\beta}}_{GLS} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Z} = (\mathbf{X}'\mathbf{R}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}$.
- $\text{var}(\hat{\boldsymbol{\beta}}_{GLS}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{R}^{-1}\mathbf{X})^{-1}$.
- If the distribution of $\boldsymbol{\epsilon}$ is multivariate normal, then the GLSE is also the best (minimum variance) unbiased estimator and the maximum likelihood estimator (MLE).
- Fitted regression equation is
$$\hat{Z} = x_1\hat{\beta}_{1,GLS} + x_2\hat{\beta}_{2,GLS} + \cdots + x_p\hat{\beta}_{p,GLS}$$
- Standard estimator of σ^2 is $\hat{\sigma}^2_{GLS} = GRSS(\hat{\boldsymbol{\beta}}_{GLS})/(n-p)$.
- To test H_0: $\mathbf{C}\boldsymbol{\beta} = \mathbf{d}$ vs. H_A: $\mathbf{C}\boldsymbol{\beta} \neq \mathbf{d}$ at α level of significance, compare
$$\frac{(\mathbf{C}\hat{\boldsymbol{\beta}}_{GLS} - \mathbf{d})'[\mathbf{C}(\mathbf{X}'\mathbf{R}^{-1}\mathbf{X})^{-1}\mathbf{C}']^{-1}(\mathbf{C}\hat{\boldsymbol{\beta}}_{GLS} - \mathbf{d})}{\hat{\sigma}^2_{GLS} \cdot (\#\text{ rows of }\mathbf{C})}$$
to upper α percentage point of $F(\#\text{ rows of }\mathbf{C}, n-p)$ distribution.

GLS estimation for toy example:

	GLS	*OLS*
$\hat{\boldsymbol{\beta}}$	2.83	2.00
$\hat{\sigma}^2$	5.15	5.50
var($\hat{\boldsymbol{\beta}}$)	1.92	2.28

(variances calculated under correct model, i.e. model with correlation)

Note:

- Accounting for the substantial correlation among observations 1, 2, 3, and 4 affects the estimation of $\boldsymbol{\beta}$.
- $\text{var}(\hat{\boldsymbol{\beta}}_{GLS}) < \text{var}(\hat{\boldsymbol{\beta}}_{OLS})$ (both calculated under correct model)

(b) Estimated generalized least squares (EGLS) estimation

In practice we don't know the true value of $\boldsymbol{\theta}$ and consequently $\mathbf{V}$ cannot be completely specified. A natural solution to this problem is to to replace $\boldsymbol{\theta}$ in the evaluation of $\mathbf{V}$ by an estimator $\hat{\boldsymbol{\theta}}$, obtaining $\hat{\mathbf{V}} \equiv \mathbf{V}(\hat{\boldsymbol{\theta}})$, and then use all the formulas associated with GLS estimation (as if $\hat{\boldsymbol{\theta}}$ was the true value).

Thus, the EGLS estimator of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}}_{EGLS} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{Z}.$$

We can *approximately* quantify the uncertainty of $\hat{\boldsymbol{\beta}}_{EGLS}$ via

$$\widehat{\mathrm{var}}(\hat{\boldsymbol{\beta}}_{EGLS}) = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}.$$

Facts:

- $\hat{\boldsymbol{\beta}}_{EGLS}$ is unbiased for $\boldsymbol{\beta}$ under rather unrestrictive conditions
- A closed-form expression for $\mathrm{var}(\hat{\boldsymbol{\beta}}_{EGLS})$ is not known
- $\widehat{\mathrm{var}}(\hat{\boldsymbol{\beta}}_{EGLS})$ tends to underestimate $\mathrm{var}(\hat{\boldsymbol{\beta}}_{EGLS})$

EGLS estimation for toy example: Suppose the estimated range is 3.0, estimated sill is 4.0, and estimated nugget is 0. Then

$$\hat{\mathbf{V}} = 4 \begin{pmatrix} 1.000 & .519 & .345 & .089 & 0 \\ & 1.000 & .519 & .345 & 0 \\ & & 1.000 & .519 & 0 \\ & \text{symm} & & 1.000 & 0 \\ & & & & 1.000 \end{pmatrix}.$$

(Compare to true $\mathbf{V}$ on page 71.) Note that the overall variation and spatial correlation are both estimated as weaker here than they really are.

Then,

$$\hat{\boldsymbol{\beta}}_{EGLS} = 2.70.$$

(c) Maximum likelihood estimation

Assume multivariate normality of the observations. Then, the log-likelihood function is

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}) = L(\boldsymbol{\beta}, \sigma^2, \mathbf{R}(\boldsymbol{\theta})) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2}\log|\mathbf{R}| - \frac{1}{2\sigma^2}(\mathbf{Z}-\mathbf{X}\boldsymbol{\beta})'\mathbf{R}^{-1}(\mathbf{Z}-\mathbf{X}\boldsymbol{\beta}).$$

The maximum likelihood estimators (MLEs) of $\boldsymbol{\beta}$, σ^2, and the parameters in $\mathbf{R}$ are given by

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{MLE} &= (\mathbf{X}'\hat{\mathbf{R}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{R}}^{-1}\mathbf{Z}, \\ \hat{\sigma}^2_{MLE} &= \frac{1}{n}(\mathbf{Z}-\mathbf{X}\hat{\boldsymbol{\beta}})'\hat{\mathbf{R}}^{-1}(\mathbf{Z}-\mathbf{X}\hat{\boldsymbol{\beta}}),\end{aligned}$$

and $\hat{\mathbf{R}}$, where the latter maximizes the “profile log-likelihood” function

$$L^*(\mathbf{R}) = -\frac{n}{2}\log(\hat{\sigma}^2_{MLE}) - \frac{1}{2}\log|\mathbf{R}|.$$

Remarks:

- $\hat{\boldsymbol{\beta}}_{MLE}$ is merely the EGLS estimator of $\boldsymbol{\beta}$ using MLEs of the variance and correlation parameters.
- $\hat{\sigma}^2_{MLE}$ is merely the residual sum of squares for EGLS, divided by the sample size.
- In general there is not an explicit formula for $\hat{\mathbf{R}}$. Grid search or other numerical optimization methods (e.g. Nelder-Mead simplex or Newton-Raphson) must be used to obtain it. For the latter, starting values and a convergence criterion must be specified.
- Care must be taken to ensure that $\hat{\mathbf{R}}$ remains in the allowable parameter space.
- Multiple modes of $L^*(\mathbf{R})$ are possible.
- Standard errors for parameter estimates have been obtained under an assumption of “increasing-domain asymptotics.” Confidence intervals for parameters can be constructed using these.
- $\hat{\mathbf{V}} = \hat{\sigma}^2_{MLE} \cdot \hat{\mathbf{R}}$.

(d) Hypothesis Testing and Model Comparisons

Can test

$$H_0 : \mathbf{A}\boldsymbol{\beta} = \mathbf{a} \quad \text{versus} \quad H_A : \mathbf{A}\boldsymbol{\beta} \neq \mathbf{a}$$

at the α level of significance (approximately) by comparing

$$\frac{(\mathbf{A}\hat{\boldsymbol{\beta}}_{EGLS} - \mathbf{a})'[\mathbf{A}(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{A}']^{-1}(\mathbf{A}\hat{\boldsymbol{\beta}}_{EGLS} - \mathbf{a})}{\mathbf{Z}'[\hat{\mathbf{V}}^{-1} - \hat{\mathbf{V}}^{-1}\mathbf{X}(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}]\mathbf{Z}}$$

to $F(\alpha, \#\text{ rows of }\mathbf{A}, n-p)$.

To compare nested models, the above test can be used. To compare non-nested models, AIC (defined previously) can be used.

10. Mean Structure or Covariance Structure?

One issue we've ignored to this point is that the choice, in practice, of a decomposition of the data into mean structure (large-scale variation) and covariance structure (small-scale variation) is not so clearcut. This is often phrased as follows:

> "One man's mean structure is another man's covariance structure."

An illustration: Five realizations from each of four one-dimensional random processes observed at the locations $s = 1, 2, \ldots, 50$.

(a) *No trend, no correlation:* $Z(s) = 0 + \epsilon(s)$, where $\{\epsilon(s)\colon\ s = 1, \ldots, 50\}$ are iid N(0,1) random variables.

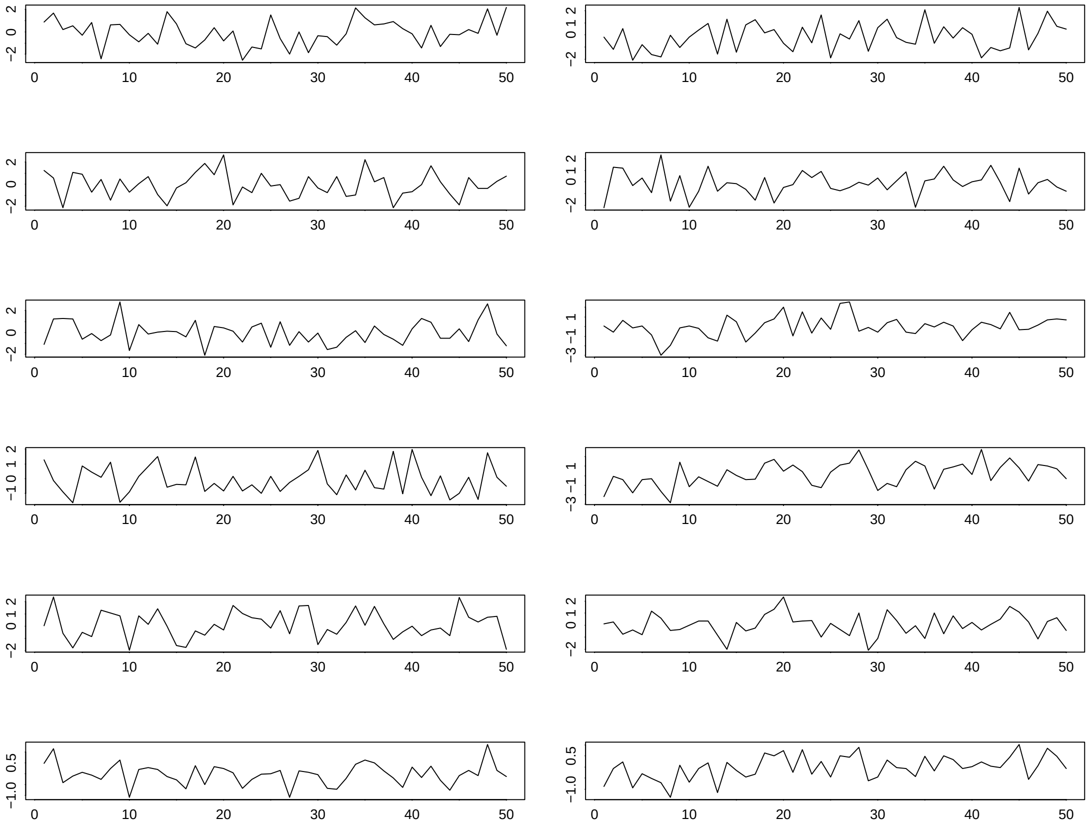
(b) *Trend but no correlation.* $Z(s) = -0.5 + 0.02s + \epsilon(s)$, where $\{\epsilon(s)\colon\ s = 1, \ldots, 50\}$ are iid N(0,1) random variables.

(c) *Correlation but no trend.* $Z(s) = 0 + \epsilon(s)$, where $\{\epsilon(s)\colon\ s = 1, \ldots, 50\}$ are normally distributed random variables with mean 0 and covariance structure determined by the exponential covariance function $C(r) = \exp(-r/5)$.

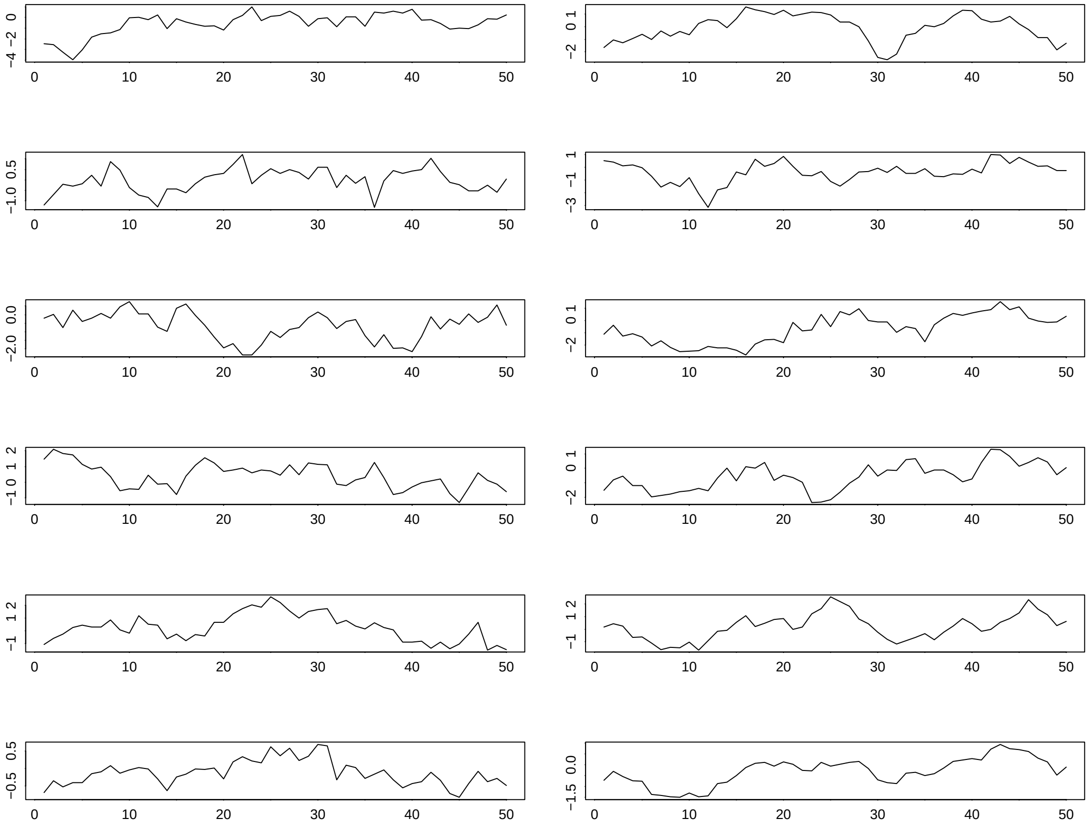
(d) *Trend and correlation.* $Z(s) = -0.5 + 0.02s + \epsilon(s)$, where $\{\epsilon(s)\colon\ s = 1, \ldots, 50\}$ are normally distributed random variables with mean 0 and covariance structure determined by the exponential covariance function $C(r) = \exp(-r/5)$.

If replications of a spatial process are available, statistical procedures exist for distinguishing between these models. In practice, however, geostatistical data are not replicated so we must settle for plausibility, rather than a high degree of certainty, of the proposed decomposition.

Realizations of (a) (left column) and (b) (right column). Last panel in each column plots the averages of the five plots above it.



Realizations of (c) (left column) and (d) (right column). Last panel in each column plots the averages of the five plots above it.



11. Spatial Prediction (Kriging)

(a) *Ordinary Kriging Problem Formulation*

Definition (Krige, 1978):

> "The name given ... to the multiple regression procedure for arriving at the best linear unbiased predictor or best linear weighted moving average predictor of the ore grade of an ore block (of any size) by assigning an optimum set of weights to all the available and relevant data inside and outside the ore block."

Krige's original method is what is now called *ordinary kriging* (OK). There have been several modifications and extensions (e.g., universal kriging, indicator kriging, disjunctive kriging, and others) but they are all based on quite similar ideas.

The theory of OK is based on the same geostatistical model we have been using all along, with two important restrictions:

1. The mean $m(\mathbf{s})$ is assumed to be constant.
2. The semivariogram $\gamma(\mathbf{h})$ is assumed to be known.

Let $\mathbf{s}_0$ denote an arbitrary location in D. Usually this will be an unsampled location but it need not be.

Goal of kriging: to predict the value of $Z(\mathbf{s}_0)$ at $\mathbf{s}_0$, using the observed responses $Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n)$.

Properties of OK predictor:

1. It is a linear combination of the data values, i.e.,

$$\hat{Z}(\mathbf{s}_0) = \sum_{i=1}^{n} \lambda_i Z(\mathbf{s}_i).$$

2. It is unbiased, i.e., it satisfies

$$E[\hat{Z}(\mathbf{s}_0)] = E[Z(\mathbf{s}_0)].$$

3. Among all functions of the data that satisfy the first 2 properties, it minimizes the variance of prediction error, $\mathrm{var}[\hat{Z}(\mathbf{s}_0) - Z(\mathbf{s}_0)]$.

The properties we have imposed on our predictor lead us to minimize

$$\text{var}[\sum_{i=1}^{n} \lambda_i Z(\mathbf{s}_i) - Z(\mathbf{s}_0)]$$

subject to the restriction

$$\sum_{i=1}^{n} \lambda_i = 1.$$

We will present OK in terms of the semivariogram; an equivalent presentation can be given in terms of the covariance function.

(b) *Ordinary Kriging Problem Solution*

The method of Lagrange multipliers, from calculus, can be used to solve the constrained minimization problem formulated previously. It can be shown that the optimal coefficients $\lambda_1, \ldots, \lambda_n$ of the OK predictor are the first n elements of the vector $\boldsymbol{\lambda}_O$ that satisfies the following system of linear equations, known as the (ordinary) *kriging equations*:

$$\boldsymbol{\Gamma}_O \boldsymbol{\lambda}_O = \boldsymbol{\gamma}_O$$

where

$$\boldsymbol{\lambda}_O = (\lambda_1, \ldots, \lambda_n, m)'$$

$$\boldsymbol{\gamma}_O = [\gamma(\mathbf{s}_1 - \mathbf{s}_0), \ldots, \gamma(\mathbf{s}_n - \mathbf{s}_0), 1]'$$

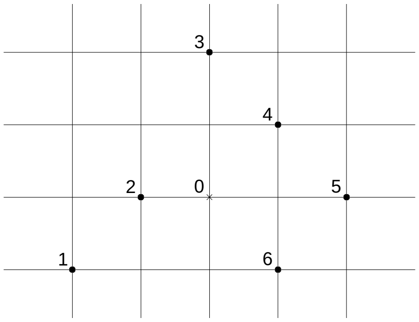
$$\boldsymbol{\Gamma}_O = \begin{cases} \gamma(\mathbf{s}_i - \mathbf{s}_j) & \text{for } i = 1, \ldots, n;\ j = 1, \ldots, n \\ 1 & \text{for } i = n+1;\ j = 1, \ldots, n \\ 0 & \text{for } i = n+1;\ j = n+1 \end{cases}$$

and m is a Lagrange multiplier and $\boldsymbol{\Gamma}_O$ is symmetric.

The minimized variance, called the *kriging variance*, is

$$\sigma^2_{OK}(\mathbf{s}_0) = \sum_{i=1}^{n} \lambda_i \gamma(\mathbf{s}_i - \mathbf{s}_0) + m = \boldsymbol{\lambda}_O' \boldsymbol{\gamma}_O.$$

Toy example:



Take $\gamma(\|\mathbf{h}\|) = 1 - \exp(-\|\mathbf{h}\|/2)$.

$$\boldsymbol{\gamma}_O = \begin{bmatrix} 1 - \exp(-\sqrt{5}/2) \\ 1 - \exp(-1/2) \\ 1 - \exp(-1) \\ 1 - \exp(-\sqrt{2}/2) \\ 1 - \exp(-1) \\ 1 - \exp(-\sqrt{2}/2) \\ 1 \end{bmatrix}$$

$$\boldsymbol{\Gamma}_O = \begin{bmatrix} 0 & 1 - \exp(-\sqrt{2}/2) & 1 - \exp(-\sqrt{13}/2) & \cdots & & 1 \\ & 0 & 1 - \exp(-\sqrt{5}/2) & \cdots & & 1 \\ & & 0 & & & 1 \\ & & & \ddots & & \vdots \\ & \text{symm} & & & 0 & 1 \\ & & & & & 0 \end{bmatrix}$$

$$\boldsymbol{\lambda}_O = \boldsymbol{\Gamma}_O^{-1}\boldsymbol{\gamma}_O = [.017, .422, .065, .218, .031, .246, .004]$$

$$\sigma^2_{OK}(\mathbf{s}_0) = \boldsymbol{\lambda}_O'\boldsymbol{\gamma}_O = .478$$

(c) *Influence of Spatial Dependence on Ordinary Kriging*

Consider the same spatial configuration as in the previous example, but with each of the following four semivariograms:

1. $\gamma(\|\mathbf{h}\|) = 1 - \exp(-\|\mathbf{h}\|/2)$ (same model as in the example)
2. $\gamma(\|\mathbf{h}\|) = 1 - \exp(-\|\mathbf{h}\|/4)$ (stronger spatial correlation)
3. $\gamma(\|\mathbf{h}\|) = 0.25 + 0.75(1 - \exp(-\|\mathbf{h}\|/2))$ (nonzero nugget, same sill)
4. $\gamma(\|\mathbf{h}\|) = 1 - \exp(-\|\mathbf{h}\|/2)$ in E–W direction, $\gamma(\|\mathbf{h}\|) = 1 - \exp(-\|\mathbf{h}\|)$ in N–S direction, geometrically anisotropic with major axis in E–W orientation

Kriging weights and variances for these models:

	Model 1	Model 2	Model 3	Model 4
λ_1	.017	.001	.083	.033
λ_2	.422	.450	.323	.504
λ_3	.065	.049	.106	.015
λ_4	.218	.230	.189	.136
λ_5	.031	.010	.083	.177
λ_6	.246	.261	.215	.135
$\sigma^2_{OK}(\mathbf{s}_0)$	.478	.254	.669	.546

(d) *Explicit Matrix Expressions for OK Predictor and OK Variance*

We can use some results from linear algebra to obtain alternative expressions for the ordinary kriging predictor and the kriging variance which do not involve the unknown Lagrange multiplier. Define

$$\begin{aligned}\boldsymbol{\lambda} &= (\lambda_1, \ldots, \lambda_n)',\\ \boldsymbol{\gamma} &= (\gamma(\mathbf{s}_1 - \mathbf{s}_0), \ldots, \gamma(\mathbf{s}_n - \mathbf{s}_0))',\\ \boldsymbol{\Gamma} &= \{\gamma(\mathbf{s}_i - \mathbf{s}_j)\}.\end{aligned}$$

It can be shown that

$$\begin{aligned}m &= -(1 - \mathbf{1}'\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma})/(\mathbf{1}'\boldsymbol{\Gamma}^{-1}\mathbf{1}),\\ \boldsymbol{\lambda} &= \boldsymbol{\Gamma}^{-1}\{\boldsymbol{\gamma} + [(1 - \mathbf{1}'\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma})/(\mathbf{1}'\boldsymbol{\Gamma}^{-1}\mathbf{1})]\mathbf{1}\}.\end{aligned}$$

Therefore the OK predictor can be expressed as follows:

$$\hat{Z}(\mathbf{s}_0) = \{\boldsymbol{\gamma} + [(1 - \mathbf{1}'\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma})/(\mathbf{1}'\boldsymbol{\Gamma}^{-1}\mathbf{1})]\mathbf{1}\}'\boldsymbol{\Gamma}^{-1}\mathbf{Z}.$$

The kriging variance can be expressed as follows:

$$\sigma^2_{OK}(\mathbf{s}_0) = \boldsymbol{\gamma}'\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma} - (\mathbf{1}'\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma} - 1)^2/(\mathbf{1}'\boldsymbol{\Gamma}^{-1}\mathbf{1}).$$

Also, it can be shown that if the process is second-order stationary, then we can express the OK predictor and its kriging variance alternatively as

$$\hat{Z}(\mathbf{s}_0) = \{\mathbf{v}_0 + [(1 - \mathbf{1}'\mathbf{V}^{-1}\mathbf{v}_0)/(\mathbf{1}'\mathbf{V}^{-1}\mathbf{1})]\mathbf{1}\}'\mathbf{V}^{-1}\mathbf{Z}$$

and

$$\sigma^2_{OK}(\mathbf{s}_0) = C(\mathbf{s}_0, \mathbf{s}_0) - \mathbf{v}_0'\mathbf{V}^{-1}\mathbf{v}_0 + (\mathbf{1}'\mathbf{V}^{-1}\mathbf{v}_0 - 1)^2/(\mathbf{1}'\mathbf{V}^{-1}\mathbf{1})$$

where

$$\begin{aligned}\mathbf{v}_0 &= (C(\mathbf{s}_1 - \mathbf{s}_0), \ldots, C(\mathbf{s}_n - \mathbf{s}_0))',\\ \mathbf{V} &= \{C(\mathbf{s}_i - \mathbf{s}_j)\}.\end{aligned}$$

(e) *Some Miscellaneous Remarks*

- We would generally want to characterize the uncertainty in our predictor. This can be achieved with a confidence interval, or more properly, a *prediction interval.* If we assume that the random field is Gaussian, then all the $Z(\mathbf{s}_i)$'s are normally distributed and so then is $\hat{Z}(\mathbf{s}_0)$. Therefore, a $100(1-\alpha)\%$ prediction interval for $Z(\mathbf{s}_0)$ is as follows:

  $$\hat{Z}(\mathbf{s}_0) \pm z_{\alpha/2}\sigma_{OK}(\mathbf{s}_0),$$

  where $z_{\alpha/2}$ is the upper $\alpha/2$ percentage point of the standard normal distribution.

- As we have described it here, the OK predictor is a linear combination of *all* the observations. In practice, often only the observations within a "moving window" or "kriging neighborhood" are used.

  The nugget/sill ratio, the range, and the sampling configuration are important factors in the choice of window size.

- Ordinary kriging is derived under an assumption of intrinsic stationarity, which assumes the mean is constant. The variation of kriging that can handle trends is called *universal kriging* (to be described shortly).

- Ordinary kriging also is derived under an assumption that the semivariogram is known. In practice, the semivariogram is unknown and must be estimated, and the estimator $\hat{\gamma}(\cdot)$ replaces $\gamma(\cdot)$ in the kriging equations and in the expression for the kriging variance.

  $$\begin{aligned}\hat{\hat{Z}}(\mathbf{s}_0) &= \{\hat{\boldsymbol{\gamma}} + [(1-\mathbf{1}'\hat{\boldsymbol{\Gamma}}^{-1}\hat{\boldsymbol{\gamma}})/(\mathbf{1}'\hat{\boldsymbol{\Gamma}}^{-1}\mathbf{1})]\mathbf{1}\}'\hat{\boldsymbol{\Gamma}}^{-1}\mathbf{Z} \\ \hat{\sigma}^2_{OK}(\mathbf{s}_0) &= \hat{\boldsymbol{\gamma}}'\hat{\boldsymbol{\Gamma}}^{-1}\hat{\boldsymbol{\gamma}} - (\mathbf{1}'\hat{\boldsymbol{\Gamma}}^{-1}\hat{\boldsymbol{\gamma}}-1)^2/(\mathbf{1}'\hat{\boldsymbol{\Gamma}}^{-1}\mathbf{1})\end{aligned}$$

  Note: The estimated kriging variance tends to underestimate the prediction error variance of the estimated OK predictor because it does not account for the estimation error incurred in estimating $\boldsymbol{\theta}$.

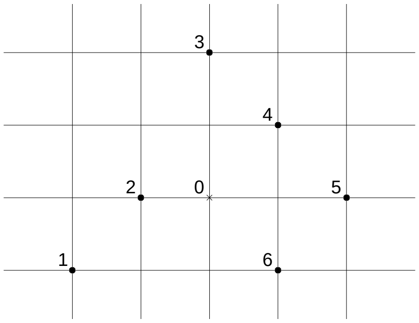
(f) *Sampling Design*

Note that the kriging variance at any given site $\mathbf{s}_0$,

$$\sigma^2_{OK}(\mathbf{s}_0) = \sum_{i=1}^{n} \lambda_i \gamma(\mathbf{s}_i - \mathbf{s}_0) + m = \boldsymbol{\lambda}'_O \boldsymbol{\gamma}_O = \boldsymbol{\gamma}'_O \boldsymbol{\Gamma}_O^{-1} \boldsymbol{\gamma}_O$$

does not depend on the observed responses. Thus, it can be used to address sampling design questions, such as where to take one more observation to maximize the reduction in σ^2_{OK} at a certain point, or where to take one more observation to minimize the maximum (or average) value of σ^2_{OK} over the entire spatial domain. This is potentially very useful for environmental monitoring programs.

Example 1: Consider the same sampling configuration as in the ordinary kriging toy example presented a few pages back: Suppose we wish to minimize the kriging variance at $\mathbf{s}_0$, and we have sufficient resources to take an observation at any one of the four remaining unsampled sites (excluding $\mathbf{s}_0$).



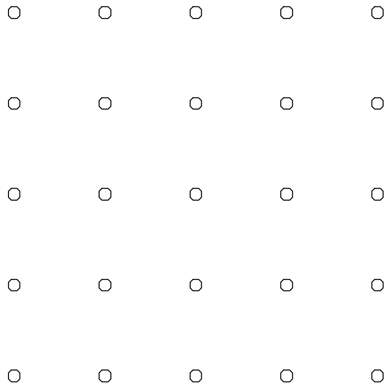
The kriging variances at $\mathbf{s}_0$ corresponding to the addition of each of the sites a, b, c, and d are as follows:

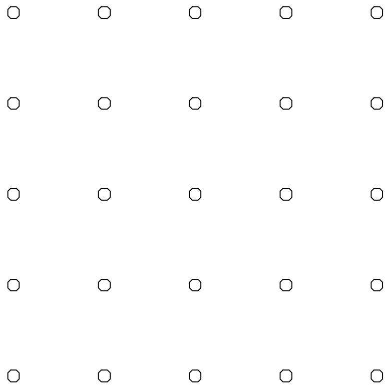
Additional site	$\sigma^2_{OK}(\mathbf{s}_0)$
a	.4687
b	.4366
c	.4368
d	.4347

Thus, the best additional site is d.

Example 2:

- Consider a situation in which there are 25 potential data locations, which are arrayed in a 5×5 square grid $\mathcal{S}$ with unit spacing.
- Suppose that potential observations at these locations are assumed to arise from a stationary process with unknown (but constant) mean and an isotropic exponential covariance function given by $C(\mathbf{s}, \mathbf{u}) = \theta_1 \exp(-\|\mathbf{s} - \mathbf{u}\|/\theta_2)$.
- Put $\rho = \exp(-1/\theta_2)$, and reparameterize the covariance function as $C(\mathbf{s}, \mathbf{u}) = \theta_1 \rho^{\|\mathbf{s}-\mathbf{u}\|}$.
- Suppose we have enough resources to take a measurement at only 4 of the potential data locations.
- Suppose we want to choose the 4 data locations to minimize the maximum kriging variance over the 25 grid locations, i.e. $\max_{\mathbf{s} \in \mathcal{S}} \sigma^2_{OK}(\mathbf{s})$.
- For each value of $\rho = 0.1, 0.2, \ldots, 0.9$, the optimal 4-point design is as follows:





(g) *Extensions of Ordinary Kriging*

- *Block Kriging.* We have considered *point kriging*, i.e., prediction at a single point. Sometimes it is desirable to predict the average value over an entire region, i.e.,

  $$Z(B) \equiv \frac{\int_B Z(\mathbf{s})\,d\mathbf{s}}{|B|},$$

  where $|B|$ is the area of the region (block) for which a predicted value is desired. For example, mining engineers are interested in this because the economics of mining require the extraction of material in relatively large units. Block kriging is a straightforward extension of OK that accomplishes this.

  The theoretical development of block kriging proceeds along similar lines as for point kriging and yields ordinary block kriging equations,

  $$\mathbf{\Gamma}_O \boldsymbol{\lambda}_{OB} = \boldsymbol{\gamma}_{OB}$$

  where

  $$\begin{aligned} \boldsymbol{\gamma}_{OB} &= [\gamma(B, \mathbf{s}_1), \ldots, \gamma(B, \mathbf{s}_n), 1]' \text{ and} \\ \gamma(B, \mathbf{s}_i) &= |B|^{-1} \int_B \gamma(\mathbf{u} - \mathbf{s}_i)\,d\mathbf{u} \end{aligned}$$

  The ordinary block kriging predictor of $Z(B)$ is given by

  $$\hat{Z}(B) = \sum_{i=1}^{n} \lambda_{B,i} Z(\mathbf{s}_i)$$

  where $\lambda_{B,1}, \ldots, \lambda_{B,n}$ are the first n elements of $\boldsymbol{\lambda}_{OB}$.

  The kriging variance is given by $\boldsymbol{\lambda}_{OB}' \boldsymbol{\gamma}_{OB} - |B|^{-2} \int_B \int_B \gamma(\mathbf{u} - \mathbf{v})\,d\mathbf{u}\,d\mathbf{v}$.

- *Universal Kriging* (UK). Suppose that

  $$Z(\mathbf{s}) = \beta_0 + \beta_1 f_1(\mathbf{s}) + \beta_2 f_2(\mathbf{s}) + \cdots + \beta_p f_p(\mathbf{s}) + \epsilon(\mathbf{s})$$

  where the $f_j(\cdot)$'s are functions of spatial location and $\epsilon(\cdot)$ is intrinsically stationary.

  Again, we wish to predict $Z(\mathbf{s}_0)$ by a linear unbiased predictor, and to do so in such a way that the variance of prediction error is minimized. That is, we seek to minimize

  $$\text{var}[\sum_{i=1}^{n} \lambda_i Z(\mathbf{s}_i) - Z(\mathbf{s}_0)]$$

  subject to the unbiasedness constraint

  $$E[\sum_{i=1}^{n} \lambda_i Z(\mathbf{s}_i)] = \beta_0 + \beta_1 f_1(\mathbf{s}_0) + \cdots + \beta_p f_p(\mathbf{s}_0) \quad \text{for all } \boldsymbol{\beta}.$$

This constraint is actually a set of $p+1$ constraints:

$$\begin{aligned}
\sum_{i=1}^{n} \lambda_i &= 1 \\
\sum_{i=1}^{n} \lambda_i f_1(\mathbf{s}_i) &= f_1(\mathbf{s}_0) \\
&\vdots \\
\sum_{i=1}^{n} \lambda_i f_p(\mathbf{s}_i) &= f_p(\mathbf{s}_0)
\end{aligned}$$

Thus, in this minimization problem there are $p+1$ Lagrange multipliers. The algebra is "messier" than with ordinary kriging but the end result is similar: the coefficients in the UK predictor are the first n elements of the vector $\boldsymbol{\lambda}_U$ that satisfies the UK equations

$$\boldsymbol{\Gamma}_U \boldsymbol{\lambda}_U = \boldsymbol{\gamma}_U.$$

Here,

$$\begin{aligned}
\boldsymbol{\lambda}_U &= (\lambda_1, \ldots, \lambda_n, m_0, m_1, \ldots, m_p)', \\
\boldsymbol{\gamma}_U &= [\gamma(\mathbf{s}_0 - \mathbf{s}_1), \ldots, \gamma(\mathbf{s}_0 - \mathbf{s}_n), 1, f_1(\mathbf{s}_0), \ldots, f_p(\mathbf{s}_0)]'
\end{aligned}$$

and $\boldsymbol{\Gamma}_U$ is a symmetric $(n+p+1) \times (n+p+1)$ matrix

$$\boldsymbol{\Gamma}_U = \begin{cases} \gamma(\mathbf{s}_i - \mathbf{s}_j) & i = 1, \ldots, n; j = 1, \ldots, n, \\ f_{j-1-n}(\mathbf{s}_i) & i = 1, \ldots, n; j = n+1, \ldots, n+p+1, \\ 0 & i = n+1, \ldots, n+p+1; j = n+1, \ldots, n+p+1. \end{cases}$$

The univeral kriging variance is

$$\sigma^2_{UK}(\mathbf{s}_0) = \boldsymbol{\lambda}_U' \boldsymbol{\gamma}_U$$

- *Indicator Kriging.* Ordinary and universal kriging both yield a predictor which is a "location estimator" of the distribution of $Z(\mathbf{s}_0)|\mathbf{Z}$; for example, the ordinary and universal kriging predictors are, under the appropriate models, estimates of $E(Z(\mathbf{s}_0)|\mathbf{Z})$. In some situations, however, the quantity $P(Z(\mathbf{s}_0) \geq z_0|\mathbf{Z})$ is of more importance than $E(Z(\mathbf{s}_0)|\mathbf{Z})$. This is often true in environmental monitoring when there are prespecified standards, such as "ozone levels in air cannot exceed 2 ppm." Indicator kriging is a method for predicting such a quantity from data at nearby locations.

- *Co-Kriging.* In some situations, measurements at data locations are taken on more than one variable. Co-Kriging simultaneously predicts all of these variables at unsampled locations, utilizing data on all variables and exploiting dependence between variables as well as spatial dependence within variables.

(h) *Kriging versus Inverse Distance Weighting (IDW)*

Kriging (in all its forms) is merely one method for spatial interpolation. Another method, highly favored by geo-scientists, is inverse distance weighting. In this approach, the predicted (or estimated) value at any location where no observation was taken is a weighted linear combination of the observed data, with weights inversely related to the distance from the observed data location to the place at which we wish to predict.

Two examples, using our notation and letting d_{0i} be the distance between $\mathbf{s}_i$ and $\mathbf{s}_0$:

1. $Z_{IDW}(\mathbf{s}_0) = \frac{\sum_i d_{0i}^{-1} Z(\mathbf{s}_i)}{\sum_i d_{0i}^{-1}}$
2. $Z_{IDSW}(\mathbf{s}_0) = \frac{\sum_i d_{0i}^{-2} Z(\mathbf{s}_i)}{\sum_i d_{0i}^{-2}}$

Note:

- Assuming that the random field is intrinsically stationary, IDW yields unbiased predictors.
- Since IDW is linear and unbiased under intrinsic stationarity, and OK yields the best linear unbiased predictor, we see that IDW is inferior to OK. Some empirical work demonstrates that the ordinary kriging variance is often 10% to 30% less than the prediction error variance associated with IDW.
- IDW does not account for how strong (or weak) the spatial correlation is.
- Prediction error variances are not usually reported with IDW interpolations (though they could be).

OK and IDW weights and variances for toy example on page 80:

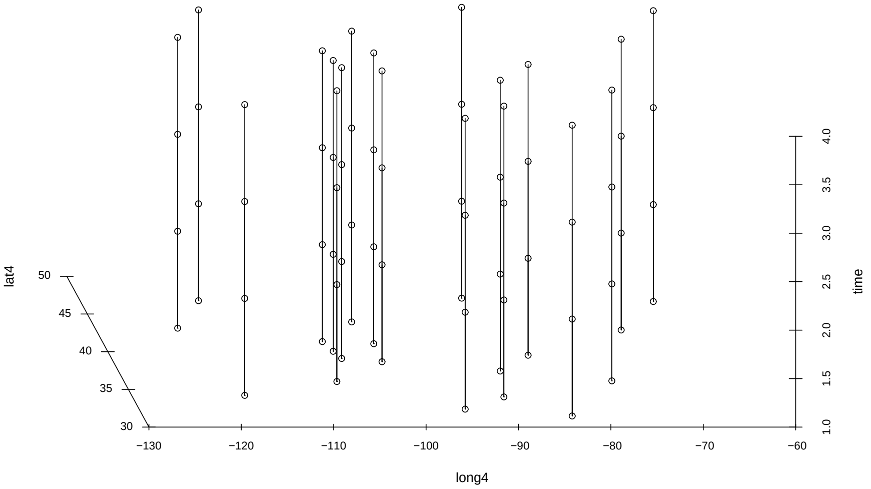
	OK	IDW	IDSW
λ_1	.017	.116	.074
λ_2	.422	.259	.370
λ_3	.065	.129	.093
λ_4	.218	.183	.185
λ_5	.031	.129	.093
λ_6	.246	.183	.185
PEV	.478	.506	.488

(i) *Spatio-Temporal Regression and Prediction*

Suppose that we have observed geostatistical data at each of m time points, i.e.,

$$\{[Z(\mathbf{s}_1, t_i), \ldots, Z(\mathbf{s}_n, t_i)] : i = 1, \ldots, m\}.$$

Here, $\mathbf{s}_1, \ldots, \mathbf{s}_n$ are the data locations (assumed to be the same across time) and $t_1 < t_2 < \cdots < t_m$ are the observation times.



We can view such data in two ways:

1. A collection of spatially correlated time series
2. A collection of temporally correlated random fields

If the temporal correlation is weak and the data are rectangular, then sampling across time gives nearly independent replications of the random field. One consequence of this is that it is no longer necessary to assume spatial stationarity. If the temporal correlation is non-neglible, however, then we generally proceed by assuming spatial and temporal stationarity of some kind.

Two important inference problems:

- Estimate parameters (particularly those in the mean structure).
- Predict $Z(\mathbf{s}_0, t_0)$. Typically, $t_0 \geq t_m$, i.e., the time at which we want to predict is either at the most recent time of observation or some time in the future; $\mathbf{s}_0$ may or may not be one of the data locations.

In principle we could use ideas from spatial regression and kriging to perform spatio-temporal regression and spatio-temporal prediction, but there are some new issues to consider:

- Data collected over time sometimes exhibit periodicity (e.g. seasonality). This can be dealt with by using a mean model that has sine and cosine terms, e.g.

$$\begin{aligned} E(Z(\mathbf{s},t)) &= \beta_1 + \beta_2 u + \beta_3 v + \beta_4 u^2 + \beta_5 uv + \beta_6 v^2 \\ &\quad + \beta_7 t + \beta_8 \cos(2\pi\beta_{10}t) + \beta_9 \sin(2\pi\beta_{10}t) \end{aligned}$$

- We must use a valid (nonnegative definite) space-time covariance function or semivariogram to model the variance-covariance structure of the residuals. Suppose $C(\|\mathbf{h}\|)$ is a valid isotropic spatial covariance function for $\mathbf{h} = (h_1, h_2) \in R^2$. To construct a valid space-time covariance function, let t now represent the time lag and append it to $\mathbf{h}$; then consider using the same isotropic function, i.e.,

  $$C(\mathbf{h},t) = C(\sqrt{h_1^2 + h_2^2 + t^2})$$

  Unfortunately, the resulting covariance function is not always valid in R^3. Even if such a covariance function were valid, it may not be sensible because t is measured in different units than spatial distance. Some methods for dealing with this are as follows:

  - Include an extra parameter to scale properly for time, i.e.,

    $$C(\mathbf{h},t) = C\left(\sqrt{h_1^2 + h_2^2 + \psi t^2}\right)$$

    where $\psi \geq 0$. This is essentially a special type of geometrically anisotropic model.
  - Assume space-time additivity, i.e.

    $$C(\mathbf{h},t) = C_S(\mathbf{h}) + C_T(t)$$

  - Assume space-time separability, i.e.

    $$C(\mathbf{h},t) = C_S(\mathbf{h})C_T(t)$$

Assuming that the aforementioned modeling issues are dealt with satisfactorily, the model can be expressed as a general linear model

$$\mathbf{Z} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \mathbf{V})$$

and we can proceed with MLE and other likelihood-based inferences, and with kriging, as we do for purely spatial data.

The only widely available software that implements this is SAS PROC MIXED. It can handle certain separable space-time covariance structures.

VII. MODELS AND INFERENCE FOR LATTICE DATA

Recall the definition of lattice data: observations are taken at a finite number of sites whose whole constitutes the entire study region. Unlike geostatistics, there is no possibility of a response "between" data locations. Therefore, there is generally no need to predict unobserved values.

Lattice data locations may be points or regions, but most cases where the data are points can be handled using geostatistics (without the kriging step). So we shall focus primarily on situations where data locations are regions. Furthermore, we shall generally consider only the two-dimensional case.

Examples:

- Presence or absence of a plant species in square quadrats over a study area
- Numbers of deaths due to SIDS in the counties of North Carolina
- Pixel values from remote sensing (satellites)

Primary goals of statistical analysis:

- formulate a model
- estimate model parameters
- test hypotheses about model parameters
- compare models

Basis for model formulation:

- EDA techniques (such as those described previously)
- Measuring and testing for spatial correlation among the observations

1. Measuring and Testing for Spatial Autocorrelation

Our objective is to measure how strong the tendency is for observations from nearby regions to be more (or less) alike than observations from regions farther apart, and then judge whether any apparent tendency is sufficiently strong that it is unlikely to be due to chance alone.

Examples of spatially autocorrelated data:

A. The General Cross-Product Statistic

Notation:

- Let Z_i denote the response at the ith location ($i = 1, \ldots, n$).
- Let Y_{ij} be a measure of how similar or dissimilar the responses are at locations i and j.
- Let W_{ij} be a measure of the spatial proximity of locations i and j.
- For future reference, define matrices $\mathbf{Y} = (Y_{ij})$ and $\mathbf{W} = (W_{ij})$.

The general cross-product statistic is

$$C = \sum_i \sum_j W_{ij} Y_{ij}.$$

Toy example: Consider trinary Z_i's, $Y_{ij} = (Z_i - Z_j)^2$, and

$$W_{ij} = \begin{cases} 1, & \text{if locations } i \text{ and } j \text{ are adjacent} \\ 0, & \text{otherwise.} \end{cases}$$

Note in this example,

- C too small $\Rightarrow$ positive spatial autocorrelation
- C too large $\Rightarrow$ negative spatial autocorrelation

B. Evaluating the Statistical Significance of C

- Comparison to randomization distribution
  - List all possible arrangements of the observed responses over the locations (obtained by permutation of responses).
  - Compute C for each arrangement, and rank these.
  - Determine where the data's C-value fits in; P-value for the test is the number of C-values in the randomization distribution as extreme or more extreme than the observed C.
  - Can do one-sided or two-sided tests.
  - Toy example (continued):

- Monte Carlo approach

  Motivated by the fact that complete enumeration of the possible arrangements may be computationally prohibitive even for moderately-sized data sets. So instead, obtain a random sample from the randomization distribution and follow the same type of procedure.

  1. Assign the numbers $1, 2, \ldots, n$ randomly to the data locations.
  2. Reassign the observations to data locations in accordance with this realization of random numbers (the smallest observed value is assigned to the data location assigned "1", the next smallest observed value is assigned to the data location assigned "2", etc.
  3. Compute C for this realization.
  4. Repeat steps 1-3 m times.
  5. Compute the P-value, which estimates the proportion of C-values as extreme or more extreme than the observed C, by

$$P = \frac{1 + \text{number of } C\text{-values} \geq \text{observed } C}{1 + m}.$$

- Normal approximation, i.e., $C \mathrel{\dot{\sim}} N(E(C), \text{var}(C))$.
  - Define
$$S_0 = \sum\sum_{i \neq j} W_{ij}, \qquad S_1 = \frac{1}{2}\sum\sum_{i \neq j}(W_{ij} + W_{ji})^2, \qquad S_2 = \sum_i (W_{i\cdot} + W_{\cdot i})^2$$
  - Define T_0, T_1, and T_2 similarly but for the Y_{ij}'s.
  - Compute
$$z = \frac{|C - E(C)| - 1}{\sqrt{\text{var}(C)}}$$
where
$$E(C) = \frac{S_0 T_0}{n(n-1)}$$
and
$$\text{var}(C) = \frac{S_1 T_1}{2n(n-1)} + \frac{(S_2 - 2S_1)(T_2 - 2T_1)}{4n(n-1)(n-2)} + \frac{(S_0^2 + S_1 - S_2)(T_0^2 + T_1 - T_2)}{n(n-1)(n-2)(n-3)} - [E(C)]^2.$$
  - Rule of thumb: To use this approximation, n should be at least 25.

C. Join-Count Statistics (for use with binary data)

Based on a map of the data coded as either 1 (black) or 0 (white).

Spatial correlation will manifest as a pattern in which neighboring locations are more likely to display the same color (or opposite colors) than would be expected in the absence of spatial autocorrelation. To quantify this:

1. Classify all of the "joins" between contiguous regions as BB, BW, or WW. The most popular definition of contiguity regards two sites as joined if they share an edge. In the context of a rectangular lattice, this is called the rook's definition. Other definitions (e.g. bishop's and queen's) are possible.
2. Count the number of joins of a specified type, e.g. the # of BW joins $\equiv BW$.
3. Test this count for significance using one of the aforementioned testing approaches.

If we define $W_{ij} = 1$ if regions i and j are joined (and 0 otherwise) and define $Y_{ij} = \frac{1}{2}(Z_i - Z_j)^2$, then

$$BW = \sum_i \sum_j W_{ij} Y_{ij},$$

i.e. BW is a special type of generalized cross-product statistic.

Likewise, BB is a special type of generalized cross-product statistic with $Y_{ij} = \frac{1}{2} Z_i Z_j$.

BW seems to be slightly more sensitive at detecting autocorrelation than the other two, so we will consider BW in further detail. Notation and results associated with the use of BW:

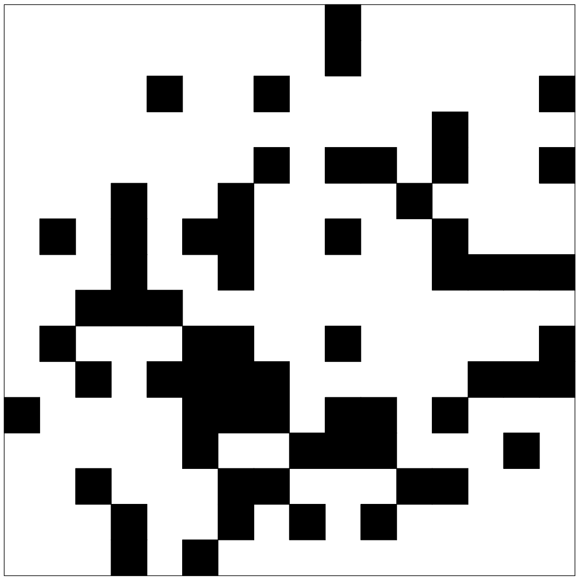
- Let $b = \#$ black regions and $w = \#$ white regions; note that $b + w = n$.
- Can be shown that

$$T_0 = bw, \qquad T_1 = T_0 = bw, \qquad T_2 = nbw.$$

- If regions form a rectangular $r \times c$ lattice, and the rook's contiguity definition is used, then

$$S_0 = 2(2rc - r - c), \qquad S_1 = 2S_0, \qquad S_2 = 8(8rc - 7r - 7c + 4).$$

Example — the *Atriplex hymenelytra* data:



Remarks:

- The same approach can be used for data at irregularly spaced and shaped locations, though the formulas given for S_0, S_1, and S_2 no longer apply. (The formulas given for T_0, T_1, and T_2 are still okay though.)
- The test presented in the example is two-sided but a one-sided test, if more appropriate, can be done easily.
- The same approach can be used for BB (and WW) joins. In the case of BB joins,

$$T_0 = \frac{1}{2}b(b-1), \qquad T_1 = T_0, \qquad T_2 = b(b-1)^2.$$

- Extensions to polytomous categorical data (i.e. a multi-colored map) are possible.

D. Moran's and Geary's Statistics (polytomous and continuous data)

Moran proposed the following autocorrelation statistic which can be used with polytomous and continuous data:

$$I = \frac{n \sum_i \sum_j W_{ij}(Z_i - \bar{Z})(Z_j - \bar{Z})}{S_0 \sum_i (Z_i - \bar{Z})^2},$$

where $\bar{Z} = \sum_i Z_i / n$.

Geary proposed a similar statistic:

$$c = \frac{(n-1) \sum_i \sum_j W_{ij}(Z_i - Z_j)^2}{S_0 \sum_i (Z_i - \bar{Z})^2}.$$

Remarks:

- I resembles the ordinary correlation coefficient.
- Note that $I = \frac{n}{S_0 \sum_i (Z_i - \bar{Z})^2} \cdot C$ if we take $Y_{ij} = (Z_i - \bar{Z})(Z_j - \bar{Z})$.
- Similarly, c may be related to C by taking $Y_{ij} = (Z_i - Z_j)^2$.
- I seems to be more popular than c, so we'll focus on I.
- $E(I) = -\frac{1}{n-1}$ under independence.
- $I > -\frac{1}{n-1} \Rightarrow$ positive autocorrelation.
- $I < -\frac{1}{n-1} \Rightarrow$ negative autocorrelation.
- Normal approximation to distribution of I under independence ($n > 25$):

$$E(I) = -\frac{1}{n-1}$$

$$\begin{aligned} \mathrm{var}(I) &= \frac{[(n^2 - 3n + 3)S_1 - nS_2 + 3S_0^2] - k[n(n-1)S_1 - 2nS_2 + 6S_0^2]}{(n-1)(n-2)} \\ &\quad - \frac{1}{(n-1)^2} \end{aligned}$$

where

$$k = \frac{\sum_i (Z_i - \bar{Z})^4}{(\sum_i (Z_i - \bar{Z})^2)^2}.$$

- For smaller sample sizes, can use randomization distribution or Monte Carlo approach to evaluate significance.

E. Generalized Proximity Values

So far our discussion of join-count statistics and the Moran and Geary statistics has assumed that the W_{ij}'s are binary (0 or 1). This is rather crude. In many situations we may be able to measure spatial proximity on a more refined scale (as we do the Y_{ij}'s in going from BW to I or c).

Possible refinements:

1. Use lengths of common boundary; this may more accurately reflect the amount of inter-site interaction.
2. Use actual distance between locations or centroids of locations, e.g. the inverse of Euclidean or city-block distance between locations. This recognizes the fact that interaction between sites does not usually terminate sharply beyond places that share a boundary.
3. Incorporate directionality (e.g. upstream vs. downstream) by allowing $W_{ij} \neq W_{ji}$.

A side benefit of using non-binary W_{ij}'s is that the distribution of the test statistic under independence is better approximated by the normal distribution.

F. Spatial Autocorrelation Functions

The statistics considered so far attempt to express information about spatial autocorrelation in a single number. Alternatively, we could consider regarding spatial autocorrelation as a function of distance. That is:

1. Divide the range of distances into q classes.
2. Compute a previously considered spatial autocorrelation measure, e.g. I, once for each of the q distance classes; in other words, we use only those pairs of locations that are within the same distance class.
3. Plot the statistic, e.g. I_d, versus d. Such a plot is called the *correlogram* corresponding to that statistic.

This last notion is essentially what we did with geostatistical data, when we measured spatial dependence via the covariance function or semivariogram.

2. Models for Lattice Data

Suppose that we find, either through EDA or formal testing, that the (lattice) data appear to be spatially correlated. Then what?

Formulate one or more models that will allow responses to be correlated with responses at nearby locations. The most useful models are:

- reasonably parsimonious
- readily interpreted
- computationally feasible to fit to the data

Lattice data can be continuous or discrete. Initially, we will restrict attention to continuous data. Modeling binary and polytomous categorical lattice data is also known as *image analysis*. Many concepts in image analysis are similar to those presented here, but it is somewhat easier to model dependence in the continuous case.

A. *Models for Data in One Dimension*

Due to the discrete nature of the spatial locations of lattice data, the most popular models are similar to commonly used models for discrete time series. Let us digress from spatial statistics for a moment and review one very important time series model: the autoregressive model of order one [AR(1)]:

$$Z_t = \rho Z_{t-1} + \epsilon_t, \qquad \{\epsilon_t\} \sim \text{iid N}(0, \sigma^2), \qquad t = 0, \pm 1, \pm 2, \ldots$$

where $\rho \in (-1, 1)$ is called the autoregressive coefficient. Recall that $\text{corr}(Z_t, Z_{t-1}) = \rho$, $\text{corr}(Z_t, Z_{t-2}) = \rho^2$, and more generally, $\text{corr}(Z_t, Z_{t-k}) = \rho^k$.

Actually, there are two ways to specify a first-order autoregressive model:

1. A *simultaneous* AR(1), as we've just given.

2. A *conditional* AR(1), as follows:

$$Z_t | Z_{t-1} \sim \text{independent } \mathrm{N}(\rho Z_{t-1}, \sigma^2(1-\rho^2)) \quad t = 0, \pm 1, \pm 2, \ldots .$$

It turns out that these two specifications are equivalent, i.e. they produce responses $Z_1, Z_2, \ldots, Z_n$ that have the same joint distribution. Note that interaction is "one-sided" in these models due to the unidirectional flow of time.

As a partial step towards generalizing the notion of an autoregressive model to spatial data, consider observations $Z_1, Z_2, \ldots, Z_n$ taken at regularly-spaced locations on a line:

Though one-dimensional, our consideration of this situation differs from the classical time series situation in two ways:

- Interactions may be "two-sided."
- The domain is bounded.

The one-sided simultaneous and conditional AR(1) models, restricted to $s = 1, \ldots, n$, could be applied to this situation; alternatively we could consider two-sided versions. In its simultaneously-specified form, the two-sided version is as follows:

$$\begin{aligned}
Z_1 &= \rho Z_2 + \epsilon_1, \\
Z_s &= \frac{\rho}{2} Z_{s-1} + \frac{\rho}{2} Z_{s+1} + \epsilon_s, \qquad s = 2, 3, \ldots, n-1 \\
Z_n &= \rho Z_{n-1} + \epsilon_n, \\
&\quad \{\epsilon_s\} \sim \text{iid } \mathrm{N}(0, \sigma^2).
\end{aligned}$$

Note that we needed to account for "edge effects," i.e. the fact that Z_1 and Z_n each have only one neighboring observation.

B. *Models for Data in Two Dimensions*

How might we generalize to two dimensions? Consider a rectangular $R \times C$ lattice of equally spaced sites and let $Z_{u,v}$ denote the observation in row u and column v. Then consider the simultaneous spatial autoregressive model

$$\begin{aligned} Z_{u,v} &= \frac{\rho}{4}(Z_{u-1,v} + Z_{u+1,v} + Z_{u,v-1} + Z_{u,v+1}) + \epsilon_{u,v}, \\ &\quad u = 2, 3, \ldots, R-1, \quad v = 2, 3, \ldots, C-1 \end{aligned}$$

with appropriate modifications for edge sites, and with errors $\epsilon_{u,v}$ satisfying $\{\epsilon_{u,v}\} \sim \text{iid N}(0, \sigma^2)$.

We can write this last representation using matrix notation, as follows:

$$\mathbf{Z} = \rho \mathbf{W} \mathbf{Z} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathrm{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}),$$

where $\mathbf{W}$ is an $n \times n$ matrix (with $n = RC$) whose nonzero elements specify the neighbors of each site. For example, if $R = C = 3$, then $\mathbf{W}$ is given by:

We can generalize this spatial autoregression idea in many ways:

1. Alternative definitions of neighbors.

$$Z_{u,v} = \rho \cdot \frac{\sum_{(j,k)\in N_{u,v}} Z_{j,k}}{|N_{u,v}|} + \epsilon_{u,v}, \qquad \{\epsilon_{u,v}\} \sim \text{iid N}(0, \sigma^2),$$

   where $N_{u,v}$ is the set of neighbors of location (u, v) and $|N_{u,v}|$ is the number of those neighbors.

2. Allow for irregular lattices.

$$Z_i = \rho \cdot \frac{\sum_{j\in N_i} Z_j}{|N_i|} + \epsilon_i, \qquad \{\epsilon_i\} \sim \text{iid N}(0, \sigma^2), \qquad i = 1, \ldots, n$$

   where N_i is the set of neighbors of location i and $|N_i|$ is the number of those neighbors.

3. Anisotropic and higher-order models.

$$Z_i = \rho_1 \cdot \frac{\sum_{j\in N_{1i}} Z_j}{|N_{1i}|} + \rho_2 \cdot \frac{\sum_{j\in N_{2i}} Z_j}{|N_{2i}|} + \epsilon_i, \qquad \{\epsilon_i\} \sim \text{iid N}(0, \sigma^2), \quad i = 1, \ldots, n.$$

4. Nonzero weights assigned to each site.

$$Z_i = \sum_j S_{ij} Z_j + \epsilon_i, \qquad \{\epsilon_i\} \sim \mathrm{N}(0, \sigma^2), \qquad i = 1, \ldots, n.$$

   However, to be useful in practice, the S_{ij}'s must be parameterized in terms of a much smaller set of parameters.

5. Allow for trend.

$$Z_i - \mu_i = \sum_j S_{ij}(Z_j - \mu_j) + \epsilon_i, \qquad \{\epsilon_i\} \sim \mathrm{N}(0, \sigma^2), \qquad i = 1, \ldots, n.$$

Spatial autoregressive models (general form):

1. Simultaneous autoregression (SAR)

$$Z_i - \mu_i = \sum_j S_{ij}(Z_j - \mu_j) + \epsilon_i, \qquad \{\epsilon_i\} \sim \mathrm{N}(0, \sigma^2), \qquad i = 1, \ldots, n$$

   where $\mathbf{S} \equiv \{S_{ij}\}$ is such that $S_{ii} = 0$ and $\mathbf{I} - \mathbf{S}$ is nonsingular. In matrix form, we can write

$$\mathbf{Z} - \boldsymbol{\mu} = \mathbf{S}(\mathbf{Z} - \boldsymbol{\mu}) + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathrm{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}).$$

2. Conditional autoregression (CAR) (also known as Markov random fields, MRFs)

$$(Z_i | Z_j, j \neq i) \sim \mathrm{N}(\mu_i + \sum_j C_{ij}(Z_j - \mu_j), \sigma^2),$$

   where $\mathbf{C} \equiv \{C_{ij}\}$ is such that $C_{ii} = 0$ and $\mathbf{I} - \mathbf{C}$ is symmetric and positive definite.

In contrast to an AR model for time series, the two specifications for spatial data yield different models, i.e., if we take $C_{ij} = S_{ij}$ the CAR yields responses whose joint distribution is different than for the SAR.

Specifically, it can be shown that

$$\mathbf{Z} \sim \begin{cases} N\left(\boldsymbol{\mu}, \sigma^2[(\mathbf{I}-\mathbf{S})^{-1}(\mathbf{I}-\mathbf{S}')^{-1}]\right) & \text{for an SAR} \\ N\left(\boldsymbol{\mu}, \sigma^2(\mathbf{I}-\mathbf{C})^{-1}\right) & \text{for a CAR} \end{cases}$$

i.e. $\mathbf{Z} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \mathbf{V})$

Note that in either case, $\mathbf{V} = \sigma^2\mathbf{R}$, but $\mathbf{R}$ is not a correlation matrix.

We can generalize still further to allow the errors ("innovations") in the two AR models to be heteroscedastic. That is, we may allow $\epsilon_i \sim$ independent $\mathrm{N}(0, \sigma_i^2)$ (for an SAR) or $\mathrm{var}(Z_i|Z_j, j \neq i) = \sigma_i^2$ (for a CAR). For these generalizations,

$$\mathrm{var}(\mathbf{Z}) = \begin{cases} \sigma^2[(\mathbf{I}-\mathbf{S})^{-1}\mathbf{D}(\mathbf{I}-\mathbf{S}')^{-1}] & \text{for an SAR} \\ \sigma^2(\mathbf{I}-\mathbf{C})^{-1}\mathbf{D} & \text{for a CAR} \end{cases}$$

where $\mathbf{D} = \mathrm{diag}(\sigma_i^2/\sigma^2)$.

Actually, for the CAR result above to hold, $(\mathbf{I}-\mathbf{C})^{-1}\mathbf{D}$ must be positive definite. Sufficient conditions on the elements of $\mathbf{C}$ and $\mathbf{D}$ for positive definiteness are apparently not easy to specify. A necessary (but not sufficient) condition is

$$c_{ij}\sigma_j^2 = c_{ji}\sigma_i^2 \quad \text{for all } i \text{ and } j.$$

Spatial moving average models (general form):

$$Z_i - \mu_i = \sum_j M_{ij}\epsilon_j, \qquad \{\epsilon_i\} \sim \text{iid N}(0, \sigma^2), \qquad i = 1, \dots, n,$$

with $M_{ii} = 1$. In matrix form, we write

$$\mathbf{Z} - \boldsymbol{\mu} = \mathbf{M}\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \text{N}_n(\mathbf{0}, \sigma^2\mathbf{I}).$$

- Have not received as much attention as spatial autoregressive models.
- For example,

$$Z_{u,v} - \mu_{u,v} = \epsilon_{u,v} + \nu(\epsilon_{u-1,v} + \epsilon_{u+1,v} + \epsilon_{u,v-1} + \epsilon_{u,v+1})$$

- Can generalize to allow heteroscedasticity of errors by taking $\boldsymbol{\epsilon} \sim \text{N}_n(\mathbf{0}, \sigma^2\mathbf{D})$.

Adapting geostatistical models:

- For point data, any geostatistical model for the covariance structure can be used without modification.
- For areal data, covariances among responses can be obtained by integration of a geostatistical covariance function.

3. Inference for Lattice Data

A. *The Likelihood Function*

Joint distributions under various models:

- For the SAR, $\mathbf{Z} \sim \mathrm{N}(\boldsymbol{\mu}, \sigma^2(\mathbf{I}-\mathbf{S})^{-1}(\mathbf{I}-\mathbf{S}')^{-1})$.

- For the CAR, $\mathbf{Z} \sim \mathrm{N}(\boldsymbol{\mu}, \sigma^2(\mathbf{I}-\mathbf{C})^{-1})$.
- For the MA, $\mathbf{Z} \sim \mathrm{N}(\boldsymbol{\mu}, \sigma^2\mathbf{M}\mathbf{M}')$.
- For geostatistical models, $\mathbf{Z} \sim \mathrm{N}(\boldsymbol{\mu}, \sigma^2\mathbf{R})$.

Corresponding to each model is a space of parameter values within which the covariance matrix is positive definite.

The log-likelihood function associated with $\mathbf{Z}$ is

$$L(\boldsymbol{\mu}, \sigma^2, \mathbf{B}) = -\frac{n}{2}\log(2\pi\sigma^2) + \frac{1}{2}\log|\mathbf{B}| - \frac{1}{2\sigma^2}(\mathbf{Z}-\boldsymbol{\mu})'\mathbf{B}(\mathbf{Z}-\boldsymbol{\mu})$$

where

$$\mathbf{B} = \begin{cases} (\mathbf{I}-\mathbf{S}')(\mathbf{I}-\mathbf{S}) & \text{for a SAR} \\ \mathbf{I}-\mathbf{C} & \text{for a CAR} \\ (\mathbf{M}\mathbf{M}')^{-1} & \text{for an MA} \\ \mathbf{R}^{-1} & \text{for a geostatistical model} \end{cases}$$

Usually the mean is parameterized by a linear model, i.e.

$$\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}.$$

B. *Maximum Likelihood Estimation*

The maximum likelihood estimators (MLEs) of $\boldsymbol{\beta}$, σ^2, and the parameters in $\mathbf{B}$ are given by

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}'\hat{\mathbf{B}}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{B}}\mathbf{Z},\\ \hat{\sigma}^2 &= \frac{1}{n}(\mathbf{Z}-\mathbf{X}\hat{\boldsymbol{\beta}})'\hat{\mathbf{B}}(\mathbf{Z}-\mathbf{X}\hat{\boldsymbol{\beta}}),\end{aligned}$$

and $\hat{\mathbf{B}}$ maximizes the "profile log-likelihood" function

$$L^*(\mathbf{B}) = -n\log(\hat{\sigma}^2) + \log|\mathbf{B}|.$$

Remarks:

- Grid search or other numerical optimization methods (e.g. Nelder-Mead simplex or Newton-Raphson) can be used.
- Care must be taken to ensure that estimates remain in the allowable parameter space.
- Multiple modes are possible.
- Standard errors for parameter estimates have been obtained under an assumption of "increasing-domain asymptotics." Confidence intervals for parameters can be constructed using these.

C. *Model Comparisons*

We may want to compare how well two or more models fit the data. For example:

- CAR vs. SAR or CAR vs. MA
- First-order CAR vs. Second-order CAR
- Isotropy vs. Anisotropy
- Different neighborhood definitions
- Constant mean vs. Planar trend

Nested models can be compared using a likelihood ratio test. Suppose the nested model imposes c independent constraints on the parameters. Let:

$$\begin{aligned}L_1 &= \text{maximized log-likelihood for the larger model,}\\ L_0 &= \text{maximized log-likelihood for the smaller model.}\end{aligned}$$

Compare $2(L_1 - L_0)$ to $\chi^2_{c,\alpha}$.

Non-nested models can be compared using penalized likelihood criteria:

$$AIC = L(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2, \hat{\mathbf{B}}) - \#\text{ parameters}$$

$$BIC = L(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2, \hat{\mathbf{B}}) - \frac{\log n}{2} \cdot (\#\text{ parameters})$$

Larger values are associated with better-fitting models.

4. Markov Random Field Models

The Gaussian CAR model we've discussed previously is a special case of a larger class of models called Markov random field models. The word "Markov" refers to the fact that the conditional distribution of $Z(\mathbf{s}_i)$, given $\{Z(\mathbf{s}_j) : j \neq i\}$, actually depends functionally only on $\{Z(\mathbf{s}_j) : j \in N_i\}$ where N_i is the set of neighbors of site i; thus, letting $f_i(\cdot)$ represent the *full* conditional probability density function of $Z(\mathbf{s}_i)$, we have

$$f_i(z(\mathbf{s}_i)|\{z(\mathbf{s}_j) : j \neq i\}) = f_i(z(\mathbf{s}_i)|\{z(\mathbf{s}_j) : j \in N_i\}).$$

Note that this does *not* say that there is no dependence marginally between $Z(\mathbf{s}_i)$ and values of Z outside $\mathbf{s}_i$'s neighborhood, but it does say that there is no conditional dependence on them.

Can we specify Markov random field models other than Gaussian? For example, could we model a random quantity that at any given site is binomial or Poisson, conditional on the quantities at that site's neighbors? The answer is yes, in some cases. However, the mathematics of deriving such models, obtaining conditions for their existence, etc. are quite difficult, so we shall not delve further into them, except to give two examples:

1. The auto-logistic model for binary data,

$$\Pr(z(\mathbf{s}_i)|\{z(\mathbf{s}_j) : j \neq i\}) = \frac{\exp[\alpha_i z(\mathbf{s}_i) + \sum_{j=1}^n \theta_{ij} z(\mathbf{s}_i) z(\mathbf{s}_j)]}{1 + \exp[\alpha_i + \sum_{j=1}^n \theta_{ij} z(\mathbf{s}_j)]}$$

2. The auto-Poisson model for count data,

$$\Pr(z(\mathbf{s}_i)|\{z(\mathbf{s}_j) : j \neq i\}) = \exp[-\lambda_i(\{z(\mathbf{s}_j) : j \neq i\})][\lambda_i(\{z(\mathbf{s}_j) : j \neq i\})]^{z(\mathbf{s}_i)}/z(\mathbf{s}_i)!,$$

where

$$\lambda_i(\{z(\mathbf{s}_j) : j \neq i\}) = \exp[\alpha_i + \sum_{j=1}^n \theta_{ij} z(\mathbf{s}_j)].$$

However, this model can be used only to model <u>negative</u> spatial dependence.

5. Disease Mapping

Some basics:

- The data are counts of disease cases, deaths, etc. over areas (e.g. counties).
- The attribute variables that are modeled are usually not the raw counts, but instead are counts adjusted for the population at risk in those areas. Let E_i be the expected count for site i, and assume that these expected counts are known. Depending on the attribute under study, the expected counts may be the number of people, live births, men 65 and older, etc.
- A hierarchical Bayesian approach is taken to modeling the data.
- A Gaussian Markov random field is used at one of the three stages of the hierarchy.

Two popular model formulations:

1. Hierarchical Gaussian-Gaussian model. First the count data are transformed to both adjust for the at-risk population size and to produce variables for which the assumption of a Gaussian distribution may be reasonable. Such transformations include the square root transformation,

   $$Y_i \equiv (Z_i/E_i)^{1/2} \quad (i = 1, \ldots, n)$$

   and the Freeman-Tukey transformation, which stabilizes the variance over a greater range of expected counts,

   $$Y_i \equiv (Z_i/E_i)^{1/2} + [(Z_i + 1)/E_i]^{1/2} \quad (i = 1, \ldots, n).$$

   Then, letting $\mu_i \equiv E(Y_i)$ and $\boldsymbol{\mu} = (\mu_i)$, we suppose that

   $$\begin{aligned} \mathbf{Y}|\boldsymbol{\mu}, \sigma^2 &\sim N(\boldsymbol{\mu}, \sigma^2\mathbf{I}), \\ \boldsymbol{\mu}|\boldsymbol{\beta}, \mathbf{C}, \mathbf{M} &\sim N(\mathbf{X}\boldsymbol{\beta}, (\mathbf{I} - \mathbf{C})^{-1}\mathbf{M}). \end{aligned}$$

   A fully Bayesian model is then completed by specifying prior distributions for σ^2, $\boldsymbol{\beta}$, and the parameters in $\mathbf{C}$ and $\mathbf{M}$.

2. Poisson-log Gaussian model.

   $$\begin{aligned} Z_i|\lambda_i, E_i &\sim \text{Poisson}(\lambda_i E_i), \quad (i = 1, \ldots, n) \\ \boldsymbol{\theta}|\boldsymbol{\beta}, \mathbf{C}, \mathbf{M} &\sim N(\mathbf{X}\boldsymbol{\beta}, (\mathbf{I} - \mathbf{C})^{-1}\mathbf{M}) \end{aligned}$$

   where $\boldsymbol{\theta} = (\theta_i) = (\log \lambda_i)$. The model is completed by specifying prior distributions for $\boldsymbol{\beta}$ and the parameters in $\mathbf{C}$ and $\mathbf{M}$.

Estimation and inference for both formulations are performed via Markov chain Monte Carlo simulation methods.

VIII. CONCEPTS AND MODELS FOR SPATIAL POINT PATTERNS

1. Basic Terminology and Set-Up

A (univariate) *spatial point pattern* (SPP) is a finite set of points in a spatial domain (study region) D whose locations are modelled as random variables. The term "univariate" refers to the fact that points in the pattern are not distinguished into various types.

Statistically, a SPP is regarded as a partial realization of a (univariate) *spatial point process*, which is a random mechanism for generating a countable set of points in D.

Note: For an ordinary SPP, there is no datum attached to the spatial label. Rather, the spatial label itself can be regarded as the datum. This is in contrast to a *marked* point pattern, for which one or more additional variables (e.g. species, soil moisture) are measured at each point.

Subsequently, we refer to the points in a spatial point pattern or process as *events*, to distinguish them from arbitrary points in D.

We will consider only $d = 2$, but many things can be specialized to $d = 1$ or extended to $d = 3$ (or higher).

Often the objects being regarded as a SPP are actually areal, i.e. they have positive area (e.g. trees, towns). However, the relative scale of the objects and the study area usually permits us to reasonably represent the objects as points.

The study area D may be irregularly shaped (e.g. an island) or "regular" (e.g. a rectangle). Typically the latter case results from taking D to be a (hopefully) representative subset of some larger region.

The characterization of spatial pattern is often an important component of a SPP analysis. When accompanied by more detailed quantitative analysis, it can be used, for example, to support or refute a hypothesis about a phenomenon of interest (e.g. territoriality, social interactions, environmental heterogeneity, allelopathy).

What do we mean by "pattern"? A working definition is as follows:

> "Pattern is the characteristic of a set of points [events] which describes the location of these points in terms of the relative distances and orientations of one point or one group of points to another point or group of points, at one or more scales of observation."

2. Four-way Classification of SPPs (qualitative, single scale of observation, admittedly simplistic):

1. Completely random (Complete spatial randomness, CSR) – No obvious structure, a random sample from the uniform distribution on D
2. Aggregated (Clustered, Clumped) – clusters of events separated by relatively empty areas
3. Regular (Overdispersed, Inhibitory, Superuniform) – events rather evenly spaced
4. Heterogeneous – events unevenly spaced, often difficult to distinguish from aggregation (exception: smooth trend in point intensity)

May need to examine pattern on more than one scale. For example:

Nevertheless the 4-way classification is useful at each scale.

3. Important Objectives of Statistical Analysis

- What is the nature of the spatial pattern? Is it aggregated, regular, or completely random? Is there a trend or some other form of spatial heterogeneity? Is one characterization sufficient for the scale of interest?
- (For sparsely sampled patterns) What is the intensity of the underlying process, i.e., how many trees are out there?
- Can we model the process that we envisage has generated the data? Can we do statistical inference on the parameters of this model?

Appropriate statistical methods for addressing these questions depend on:

- the extent of sampling (completely mapped or sparsely sampled)
- the type of sampling (areal or distance sampling)

4. Areal and Distance Methods

(a) Areal methods:

- Based on a reduction of the SPP to counts of events within nonoverlapping subregions, i.e. *quadrats*, of equal size

- Quadrats need not be rectangular (but usually are)
- Quadrats may or may not constitute an exhaustive partition of D.

(b) Distance methods:

- Based on a reduction of the SPP to distances to events
- May utilize interevent distances (e.g. distance of an event to its nearest neighbor) or point-to-event distances, or both
- May utilize distances only to nearest events, or to events beyond the nearest, or both

(c) Advantage and disadvantages

- Areal methods emphasize "global" information at the expense of "local" information; vice versa for distance methods.
- Size and shape of quadrats are arbitrary, and different choices can give you different answers.
- Two problems with distance methods are edge effects and overlap effects.

Edge effects: Distance measurements taken near the boundary of D will tend to be larger than those taken in the interior, since points or events near the boundary are denied the possibility of neighboring events outside the boundary.

Possible remedies:

- Restrict attention to points or events in interior, surrounded by "buffer zone"

- If D is rectangular, connect opposite edges (toroidal edge correction)

- Truncate the search distance and incorporate this into the distribution of distance measurements

Overlap effects: "Search areas" for the nearest event can overlap, resulting in dependent measurements.

Possible remedies:

- Use sparse sampling (undesirable, however, for completely mapped patterns)
- Truncate search distances to prevent overlap

5. Notation and Some Important Descriptive Functions and Properties

- $N(B)$: # of events in an arbitrary region $B \subset D$
- $|B|$: area of B
- $d\mathbf{s}$: infinitesimal region containing $\mathbf{s} \in D$
- Intensity function (first-order):

$$\lambda(\mathbf{s}) = \lim_{|d\mathbf{s}| \to 0} \left(\frac{E[N(d\mathbf{s})]}{|d\mathbf{s}|} \right)$$

- Second-order intensity function:

$$\lambda_2(\mathbf{s}, \mathbf{t}) = \lim_{|d\mathbf{s}|, |d\mathbf{t}| \to 0} \left\{ \frac{E[N(d\mathbf{s})N(d\mathbf{t})]}{|d\mathbf{s}||d\mathbf{t}|} \right\}$$

- Covariance intensity function:

$$C(\mathbf{s}, \mathbf{t}) = \lambda_2(\mathbf{s}, \mathbf{t}) - \lambda(\mathbf{s})\lambda(\mathbf{t})$$

The intensity function is analogous to the mean function of a random field, and the covariance intensity function is analogous to the covariance function of a random field.

Stationarity

- A process is stationary if all probability statements about it in any region $B \subset D$ are invariant under arbitrary translations of B.

- If the process is stationary, then $\lambda(\mathbf{s}) \equiv \lambda$ for all $\mathbf{s}$, where λ, the *intensity*, is the mean number of events per unit area. Also, $C(\mathbf{s}, \mathbf{t}) = C(\mathbf{s} - \mathbf{t})$ for all $\mathbf{s}, \mathbf{t} \in D$.

Isotropy

- A process is isotropic if the invariance required for stationarity holds under rotation as well as translation.

- If the process is isotropic, then $C(\mathbf{s},\mathbf{t}) = C(h)$, where $h = [(\mathbf{s}-\mathbf{t})'(\mathbf{s}-\mathbf{t})]^{1/2}$, for all $\mathbf{s},\mathbf{t} \in D$.

For an isotropic (hence stationary) SPP, three additional descriptive functions are useful:

- Nearest-neighbor distance function, $G(y)$

  Let Y denote the distance from an arbitrary event to its nearest neighbor. Then

  $$G(y) = P(Y \leq y).$$

- Point-to-nearest-event distance function, $F(x)$

  Let X denote the distance from an arbitrary point to the nearest event to that point. Then

  $$F(x) = P(X \leq x).$$

- The second-moment cumulative function, $K(h)$

  $$K(h) = \frac{1}{\lambda}E(\#\text{ of additional events within } h \text{ of a randomly chosen event})$$

  where λ is the intensity.

6. Homogeneous Poisson Process (HPP, CSR)

Defining characterization:

- For every $B \subset D$, $N(B)$ has a Poisson distribution with mean $\lambda|B|$ for some $\lambda > 0$.
- For any two disjoint regions B_1 and B_2, $N(B_1)$ and $N(B_2)$ are independent.

Implication:

- Conditional on $N(D) = n$, the n events are a random sample from a uniform distribution on D.

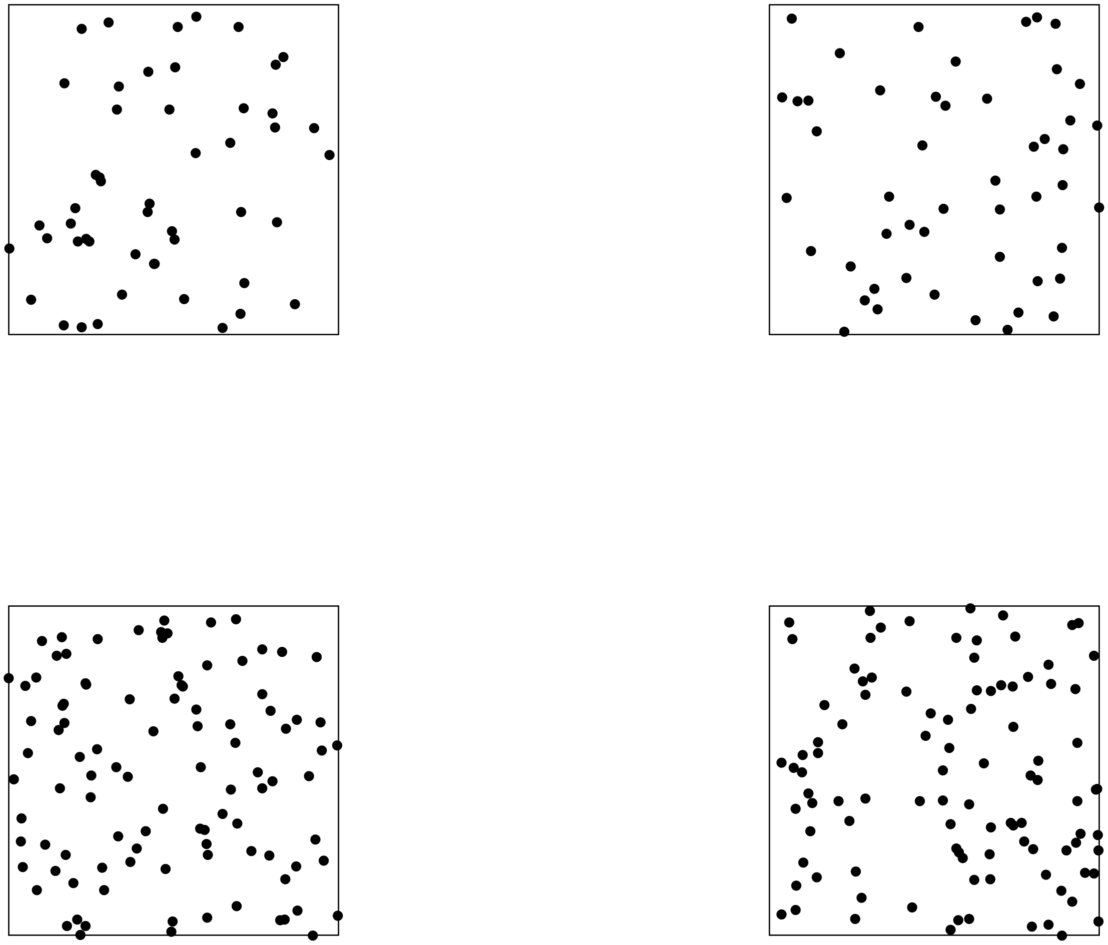
Attributes of HPP:

- Stationary and isotropic
- Intensity $= \lambda$
- $\lambda_2(\mathbf{s}, \mathbf{t}) = \lambda^2$
- $C(h) = 0$ for $h \neq 0$

Computer simulation (conditioned on n events in D):

- For the common case of a unit square, merely generate 2 independent uniform random variates, pair them up to get the coordinates of a single event, and repeat this independently n times.
- For the case of a rectangle, simply rescale one of the coordinates.
- For an irregularly shaped D, simulate on a rectangle containing D and retain only those events that lie within D.
- In the `spatstat` package of R, the `rpoispp()` function can be used to simulate a homogeneous Poisson process with given intensity λ on a rectangle, which is *not* conditioned on n events. The `runifpoint()` function yields a realization from the same process, but conditioned on n events. The next page gives some realizations of an HPP, conditioned on either 50 or 100 events on the unit square:

Realizations of a HPP, conditioned on $N = 50$ (top two plots), or $N = 100$ (bottom two plots):



7. Poisson Cluster Process (PCP)

The process generates clusters according to three rules:

1. Cluster centers form a HPP with intensity ρ.
2. The #'s of events in each cluster are iid variates with mean μ.
3. Positions of events within a cluster, relative to its center, are iid $\sim$ pdf $f(\cdot)$.

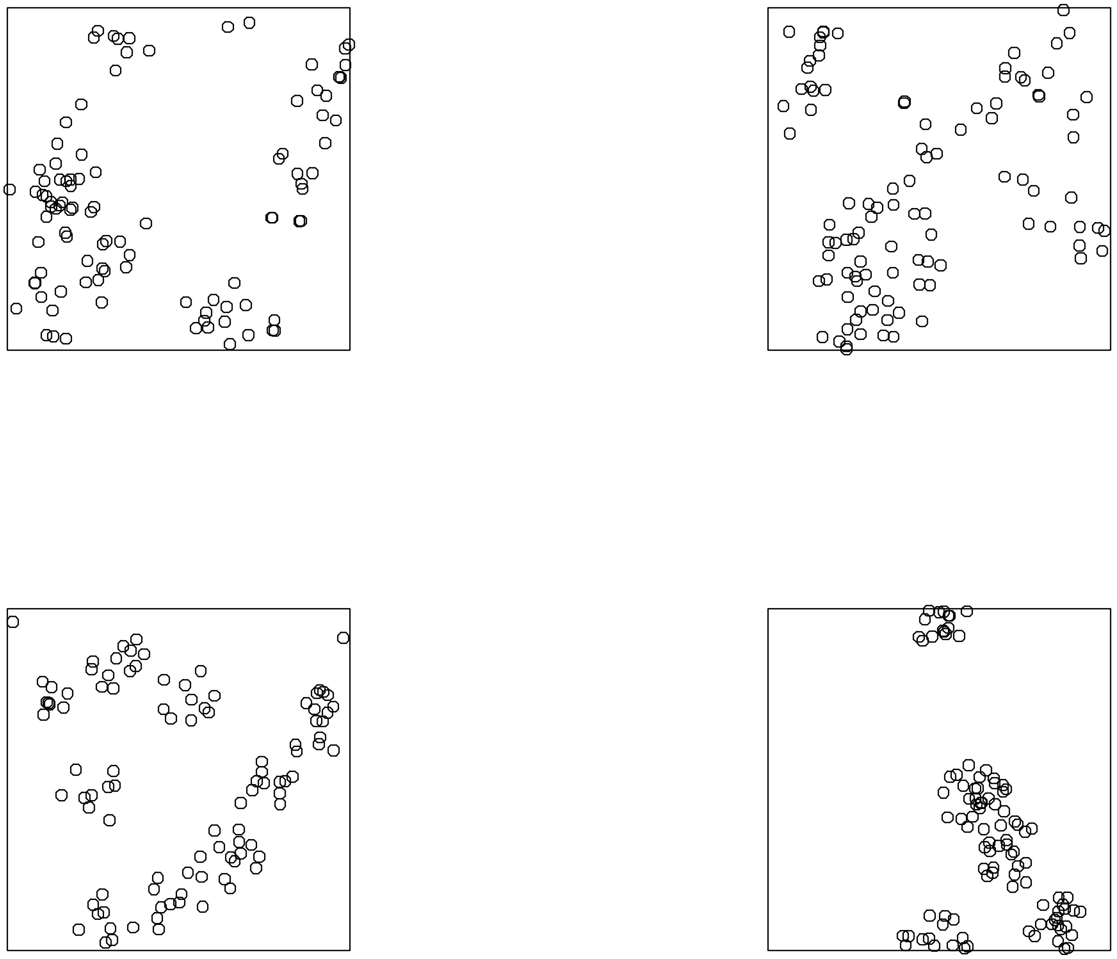
Attributes:

- Stationary, with intensity $\lambda = \rho\mu$.
- Isotropic $\Leftrightarrow$ $f(\cdot)$ is radially symmetric
- Can be shown that $C(\mathbf{s}-\mathbf{t}) = \rho E\left\{S(S-1)\right\} f^*(\mathbf{s}-\mathbf{t}) \geq 0$, where S is the number of events in an arbitrary cluster and $f^*(\cdot)$ is a probability density function.

Computer simulation (conditioned on n events in D):

- Straightforward if the #'s of events in each cluster are taken to be Poisson random variates, for in that case randomly allocating the n events among the clusters effectively conditions on N.
- The displacement of events from the cluster center may "go off the edge" if the cluster center is near the edge of D. Some type of edge correction may be necessary.
- The `rMatClust()` and `rThomas()` functions in the `spatstat` package can be used to simulate realizations of Poisson cluster processes. For `rMatClust`, the displacements have uniform distributions over a circle of specified radius; for `rThomas`, the displacements have an isotropic bivariate normal distribution with a specified common standard deviation (and zero correlation). Their edge corrections, if any exist, are unspecified.

Realizations of a PCP, conditioned on $N = 100$, with Poisson #'s of events within clusters, with $h(\cdot)$ equal to the uniform density on a circle of radius 0.10, and with toroidal edge correction:
Top two plots: $\rho = 20, \mu = 5$
Bottom two plots: $\rho = \mu = 10$



8. Simple Sequential Inhibition Process (SSIP)

- First event is uniformly distributed in D.
- The distribution of each subsequent event, conditional on all previously realized events, is uniform on that portion of D that lies no closer than δ to any previously realized event.

Attributes:

- Stationary and isotropic
- For any desired # of events n, δ cannot be too large or else it becomes impossible to add further events (related to maximum packing intensity). For a square study region D, the maximum permissible value of δ is

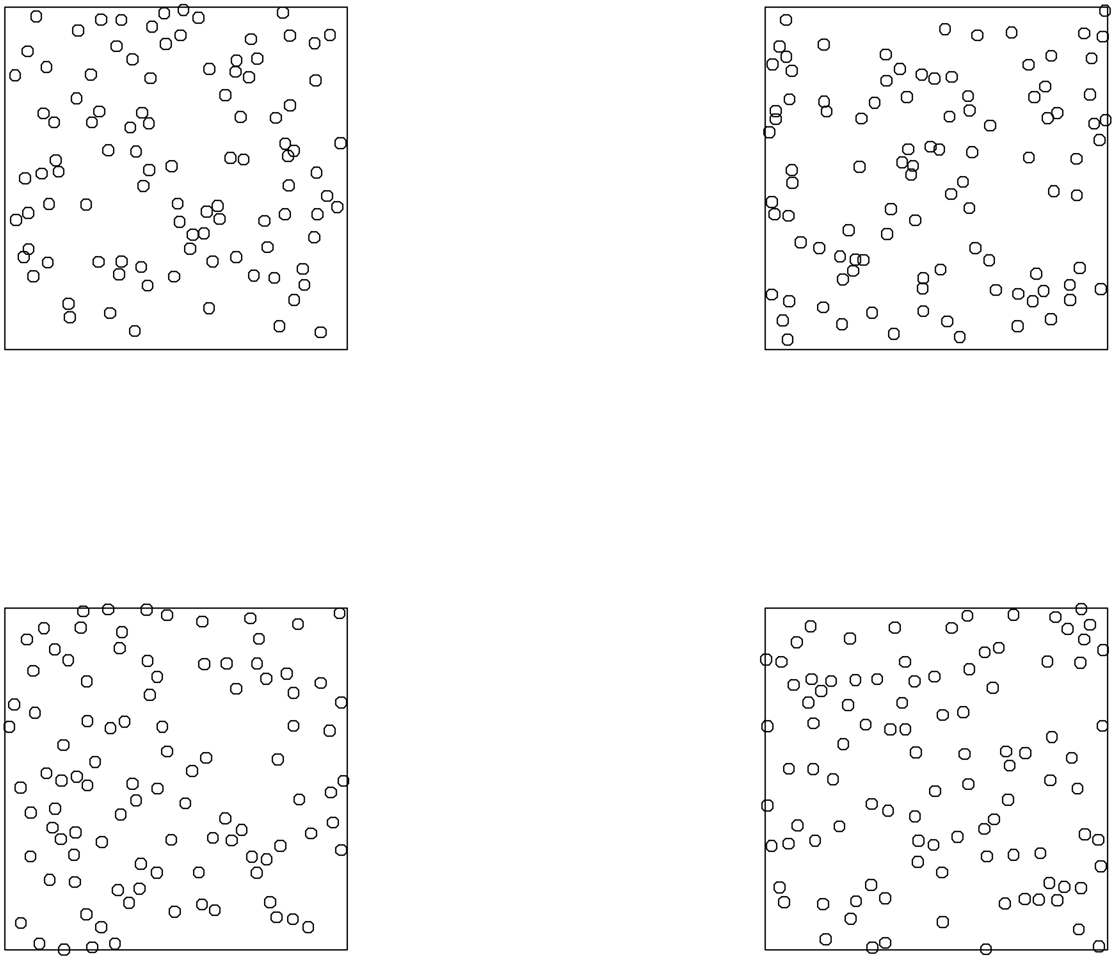
$$\sqrt{\frac{2|D|\sqrt{3}}{3n}}.$$

The above describes a *hard-core* process, meaning that absolutely no event is allowed within δ of any other event. *Soft-core* processes can also be defined, which allow events to be closer than δ units apart, but with small probability.

Computer simulation (conditional on n events):

- Simply generate events as for a HPP, but retain only those events that are no closer than δ to all previously generated events.
- Thus, you generally must generate many more than n events to retain n events.
- Make sure δ is not too large!
- The function `rSSI()` in `spatstat` can be used to simulate a realization from a simple sequential inhibition process.

Realizations of a SSIP, conditioned on $N = 100$:
Top two plots: $\delta = 0.02$
Bottom two plots: $\delta = 0.04$



9. Inhomogeneous Poisson Process (IPP)

This is a nonstationary process with non-constant intensity function $\lambda(\mathbf{s})$.

- Definition: For every $B \subset D$, $N(B) \sim$ Poisson with mean
$$\int_B \lambda(\mathbf{s}) d\mathbf{s},$$
and for any two disjoint regions B_1 and B_2, $N(B_1)$ and $N(B_2)$ are independent.

- Implication: Conditional on $N(D)$, the events are a random sample from a continuous distribution on D with pdf $\propto \lambda(\mathbf{s})$.

Remarks:

- A monotone or otherwise smooth $\lambda(\cdot)$ may be useful for accounting for global trends in intensity.

- The IPP provides a possible framework for the incorporation of q covariates, via an intensity function
$$\lambda(\mathbf{s}) \equiv \lambda(z_1(\mathbf{s}), \ldots, z_q(\mathbf{s})).$$

- A *Cox process* is obtained by first generating a realization $\lambda(\mathbf{s})$ from a nonnegative-valued random field $\{\Lambda(\mathbf{s}): \mathbf{s} \in R^2\}$ and then generating events from an IPP with intensity function $\lambda(\mathbf{s})$.

Computer simulation (conditioned on n events in D):

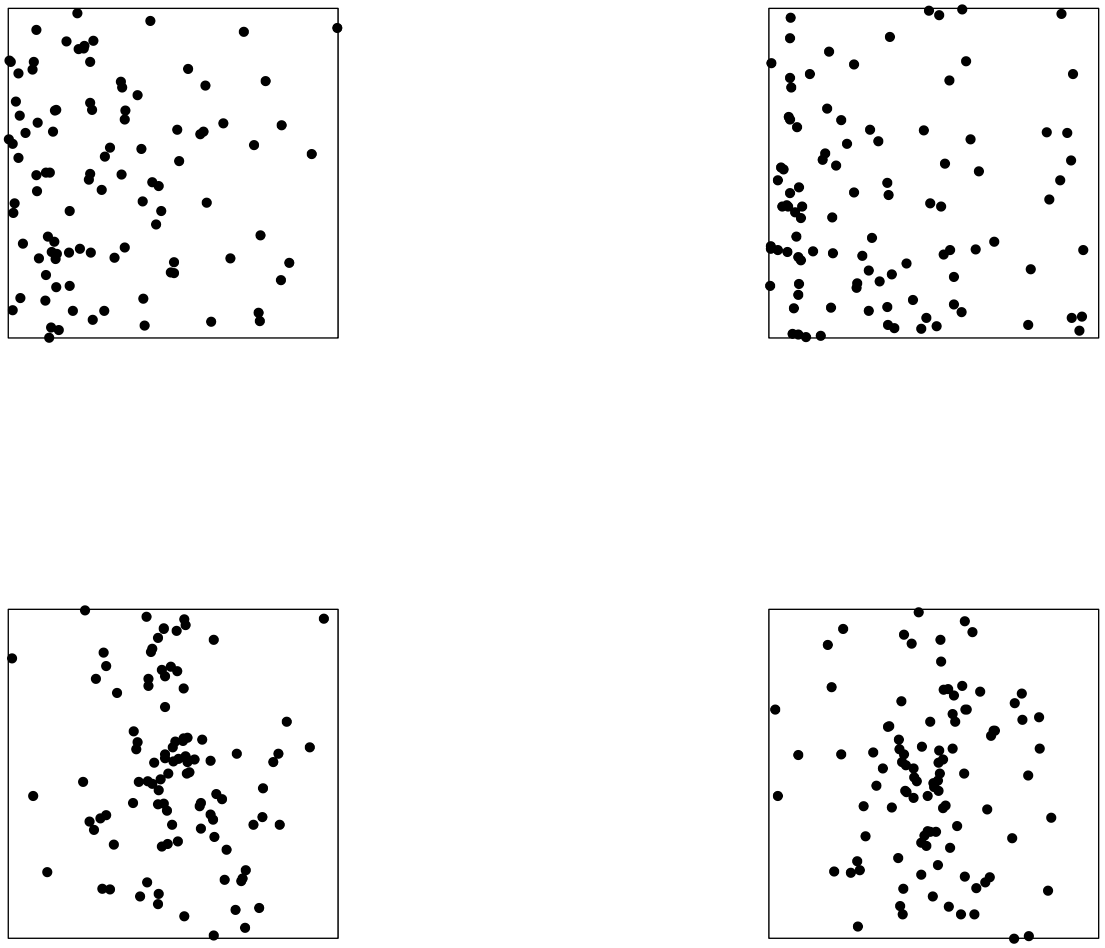
1. Generate an event from the uniform distribution on D. Call its coordinate vector $\mathbf{s}$.

2. Retain the event at $\mathbf{s}$ with probability $\lambda(\mathbf{s})/\lambda_0$, where $\lambda_0 \equiv \max_{\mathbf{s} \in D} \lambda(\mathbf{s})$. This is called "thinning."

3. Repeat steps 1 and 2 until n events have been retained.

The function `rpoispp()` is flexible enough to simulate realizations from an inhomogeneous Poisson process, but if you want the realizations to be conditioned on n events you must write your own procedure using the `runifpoint()` and `rthin()` functions.

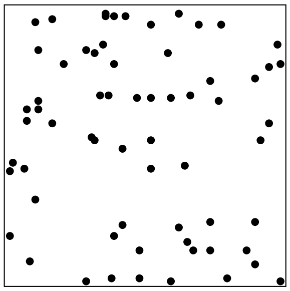
Realizations of an IPP, conditioned on $N = 100$:
Top two plots: $\lambda(x, y) = \exp(-2x - y)$
Bottom two plots: $\lambda(x, y) = \exp(-6|x - \frac{1}{2}| - 2|y - \frac{1}{2}|)$



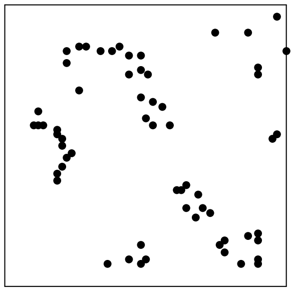
IX. INFERENCE FOR SPATIAL POINT PATTERNS

A. Testing Mapped Patterns for Complete Spatial Randomness

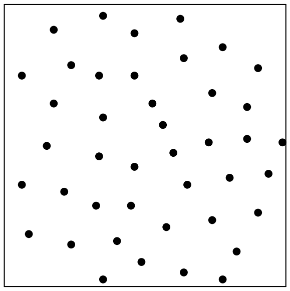
Some examples of mapped patterns:



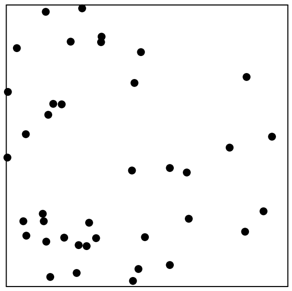
Pines



Redwoods



Cells



Rushes

1. *Areal Methods*

Let $N_1, \ldots, N_m$ denote the counts from a partitioning of D into m equally-sized quadrats. Write $\bar{N} = \sum N_i/m$ for the sample mean of the N_i's. Then compute the "index of dispersion,"

$$X^2 = \sum_{i=1}^{m} (N_i - \bar{N})^2/\bar{N}.$$

If the spatial point process is HPP, then the distribution of X^2 is, to a good approximation, χ^2_{m-1} (provided that $\bar{N}$ is not too small, say $\bar{N} \geq 5$).

Two interpretations of X^2:

1. Pearson's chi-square statistic, since $E(N_i | \sum_i N_i) = \bar{N}$ under the uniformity implied by CSR.
2. $\frac{(m-1)S^2}{\bar{N}}$, i.e. $(m-1)$ times the sample variance-to-mean ratio, which makes sense since the mean and variance of a Poisson distribution are equal.

The test is two-sided:

- X^2 too large indicates aggregation or heterogeneity
- X^2 too small indicates regularity

Analysis of Japanese black pines: Divide the study area into a 3×3 square grid of quadrats. Counts and result of test are as follows:

$X^2_8 = 8.80$, $P_{CSR}(2.733 < X^2_8 < 15.51) \doteq 0.90$, CSR not rejected.

Results for other patterns:

- Redwood seedlings: $X_8^2 = 23.54$, reject CSR.
- Cells: $X_8^2 = 5.57$, CSR not rejected.
- Scouring rushes: $X_8^2 = 19.40$, reject CSR.

Criticisms of areal methods:

- Insensitive to regular departures from CSR
- Conclusion can depend on quadrat size and shape, the choice of which is quite arbitrary. For example, if we repeat the procedure for the redwood data using a 2×2 grid, we obtain $X_3^2 = 5.56$, and we do not reject CSR ($P \doteq 0.14$).
- Too much information is lost by reducing the pattern to areal counts

Consequently, an analysis based on quadrats of merely one size and shape is not recommended for use with completely mapped patterns.

However, an analysis based on combining contiguous quadrats can be useful for characterizing pattern at different scales. That is:

1. Successively combine quadrats into 2×2, $4 \times 4, \ldots,$ blocks
2. Plot X^2 for each block size, vs. block size
3. Peaks or troughs in the plot may be interpreted as evidence of scales of pattern (aggregated or regular, respectively).

2. *Distance Methods*

(a) Clark-Evans test

- Based on the mean nearest-neighbor (NN) distance, $\bar{Y}$.
- Derivation of density of Y under CSR (ignoring edge effects):

- $\bar{Y}$ too small indicates aggregation (small-scale); $\bar{Y}$ too large indicates regularity (small-scale).
- Test statistic is

$$CE = \frac{\bar{Y} - \frac{1}{2\sqrt{\lambda}}}{\sqrt{\frac{4-\pi}{4\lambda\pi n}}},$$

  where $\lambda = n/|D|$.

- Quantity in numerator subtracted from $\bar{Y}$ is $E(Y)$ ignoring edge and overlap effects; quantity in denominator is standard error.
- Under CSR, and if edge and overlap effects are ignored, the distribution of CE is, to a fairly good approximation, $N(0,1)$.
- There are various fix-ups for edge and overlap effects. One fix-up is as follows:

$$\begin{aligned} E(\bar{Y}) &= 0.5\sqrt{\frac{|D|}{n}} + 0.0514\frac{l(D)}{n} + 0.041\frac{l(D)}{n^{3/2}}, \\ \text{var}(\bar{Y}) &= 0.0703\frac{|D|}{n^2} + 0.037\sqrt{\frac{|D|}{n^5}}l(D) \end{aligned}$$

  where $l(D)$ is the length of the study region's perimeter.

- Test tends to be powerful for detecting aggregation and regularity, weak at detecting heterogeneity.

Examples:

- Japanese black pines — $CE = -0.11$, accept CSR
- Redwood seedlings — $CE = -5.96$, emphatically reject CSR in favor of aggregation
- Biological cells — $CE = 8.33$, emphatically reject CSR in favor of regularity
- Scouring rushes — $CE = 1.04$, accept CSR

(b) Diggle's Refined NN analysis

- Motivation: CE will perform poorly when there are more large and small, but fewer intermediate, NN distances than expected under CSR but $\bar{Y}$ is still about $(2\sqrt{\lambda})^{-1}$. This possibility suggests that a test based on the entire empirical distribution function (EDF) of the NN distances may be more sensitive.
- Let
$$\hat{G}(y) = \frac{1}{n}\#(Y_i \leq y).$$
If CSR holds, $\hat{G}(y)$ should be "close" to $G(y) = 1 - \exp(-\lambda\pi y^2)$ for all $y > 0$, and a plot of $\hat{G}(y)$ vs. $G(y)$ should be nearly a straight line.
- $\hat{G}(y) > G(y)$ for small y indicates aggregation (at small scale)
- $\hat{G}(y) < G(y)$ for small y indicates regularity (at small scale)
- Measures of discrepancy between $\hat{G}(\cdot)$ and $G(\cdot)$:
  (a) $\Delta G = \max_y |\hat{G}(y) - G(y)|$ (Kolmogorov-Smirnov type)
  (b) $\int\{\hat{G}(y) - G(y)\}^2\, dy$ (Cramer-von Mises type)
- How do we judge significance? Because distribution theory for these statistics is too difficult, we use Monte Carlo testing. That is, we compare the measure's value for our data to the measure's values for s simulations (typically take $s = 99$ or 999) of an HPP.
- Because we don't know the true cdf G (due to edge and overlap effects), the use of
$$\bar{G}_i(y) = \frac{1}{s-1}\sum_{j\neq i}\hat{G}_j(y)$$
in place of $G(y)$ is recommended. That is, take
$$u_i = \max_y |\hat{G}_i(y) - \bar{G}_i(y)| \qquad (i = 1, \ldots, s).$$
- Alternatively, Koen (1990, *Biometrical Journal*) has tabulated the distribution of ΔG using simulation.

- Rather than reducing the EDF to a single summary statistic, it may be more informative to look at a plot of the EDF. If the SPP is consistent with CSR, then a plot of $\hat{G}(y)$ vs. $G(y)$ should be nearly a straight line from $(0,0)$ to $(1,1)$. Departures from CSR can be detected by means of *simulation envelopes*, whose upper and lower endpoints are defined as

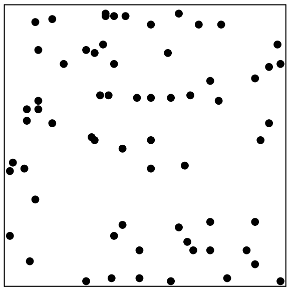
$$U(y) = \max_{i=1,\ldots,s}\{\hat{G}_i(y)\},\ L(y) = \min_{i=1,\ldots,s}\{\hat{G}_i(y)\}$$

  where s is the number of simulated HPP patterns having the same number of events (s is usually taken to be 99), and $\hat{G}_i(\cdot)$ is the NN-distance EDF for the ith simulation. For each $y > 0$,

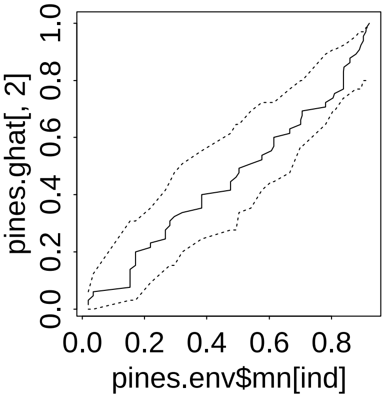
$$P[\hat{G}(y) > U(y)] = P[\hat{G}(y) < L(y)] = \frac{1}{s+1}.$$

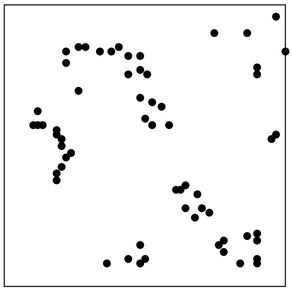
- Simulation envelopes also indicate the distance at which a deviation, if any, from CSR occurs.

Examples:

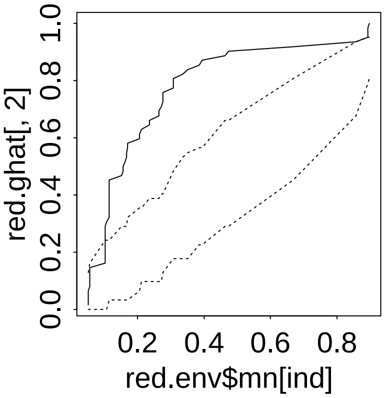


Pines





Redwoods

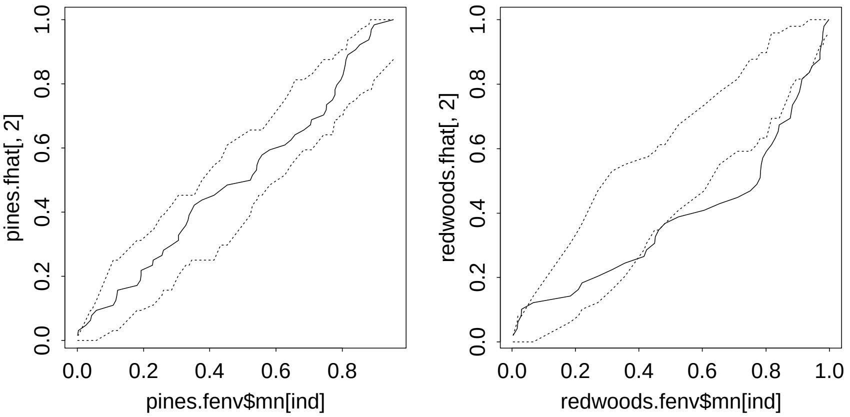


We can do precisely the same kinds of tests using the EDF of point-to-nearest event distances $X_1, \ldots, X_m$ from m randomly or systematically placed sample points. The important thing is that the sample points should be chosen without reference to the observed spatial point pattern.

- Let
$$\hat{F}(x) = \frac{1}{m} \#(X_i \leq x).$$
If CSR holds, $\hat{F}(x)$ should be close to $F(x) = 1 - \exp(-\lambda \pi x^2)$ for all $x > 0$, and a plot of $\hat{F}(x)$ vs. $F(x)$ should be nearly a straight line.
- $\hat{F}(x) < F(x)$ for small x indicates aggregation (small-scale)
- $\hat{F}(x) > F(x)$ for small x indicates regularity (small-scale)
- Again, Monte Carlo testing is needed to judge significance.

The use of both $\hat{G}(y)$ and $\hat{F}(x)$ is what Diggle called refined NN analysis.

Examples:



(c) Ripley's K-function approach

- Recall that the "K-function" (second-moment cumulative function) is defined as

$$K(h) = \frac{1}{\lambda} E(\# \text{ of additional events within } h \text{ of a randomly chosen event}).$$

- $K(h)$ combines distance measurement with areal counting, so we might expect it to contain more information than the NN distances and thus provide a more sensitive analysis.

- For a HPP, $K(h) = \pi h^2$ and

$$L(h) \equiv h - \left(\frac{K(h)}{\pi} \right)^{1/2} = 0.$$

- $L(h) < 0$ $(K(h) > \pi h^2)$ for small h indicates aggregation (small-scale)

- $L(h) > 0$ $(K(h) < \pi h^2)$ for small h indicates regularity (small-scale)

- Ripley proposes a nonparametric estimator $\hat{K}(h)$ of $K(h)$ (whose exact form we will not go into). He then suggests looking at the plot of $\hat{L}(h) \equiv h - \{\hat{K}(h)/\pi\}^{1/2}$ vs. h and computing a test statistic

$$L_{\max} = \max_{h<h_0} |\hat{L}(h)|.$$

  The upper bound h_0 is used to account for the scarcity of information about $K(h)$ at "large" distances.

- A Monte Carlo approach can be used to assess significance.

Examples:

- Japanese black pines — $P = .41$

- Redwood seedlings — $P < .01$

- Biological cells — $P < .01$

- Scouring rushes — $P = .06$

3. *Coordinate-based Goodness-of-Fit Methods* (for rectangular D only)

Under CSR the marginal distribution of the events' u-coordinates is uniform; likewise for the v-coordinates. Thus, we could use standard univariate goodness-of-fit tests for uniformity separately on each set of coordinates.

But a bivariate distribution with uniform marginals need not be uniform, so it would be much better to use an approach that takes account of how the u- and v-coordinates vary jointly.

Zimmerman's Modified Cramér-von Mises test (Zimmerman, 1993, *Applied Statistics*)

- Test is based on the discrepancy between the bivariate EDF of the (u, v) coordinates of events and the distribution of the same under CSR.
- Assume D is the unit square. Under CSR, $F(u, v) = uv$.
- Consider measuring discrepancy by Cramér-von Mises statistic

$$\omega^2 = n \int_0^1 \int_0^1 \{\hat{F}(u, v) - uv\}^2 \, du \, dv.$$

- Unfortunately, ω^2 is not invariant to which corner of the square we identify as the origin.
- So alternatively, measure discrepancy by $\bar{\omega}^2$, the average of the four Cramér-von Mises statistics corresponding to each of the four corners.
- Computing formula:

$$\begin{aligned}\bar{\omega}^2 &= \frac{1}{4n} \sum_{i=1}^{n} \sum_{i=1}^{n} (1 - |u_i - u_j|)(1 - |v_i - v_j|) \\ &\quad - \frac{1}{2} \sum_{i=1}^{n} (u_i^2 - u_i - \frac{1}{2})(v_i^2 - v_i - \frac{1}{2}) + \frac{1}{9} n\end{aligned}$$

- Test is two-sided; large values indicate aggregation or heterogeneity, small values indicate regularity.
- Percentiles of the distribution of $\bar{\omega}^2$ are given below; the values for $n = \infty$ seem valid for use with quite small samples.

$P(\bar{\omega}^2 \leq x)$	$n = 10$	$n = 20$	$n = \infty$
.01	.049	.046	.041
.05	.061	.059	.054
.10	.069	.067	.068
.25	.090	.088	.088
.50	.121	.121	.122
.75	.169	.170	.172
.90	.229	.232	.233
.95	.273	.280	.280
.99	.372	.392	.388

Advantages of $\bar{\omega}^2$:

- Easy to compute
- No simulation necessary
- No edge-effect or overlap-effect adjustment is necessary
- Powerful against heterogeneous alternatives
- Can test for heterogeneity (of certain kinds) in addition to testing for CSR

Limitations of $\bar{\omega}^2$:

- Requires rectangular D
- Weak against regularity and aggregation

Examples:

- Japanese black pines — $P = .67$
- Redwood seedlings — $P = .67$
- Biological cells — $P = .017$
- Scouring rushes — $P = .006$

4. *Comparisons of Tests*

Some tests for CSR are more powerful than others against specific alternatives. How powerful a test is against a specific alternative is affected by whether the test statistic is primarily a function of "local" or "global" characteristics.

- Tests based on distances to the nearest event emphasize local characteristics and thus do well against aggregation and regularity, but not against intensity trends (heterogeneity). The Clark-Evans test and Diggle's refined NN analysis are examples.
- Bivariate EDF tests (such as $\bar{\omega}^2$) and areal count-based tests emphasize global characteristics and are thus more powerful against large-scale heterogeneity but weaker against aggregation and regularity.
- Tests that combine distance measurement with areal counting, like the $L_{\max}$-test based on Ripley's K-function, give some weight to both local and global characteristics, and thus might be regarded as good all-purpose tests.

B. Modeling Completely Mapped Patterns

For completely mapped patterns, testing for CSR may be only the first step of an analysis. If CSR is rejected, we may then want to fit an alternative model (a PCP or SIP, for example) to the data, and assess the model's goodness-of-fit.

Methods used to fit models to patterns may be different for different models. For example, for some models maximum likelihood estimation is possible but for others the likelihood function is intractable or computationally burdensome to evaluate.

1. *The Inhomogeneous Poisson Process (IPP)*

(a) Maximum likelihood estimation

Consider a parametric family of intensity functions $\{\lambda_\theta(x,y)\colon\ \theta \in \Theta\}$. For this family, the likelihood function is proportional to

$$l(\theta; D) = \{\prod_{i=1}^{n} \lambda_\theta(x_i, y_i)\} \exp\{-\int_D \lambda_\theta(u,v)\,du\,dv\}.$$

A MLE of θ is a value $\hat{\theta}$ that maximizes $l(\theta; D)$.

Remarks:

- Usually, the likelihood equations do not yield an explicit solution, so numerical techniques are necessary (e.g. Newton-Raphson).
- A particularly useful family of intensity functions is

  $$\lambda(x,y;\theta) = \exp\{\theta' \mathbf{z}(x,y)\}$$

  where $\mathbf{z}(x,y)$ is a vector whose components may be values of concomitant environmental variables (e.g. elevation, soil moisture), known functions of the coordinates themselves, and/or distances to known environmental features (e.g. coastlines).
- Estimation of θ could help answer questions like "How much more likely is a particular plant species to occur at 2000m than at 3000m?"
- Special case of HPP:

(b) Nonparametric estimation

As an alternative to parametric estimation, nonparametric methods for multivariate density estimation can be applied to the problem of estimating $\lambda(\cdot)$.

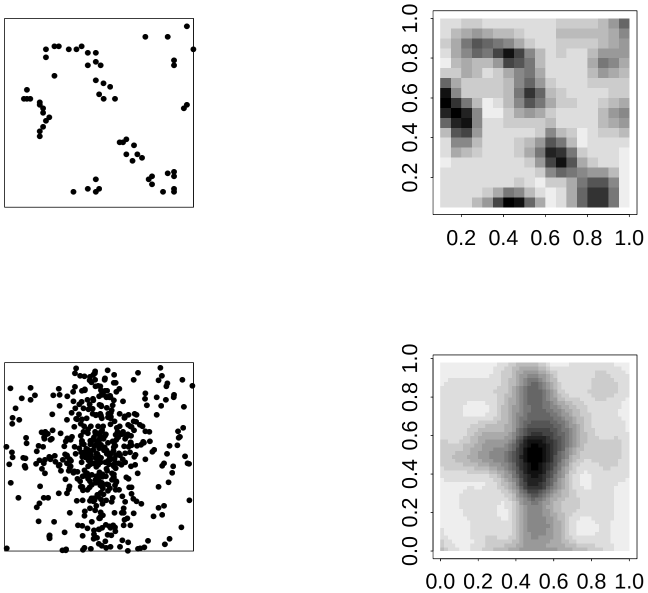
An edge-corrected kernel estimator for $\lambda(x, y)$ is given by

$$\hat{\lambda}_h(x, y) = \frac{1}{p_h(x, y)} \sum_{i=1}^{N(A)} \frac{1}{h^2} \kappa \left(\frac{\sqrt{(x - x_i)^2 + (y - y_i)^2}}{h} \right)$$

where $\kappa(\cdot)$ is a probability density (kernel) function symmetric about the origin, $h > 0$ is a bandwidth that determines the amount of smoothing, and $p_h(x, y)$ is an edge correction.

For more details on this topic, see Waller and Gotway (2004, section 5.2.5) or Silverman (1986), *Density Estimation*, Chapman and Hall, New York.

Plots of kernel intensity estimates:



2. *Stationary Processes (e.g. PCP, SIP)*

Let $\hat{K}(t)$ be a nonparametric estimator of $K(t)$, and suppose that we wish to fit a family of stationary models whose K-function is a known function of a parameter vector θ.

A (modified) least squares estimator for θ is obtained by minimizing

$$Q(\theta) = \int_0^{t_0} \{[\hat{K}(t)]^c - [K(t;\theta)]^c\}^2 \, dt$$

where c and t_0 are "tuning constants."

Remarks:

- In practice the integral must be discretized to a sum in order to evaluate it.
- c is used to control for heterogeneity of variance of $\hat{K}(t)$; $c = \frac{1}{4}$ has been suggested for aggregated patterns, and $c = \frac{1}{2}$ has been suggested for regular patterns.
- t_0 is used as an upper limit since the pattern supplies increasingly limited information as t increases.
- A defect of the criterion $Q(\theta)$ is that it does not take into account the strong statistical dependence between $\hat{K}(t_1)$ and $\hat{K}(t_2)$ at two distances t_1 and t_2.
- Goodness-of-fit can be assessed by comparing the minimized value, $Q(\hat{\theta})$, to values obtained by Monte Carlo simulation of the process in the family under consideration with parameter $\hat{\theta}$.
- In principle we could define similar estimators of θ using F or G, but their expressions either aren't known or aren't as simple as those of K for commonly used non-CSR processes.
- Some computable K-functions:
  - PCP with Poisson number of offspring per parent:

    $$K(t) = \pi t^2 + H(t)/\rho$$

    where $H(t)$ is a nonnegative-valued function.
  - Static inhibition process:

    $$K(t) = 2\pi \exp(2\pi\rho\delta^2) \int_\delta^t \exp\{-\rho U_\delta(x)\} x \, dx$$
- Bayesian modelling and estimation of such processes is more promising.

C. Testing Sparsely Sampled Patterns for Complete Spatial Randomness

Now we suppose that there were not sufficient resources to completely map the events. Rather, we suppose that the pattern was, in some manner, sampled. The sample may be a completely random sample (i.e. a realization of a uniform distribution on D) or systematic, but it must be taken completely independently of the observed events.

Advantages of systematic sampling:

- More practical in the field
- Can reduce the impact of edge and overlap effects (for distance methods)

Disadvantage of systematic sampling:

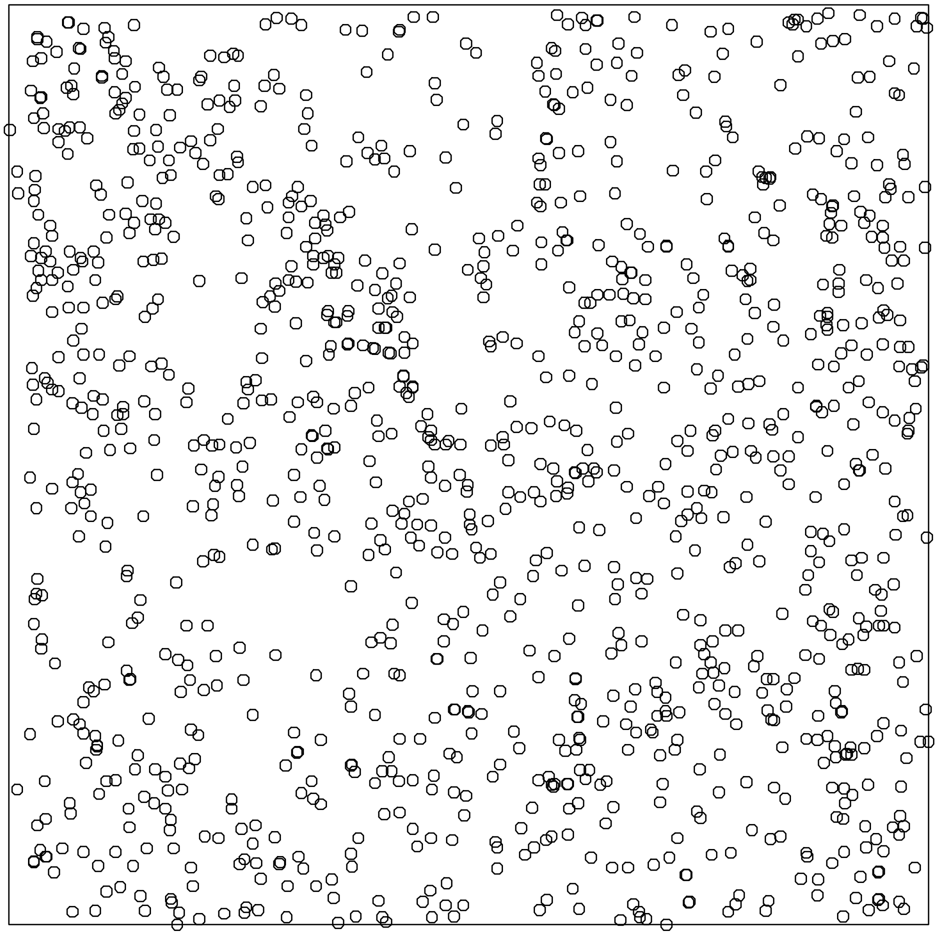
- The grid spacing may coincide with periodicities in the pattern

1. *Areal Methods*

- In this context $n_1, \ldots, n_m$ are counts from m non-overlapping equally-sized quadrats in D with relatively sparse coverage, rather than from a complete partition of D.
- Nevertheless, the test statistic and its limiting distribution under CSR are the same as in the completely mapped context:

$$X^2 = \frac{\sum_{i=1}^{m}(n_i - \bar{n})^2}{\bar{n}} \mathrel{\dot{\sim}} \chi^2_{m-1} \text{ under CSR.}$$

- Moreover, this approach is more competitive in this context (arguably it doesn't ignore as much information; non-uniqueness of quadrat size and shape is not a problem).
- Generally quite powerful against aggregation and heterogeneity, but weak against regularity (as was the case for completely mapped patterns).
- Example: Lansing woods data (next page)



2. *Distance Methods*

- Edge effects are still a concern; usually dealt with by sampling from an interior subregion of D.
- Overlap effects are not as serious in this context provided the sampling intensity is not too large; it is recommended that the # of samples be less than 10% of the # of events.
- Data, if it is to be regarded as a random sample, generally cannot be NN distances, as the act of randomly sampling a subset of events from which to measure NN distances implies a complete enumeration of events within D.

So consider methods based on m sample point-to-nearest-event distances $X_1, \ldots, X_m$.

- Recall that each X_i has cdf $F(x) = 1 - \exp(-\lambda\pi x^2)$ under CSR (ignoring edge effects).
- Let $U_i = \pi X_i^2$ be the circular search region obtained from point i. Then each $U_i \sim$ exponential(λ), or equivalently
$$2\lambda U_i \sim \chi^2_2.$$
- So if $X_1, \ldots, X_m$ are independent (as is the case if overlap effects are ignored), then
$$2\lambda \sum_{i=1}^{m} U_i \sim \chi^2_{2m}.$$
- Unfortunately, an exact test for CSR cannot be based on this result since we don't know N (or λ) in this context.

Suppose, for the sake of argument, that we <u>could</u> measure NN distances $Y_1, \ldots, Y_m$ from a randomly selected subset of m events. Then by the same kinds of arguments,

$$2\lambda \sum_{i=1}^{m} \pi Y_i^2 \sim \chi^2_{2m}.$$

If we ignore overlap effects, $2\lambda \sum_i \pi X_i^2$ and $2\lambda \sum_i \pi Y_i^2$ are independent under CSR, so the scale-free statistic

$$H \equiv \frac{2\lambda \sum_i \pi X_i^2 / 2m}{2\lambda \sum_i \pi Y_i^2 / 2m} = \frac{\sum_i X_i^2}{\sum_i Y_i^2} \sim F_{2m,2m}.$$

A test for CSR — Hopkins' test — consists of comparing H to $F_{2m,2m}$.

- H large ⇒ aggregation
- H small ⇒ regularity

But as noted previously, we cannot actually get a random sample of Y_i's, so Hopkins test isn't completely sound. Is there any alternative?

Consider T-square sampling:

- x_i = distance from sample point to nearest event
- z_i = distance from that nearest event to its NN within the half-plane "perpendicular" to the chord from the point to the nearest event
- Thus, the search area associated with Z_i is a semicircle, not a circle.
- The Z_i's are not a random sample of NN distances either, but they represent a reasonable attempt to deal with the problem of not being able to obtain such a random sample.
- By an argument similar to one already given,

$$\lambda\pi \sum_{i=1}^{m} Z_i^2 \sim \chi^2_{2m}.$$

- Thus we could test for CSR using

$$t = \frac{2\sum_i X_i^2}{\sum_i Z_i^2} \sim F_{2m,2m}.$$

At least 20 other distance-based tests for CSR have been proposed.

D. Estimation of Intensity (sparsely sampled data)

Assume that the spatial point pattern arises from a stationary process, so that the constant intensity parameter λ is well-defined.

1. Quadrat methods

Consider the problem of estimating N, or equivalently $\lambda = N/|D|$, from counts $n_1, \ldots, n_m$ from m sparsely placed quadrats, each of area a. An intuitively reasonable estimator of λ is

$$\tilde{\lambda} = \frac{\sum_{i=1}^{m} n_i}{ma}.$$

Remarks:

- $\tilde{\lambda}$ is the MLE under CSR

- $\tilde{\lambda}$ is unbiased, regardless of whether CSR holds

- Under CSR, $\text{var}(\tilde{\lambda}) = \lambda/ma$ and we can estimate this by substituting $\tilde{\lambda}$ for λ. Otherwise, however, it is recommended that we estimate $\text{var}(\tilde{\lambda})$ by s^2/ma, where s^2 is the sample variance of the n_i's.

- The MLE of N and its variance are easily obtained from these results

- Confidence intervals for λ or N are readily obtained from the Central Limit Theorem and these variance results. Example: under CSR, an approximate 95% confidence interval for λ is
$$\tilde{\lambda} \pm 1.96\sqrt{\frac{\tilde{\lambda}}{ma}}.$$

Example: Lansing woods data

2. Distance methods

Consider the problem of estimating λ from m point-to-nearest-event distances $X_1, \ldots, X_m$. Ignoring edge and overlap effects, $\pi X_i^2 \sim$ iid exponential(λ) under CSR, so the MLE under CSR is easily found to be

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

Remarks:

- $\hat{\lambda}$ can be interpreted as the reciprocal of the average area searched to find the nearest event
- Under CSR, $\hat{\lambda}$ is slightly biased; an unbiased (under CSR) estimator is $(1 - \frac{1}{m})\hat{\lambda}$
- An exact (under CSR) $100(1-\alpha)\%$ confidence interval for λ is:

$$\left(\hat{\lambda}\frac{\chi^2_{\alpha/2,2m}}{2m}, \hat{\lambda}\frac{\chi^2_{1-\alpha/2,2m}}{2m}\right)$$

- For aggregated patterns, $\hat{\lambda}$ is negatively biased
- For regular patterns, $\hat{\lambda}$ is positively biased

Again, for the sake of argument suppose that we had available a random sample of NN distances $Y_1, \ldots, Y_m$. The MLE for λ based on these distances would be $\bar{\lambda} \equiv m / \sum_i \pi Y_i^2$. It would also be biased under departures from CSR, but in opposite directions:

- For aggregated patterns, $\bar{\lambda}$ is positively biased
- For regular patterns, $\bar{\lambda}$ is negatively biased

We could get a maximum likelihood estimator from T-square measurements also.

Robust estimation

- The tendency for the MLE's based on the X_i's and the Y_i's to be biased in opposite directions under departures to CSR suggests that a more robust (with respect to this bias) estimator may result from combining the two estimators.
- Such estimators are called "compound estimators."
- Two compound estimators, based on T-square measurements (X_i, Z_i), are:

$$\hat{\lambda}_D = \frac{m\sqrt{2}}{\pi\sqrt{(\sum_i X_i^2)(\sum_i Z_i^2)}} \quad \text{(Diggle, 1975, \textit{Biometrika})}$$

$$\hat{\lambda}_B = \frac{m^2}{\sqrt{2}(\sum_i X_i)(\sum_i Z_i)} \quad \text{(Byth, 1982, \textit{Biometrics})}$$

- Simulation studies show that $\hat{\lambda}_B$ performs better than $\hat{\lambda}_D$.

Censoring (Zimmerman, 1991, *Biometrika*)

A final concept that can be useful for distance methods is that of censoring, i.e., imposing an upper limit on the radius of search for the nearest event.

- May be motivated by practical considerations such as reduction of sampling effort, or for the purpose of eliminating edge and overlap effects.
- Censored systematic sampling: Consider a square grid of spacing $2L$ overlaid upon D, with the restriction that the shortest distance from any grid point to the boundary is equal to L. Distance X_i from each point to the nearest event is measured, subject to a maximum allowable search distance L.

- Data are m pairs of measurements (x_i, δ_i), where
$$x_i = \min(X_i, L)$$
and δ_i is the indicator function for the set $\{X_i \leq L\}$.
- Let $U_i = \pi X_i^2$ and $u_i = \pi x_i^2$.
- Under CSR, the (u_i, δ_i)'s are iid with pdf
$$f(u, \delta) = \lambda^\delta e^{-\lambda u}$$
- It follows that the MLE of λ is
$$\hat{\lambda}_C = \frac{r}{\sum_{i=1}^m u_i},$$
where $r = \sum \delta_i$ is the number of uncensored distance measurements.

X. ANALYSIS OF MULTIVARIATE POINT PATTERNS

A. Terminology and Basic Concepts

So far we have considered situations in which the events are members of a single population. Now we consider *multivariate* spatial point patterns, for which each event can be classified into one of a finite number of categories. We shall consider only the bivariate case, where events are of only two types, e.g., oaks and maples, or adults and juveniles. We shall refer to the two types generically as Type 1 and Type 2.

There are four aspects of a bivariate pattern that may be of interest:

- The pattern of Type 1 events only
- The pattern of Type 2 events only
- The combined pattern of intermingled Type 1 and Type 2 events
- The interrelation between Type 1 and Type 2 events

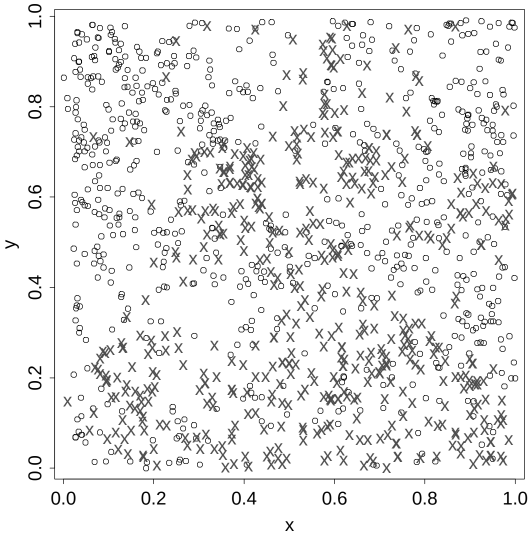
The first three of these refer to a single point pattern, so we could use methods for the analysis of univariate point patterns to classify and model them. If we consider the four-way classification given previously for each type of event, there are numerous possible combinations of the first three aspects.

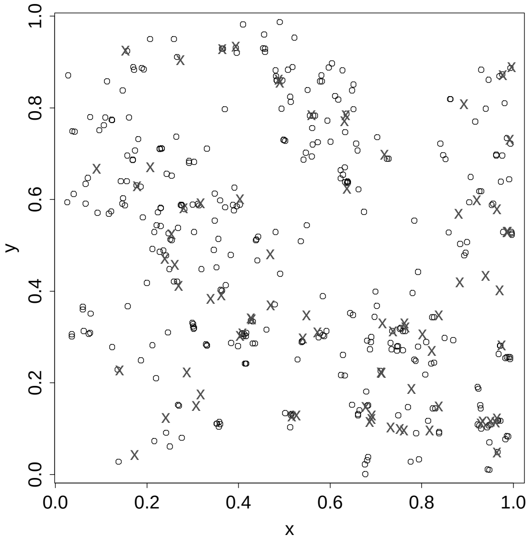
The fourth aspect, however, is something different. We can broadly classify it into three possibilities, according to how the locations of Type 1 events are associated with the locations of Type 2 events:

- Attraction (positive association, due e.g. to similar responses to environmental heterogeneity or to a mutualistic relationship)
- Independence (no association)
- Repulsion (negative association, due e.g. to complementary responses to environmental heterogeneity or to competition)

We shall consider methods for addressing this fourth aspect.

Examples: Lansing Woods data and bramble canes data





B. Models and Important Theoretical Quantities

Models

Models for bivariate point processes have typically been constructed by linking univariate models independently or with either positive or negative dependence. We briefly describe just two:

1. Bivariate Poisson process
    - Each component process is an HPP
    - Independence model is simply a superposition of the two HPP's
    - Positive dependence can be built in by, for example, displacing each Type 2 event randomly about a Type 1 event, according to a radially symmetric bivariate distribution with mode (0,0) [Type 1 event = Parent, Type 2 event = Offspring].
2. Mutual inhibition process
    - Events of each type are generated in an alternating sequence over D.
    - At each stage, the next event of a given type is realized from a uniform distribution over that portion of D that is at least distance δ away from any previously realized events of the *opposite* type.

Theoretical Quantities

Consider stationary, isotropic, and orderly processes only (both for the component processes separately and for their superposition). Then we have:

- Intensities: λ_1, λ_2 (defined as before, but for each process)
- Second-order intensity function:

$$\lambda_{ij}(u) = \lim_{|d\mathbf{x}||d\mathbf{y}| \to 0} \left\{ \frac{E[N_i(d\mathbf{x})N_j(d\mathbf{y})]}{|d\mathbf{x}||d\mathbf{y}|} \right\}$$

  where $u = \sqrt{(\mathbf{x}-\mathbf{y})'(\mathbf{x}-\mathbf{y})}$.
- Note that $\lambda_{12}(u) = \lambda_{21}(u)$.
- Second-moment cumulative functions:

$$K_{ij}(t) = \frac{1}{\lambda_j} E[\# \text{ of Type } j \text{ events within } t \text{ of an arbitrary Type } i \text{ event}]$$

  - $K_{12}(t)$ "large" $\Rightarrow$ Positive association of Type 2 with Type 1
  - $K_{12}(t)$ "small" $\Rightarrow$ Negative association of Type 2 with Type 1
  - Note that $K_{12}(t) = K_{21}(t)$.
- Cdf's:

$$\begin{aligned} G_{ij}(y) &= \text{cdf of distance to nearest Type } j \text{ event} \\ &\quad\ \text{from an arbitrary Type } i \text{ event} \end{aligned}$$

$$\begin{aligned} F_j(x) &= \text{cdf of distance to nearest Type } j \text{ event} \\ &\quad\ \text{from an arbitrary point} \end{aligned}$$

- Note that $G_{12}(y) \neq G_{21}(y)$ and $F_1(x) \neq F_2(x)$, in general.

Under independence (but regardless of whether CSR holds):

- $K_{12}(t) = K_{21}(t) = \pi t^2$ (the first equality holds without independence, as noted above)
- $F_1(x) = G_{21}(x)$
- $F_2(x) = G_{12}(x)$

C. Tests for Independence

1. Quadrat methods

Suppose we have quadrat count data $\{(n_{i1}, n_{i2}) : i = 1, \ldots, m\}$ where n_{ij} is the number of Type j events in quadrat i. The quadrats may constitute a complete partition of D or merely a sparse sample.

(a) Presence-Absence table [Greig-Smith (1964), *Quantitative Plant Ecology*]

Test for independence using

$$X^2 = \frac{m(|ad - bc| - \frac{1}{2}m)^2}{(a+b)(c+d)(a+c)(b+d)}.$$

For large m, and under independence, $X^2 \mathrel{\dot\sim} \chi^2_1$.

Test is two-sided, but we determine significance using only the right tail of the χ^2 distribution.

Potential problem: If the two types are present over only a small proportion of D, then the (absent, absent) cell may dominate and lead to a spurious conclusion of attraction.

(b) Correlation coefficient, r, between n_{i1}'s and n_{i2}'s.

- Motivated by the fact that the presence-absence table represents a severe reduction of the data — we ought to be able to make better use of quadrat counts.
- $r > 0 \Rightarrow$ Attraction
- $r < 0 \Rightarrow$ Repulsion
- We can test for significance using a standard test for zero correlation, i.e., compare $t \equiv \frac{r\sqrt{m-2}}{\sqrt{1-r^2}} \sim t_{m-2}$.
- The test can be affected by a lack of independence between neighboring quadrat counts.
- The problem of $(0,0)$-dominance occurs here as well.

2. Distance methods

(a) Nearest-neighbor table [Pielou (1961), *Journal of Ecology*, 49, pp. 255-269]

The patterns are reduced by determining the type of the NN from each of m events (some Type 1, others Type 2), and constructing the following 2×2 table.

Test for independence using an X^2 statistic given by the same formula as the X^2 statistic for the quadrat presence-absence table. Same asymptotic reference distribution too (χ_1^2).

Remarks:

- Can be used for completely mapped or sparsely sampled data
- Should only be used when each of the two types satisfies CSR, at least approximately.

(b) Comparison of point-to-nearest-event and NN distance distributions [Goodall (1965), *Journal of Ecology*, 53, pp. 197-210]

Based on the idea that under independence,

$$F_1(x) = G_{21}(x) \text{ and } F_2(x) = G_{12}(x).$$

Can quantify the discrepancy between the two distributions in either pair in various ways:

- Compare sample means of the distances, either using a t-test or a nonparametric alternative (Mann-Whitney two-sample rank sum test).
- Use a two-sample Kolmogorov-Smirnov type of statistic

  $$\max_x |\hat{F}_j(x) - \hat{G}_{ij}(x)| \text{ for } i = 1, 2; j = 3 - i.$$

  Note: Goodall gives some approximate 5% and 1% critical values for the test. Alternatively, the method of toroidal shifts could be used (see next page).

(c) Correlation of paired point-to-nearest-event distances [Diggle and Cox (1983), *Int. Stat. Rev.*, 51, pp. 11-23]

Based on data $\{(X_{i1}, X_{i2})\colon\ i = 1, \ldots, m\}$, the distances from arbitrary points in D to the nearest events of each type. Under independence, the correlation between the X_{i1}'s and the X_{i2}'s is zero.

Remarks:

- Positive correlation $\Rightarrow$ attraction
- Negative correlation $\Rightarrow$ repulsion
- The distances are not normally distributed so for small or moderate m, the use of Spearman's rank correlation or Kendall's tau is advised.

(d) Use of bivariate K-functions [Lotwick and Silverman (1982), *Journal of the Royal Statistical Society-B*, 44, pp. 406-413]

Recall that $K_{12}(t) = K_{21}(t) = \pi t^2$ if the two processes are independent. Thus, we could base a test for independence on a plot of

$$\hat{L}_{12}(t) = \sqrt{\frac{\hat{K}_{12}(t)}{\pi}} - t$$

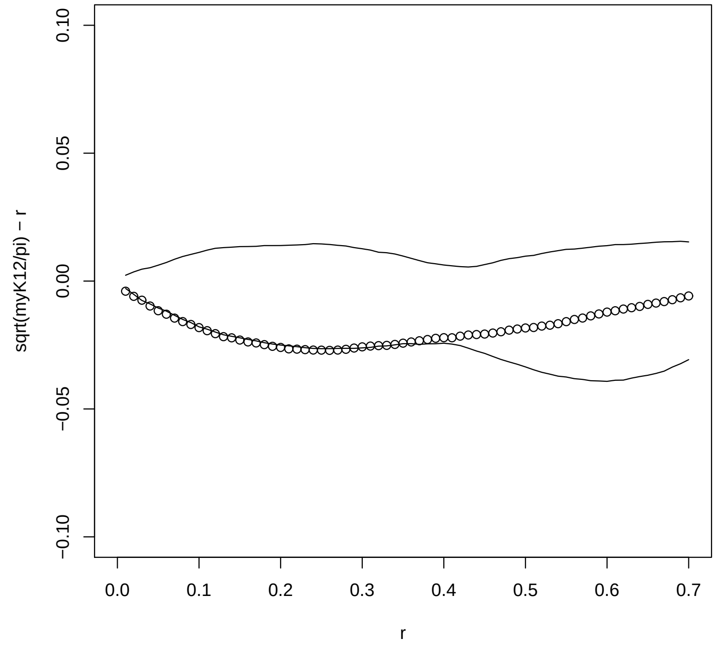
versus t, or on, say, the maximum of $|\hat{L}_{12}(t)|$ for $t \leq t_0$.

Remarks:

- $\hat{L}_{12}(t) < 0 \Rightarrow$ repulsion at that t
- $\hat{L}_{12}(t) > 0 \Rightarrow$ attraction at that t
- An estimator of $K_{12}(t)$ is proposed by Lotwick and Silverman.
- To assess significance, use the method of *toroidal shifts*, a Monte Carlo testing approach which preserves the observed patterns of each type separately. Realizations of "new" patterns are obtained by perturbing each event of one type (but not the other) a random amount $(\Delta x, \Delta y)$, using toroidal edge correction if necessary. (Assumes rectangular D).
    1. Hold the Type 1 pattern in place, but shift the entire Type 2 pattern from its original locations by a random amount $(\Delta x, \Delta y)$, using toroidal edge correction if necessary.
    2. Recalculate $\hat{L}_{12}(t)$.
    3. Repeat steps 1 and 2 a large number, say s, of times to get a simulation envelope or empirical distribution for assessing significance.
- See next page for an example, with code.

3. Comparisons of Tests

No thorough power studies have been done (a good research area!).



Splancs code within R for implementing the method of toroidal shifts on the Lansing Woods hickory and maple data:

```
hick <- lansing[lansing$marks=="hickory",]
hick <- as.points(cbind(hick$x,hick$y))
maple <- lansing[lansing$marks=="maple",]
maple <- as.points(cbind(maple$x,maple$y))
r <- 1:70/100
poly <- cbind(c(0,0,1,1),c(0,1,1,0))
myK12 <- k12hat(hick,maple,poly,r)
plot(r,sqrt(myK12/pi)-r,xlim=c(0,0.7),ylim=c(-.1,.1))
my12env <- Kenv.tor(hick,maple,poly,99,r)
lines(r,sqrt(my12env$upper/pi)-r)
lines(r,sqrt(my12env$lower/pi)-r)
dev.print()
```

D. Tests for Random Labelling

Random labelling versus independence:

- Random labelling hypothesis
  - Given the "population" of $N_1 + N_2$ events of two types, the N_1 Type 1 events are a random sample from this population.
  - Conceptual framework: locations are determined by a univariate point process, and then types are determined by a second random mechanism that operates independently of the point process.
  - Biological example: Colonization by seedlings of a single species, followed by transmission of disease between seedlings, with the end result that some seedlings are diseased and some are not.
- Independence
  - Events of Type 2 are located independently of events of Type 1.
  - Conceptual framework: locations of two univariate point processes are determined completely independent of one another.
  - Biological example: Simultaneous colonization of a region by two different plant species.

Random labelling and independence are <u>not</u> equivalent, in general:

- Random labelling does not imply independence; for example,

- Independence does not imply random labelling; for example,

- Random labelling plus two HPP's $\Leftrightarrow$ Independence plus two HPP's

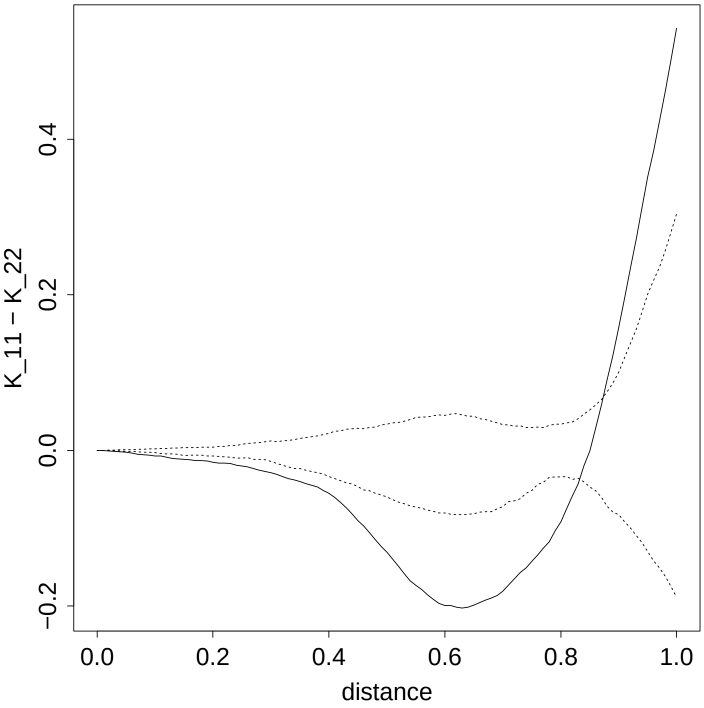
Key Result: Under random labelling,

$$K_{11}(t) = K_{22}(t) = K_{12}(t)$$

Measures of departure from (i.e. test statistics for) the random labelling hypothesis:

- $T = \#$ of Type 1 events which have other Type 1 events as their NN's (Cuzick and Edwards, 1990, *JRSS-B*, 52, 73-104)

- $\hat{D}(t) = \hat{K}_{11}(t) - \hat{K}_{22}(t)$ or $\hat{D} = \sum_{q=1}^{m} \hat{D}(t_q)/\sqrt{\text{var}(\hat{D}(t_q))}$ (Diggle and Chetwynd, 1991, *Biometrics*, 47, 1155-1163), where $t_1, \dots, t_q$ are regularly spaced distances.
- $\hat{D}_1(t) = \hat{K}_{11}(t) - \hat{K}_{12}(t)$ or $\hat{D}_2(t) = \hat{K}_{22}(t) - \hat{K}_{12}(t)$ (Dixon, 1994, unpublished report)
- Others, e.g. $\hat{K}_{11}(t)/\hat{K}_{22}(t)$, $\hat{D}(t)/\hat{K}_{22}(t), \dots$
- To evaluate significance, can use a normal approximation for numerical quantities like T and $\hat{D}$, making use of formulas for means and variances of these quantities supplied by the appropriate authors. Alternatively, can use a Monte Carlo approach, in which each simulated pattern is obtained by permuting event labels completely at random.
- An example, with code, is given on the following page.



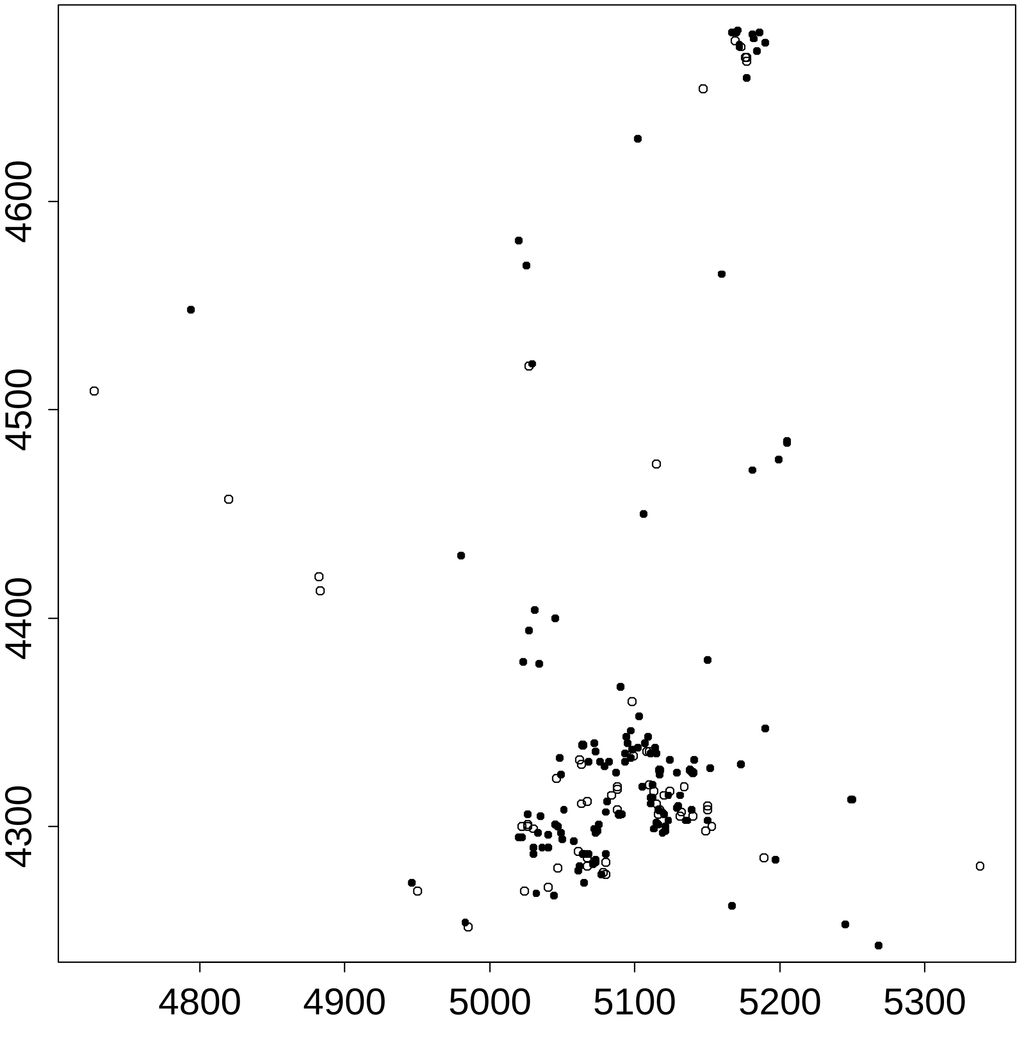
Splancs code within S+ for implementing a Monte Carlo test for random labelling on the Lansing Woods hickory and maple data:

```
unitsquare <- spoints(c(0,0,1,0,1,1,0,1))
hick.khat <- khat(hick.spp,unitsquare,seq(0,1,.01))
maple.khat <- khat(maple.spp,unitsquare,seq(0,1,.01))
khat.diff <- hick.khat-maple.khat
plot(seq(0,1,.01),khat.diff,xlab="distance",ylab="K_11 - K_22",type="l")
diff.lab <- Kenv.label(hick.spp,maple.spp,unitsquare,nsim=99,seq(0,1,.01))
lines(seq(0,1,.01),diff.lab$upper,lty=2)
lines(seq(0,1,.01),diff.lab$lower,lty=2)
```

E. Three Other Important Hypothesis Tests

1. Testing for spatial clustering of diseases

   - Background population of interest typically has inhomogeneous intensity, so the pattern will inevitably be clustered within the study area, and the methods we've learned so far will not be directly applicable.
   - Example: North Humberside leukemia data. Data consist of 62 cases of childhood leukemia and lymphoma in North Humberside, UK from 1974-1986; and 141 controls selected at random from entries on the birth register for January 7 or June 7 for each of the years 1974-1986.

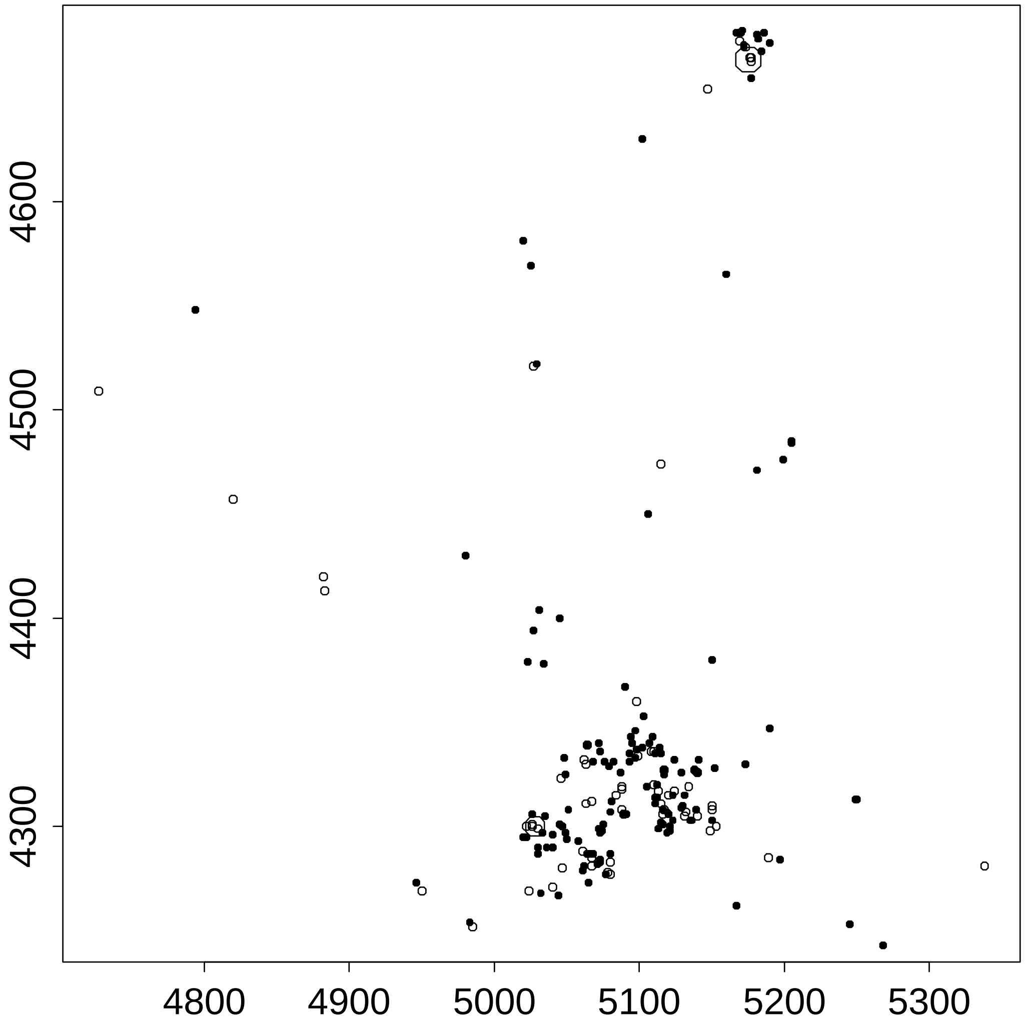


- One approach to testing for clustering: First ascertain locations of "cases" (Type 1 events); then randomly sample "controls" (Type 2 events) from the population of interest. From each case, determine whether the nearest other event is a case or a control. Let $T = \#$ of cases which have other cases as their NN's. Evaluate significance using permutation distribution of T. If T is too "large," then there is spatial clustering of the disease over and above that attributable to environmental inhomogeneity.
- More power to detect larger clusters might result if we measure not only NN's but also second NN's, third NN's, etc. I.e., $T_1 \equiv T$, $T_2 = \#$ of cases which have other cases as their second NN's, etc.
- For the North Humberside data, the permutation-based p-values for T_1, T_2, and T_3 were 0.055, 0.006, and 0.003, respectively.

2. Testing for clusters, i.e. testing whether a specific subset of events are clustered (rather than testing whether events generally tend to occur near other events)

   - A very popular approach for testing for clusters is based on the scan statistic.
   - Move a circular window across the study area and compare the ratio of cases to controls inside the circle to that observed outside the window.
   - In order to capture different potential cluster sizes, we consider a range of radii for the circles (perhaps ranging from the minimum interevent distance to half the length of the study region).
   - The test statistic is the maximized case/control ratio, and the cluster corresponding to this ratio is the "most likely cluster."
   - A permutation-based approach can be used to assess significance.
   - A software package, SaTScan, is available for this kind of analysis.
   - For the North Humberside data, the observed case/control ratio (over a range of cluster sizes) is 3.274, i.e. over a tripling of disease risk. (See next page for the locations of the corresponding two most likely clusters.) However, based on 999 random permutations, the p-value is only 0.691. Thus, approximately 69% of the permutations resulted in most likely clusters with higher case/control ratios than that seen in the original data. Thus, the most likely cluster is not a statistically unusual cluster of cases.
   - How do we reconcile this result with the results from testing for clustering?

North Humberside data, with most likely clusters indicated:



3. Testing for clustering around prespecified points (nuclear reactors, toxic waste dumps, etc.)

   - One simple approach is to measure the distance from each of the cases to the nearest putative hazard site, measure the distance from each of the same number of controls (selected randomly from the at-risk population) to the nearest putative hazard site, and compare the mean distance for cases to the mean distance for controls.