Spatial Statistics

PM569 Lecture 2: Geostatistical Data 1

Meredith Franklin

Division of Biostatistics, University of Southern California

September 7th, 2018

Geostatistical Data

In this lecture we will cover:

- Mapping geostatistical data
- Exploratory tools for geostatistical data
- Definition of distance and projections
- The semivariogram
- Stationarity
- Empirical semivariograms
- Theoretical semivariograms

The R code that accompanies this lecture is Lecture2Geostats.R

Geostatistical Data: Description

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

Examples:

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

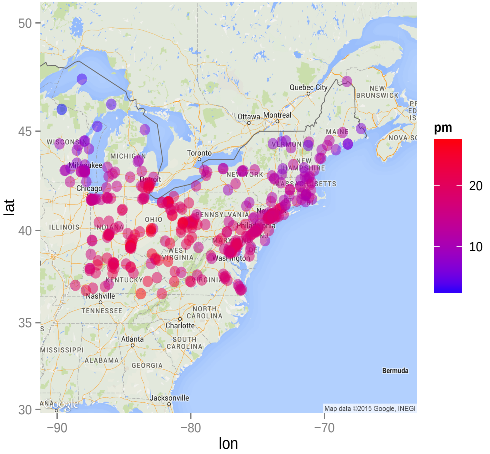
The common thread that links the data is a random process (also called stochastic process or random field)

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

where D is a domain in $\mathbb{R}^d$ (d typically 2)

Geostatistical Data: Example

Recall our example of $PM_{2.5}$ in the northeast US. Here we have each point representing a location (latitude, longitude), and an associated Z(s) which is monthly $PM_{2.5}$ concentrations.



Geostatistical Data: Analytical Goals

Goals of spatial statistics applied to point referenced data

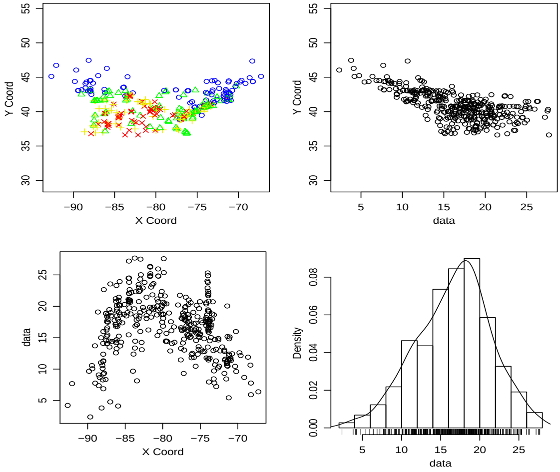
- Visualization of points on a map to look at distribution. Add colour scale to represent Z(s) values.
- Exploring the data to determine if there is a spatial pattern in the observations. (Often called spatial "structure")
- Testing null hypothesis of no spatial structure.
- Modeling the spatial correlation/covariance in the observations.
- Making predictions at unobserved locations: interpolation, smoothing.
- Accounting for spatial structure in regression models.

Exploratory Data Analysis

- Exploratory analysis is critical in applied statistics. We want to know what are the spread and distribution of the data, outliers, and in spatial statistics outlying locations
- Might need to transform since many methods rely on normality/symmetry of the data.
- As in regression, assumptions are generally based on residuals and not on original data.
- Log and square-root transformations are most common, primarily used to deal with skewed and non-negative data. However, there are sometimes issues of interpretation in moving back to the original scale if this is required for the analysis. One effect can be underestimation of peaks after back-transformation.

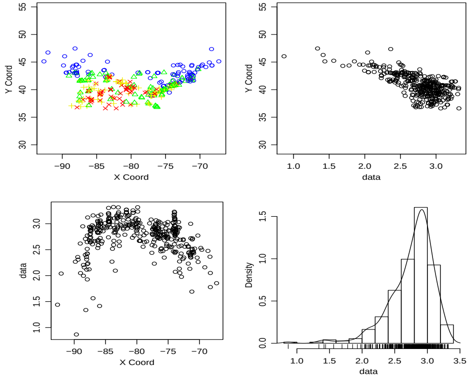
Geostatistical Data: Exploratory Analysis

Using the geoR library, we can do quick explorations of spatial data. Here we show the $PM_{2.5}$ concentrations, the $PM_{2.5}$ trend in each x (longitude) and y (latitude) directions, and a histogram of the $PM_{2.5}$ concentrations.



Geostatistical Data: Exploratory Analysis

Logging the $PM_{2.5}$ concentrations, we see that the map (top left) remains the same, but that the trend plots and histogram are different.



In this case, it looks like the untransformed data are better (less skewed).

Geostatistical Data: Statistical Definition

Statistical formulation

- Spatial pattern as a random process: $Z(\mathbf{s}) : \mathbf{s} \in D$ where the spatial domain D is fixed (e.g. Northeast US) and $\mathbf{s}$ are the spatial locations $s_1, s_2, ..., s_n$ in D (e.g. GPS locations of $\text{PM}_{2.5}$ monitor). The process is the collection of random variables $Z(\mathbf{s})$ (e.g. $Z(s_1) = PM_{2.5}$ concentration at location s_1)
- Since an infinite number of measurements could have been taken over the domain, D we think of the spatial locations we have measured $Z(\mathbf{s})$ as a realization of the random process

Geostatistical Data: Distance

- Understanding spatial structure based on covariance/correlation is often more difficult than understanding temporal structure (time series).
- How is distance defined?
- Irregular spacing leads to few (usually one) pair of points for a given distance.
- In time, there is only one dimension, but in space there are two, so direction can matter as well as distance.

Geostatistical Data: Covariance

- Covariance tells us whether knowing one observation gives us any information about another observation.
- Covariance allows borrowing of strength for local prediction and estimation, usually increasing efficiency (reduced uncertainty).

Definition of covariance and correlation between two variables

$$\text{Cov}(X_1, X_2) = E[(X_1 - E(X_1))(X_2 - E(X_2))]$$
$$\text{Corr}(X_1, X_2) = \frac{\text{Cov}(X_1, X_2)}{\sqrt{\text{Var}(X_1)\text{Var}(X_2)}}$$

- In the spatial setting, **covariance is a function of distance**.

Geostatistical Data: Covariance

Covariance and correlation for random processes are often called autocovariance and autocorrelation

$$C(h) = E[(Z(s) - E(Z(s))) \cdot (Z(s+h) - E(Z(s+h)))]$$
$$\text{Cov}(Z(s_i), Z(s_j)) = C(s_i - s_j) = C(h)$$

The covariance depends only on the distance, h, between locations s_i and s_j, not on the locations themselves. We will revisit this assumption later.

$$\rho(h) = \frac{C(h)}{C(0)} = \frac{C(h)}{Var(Z(s))}$$

Geostatistical Data: Distance

- How is distance, h, defined?
- Several ways to define spatial distances but they must satisfy several technical conditions
  1. symmetry $(d(s_i, s_j) = d(s_j, s_i))$
  2. the distance between a point and itself is zero
  3. the triangle inequality $(d(s_i, s_j) \leq d(s_i) + d(s_i))$
- Euclidean distance
  $d(s_i, s_j) = \sqrt{(s_{ix} - s_{jx})^2 + (s_{iy} - s_{jy})^2}$

Euclidean distance is by far the most common way to represent distance in spatial analysis.

Geostatistical Data: Distance

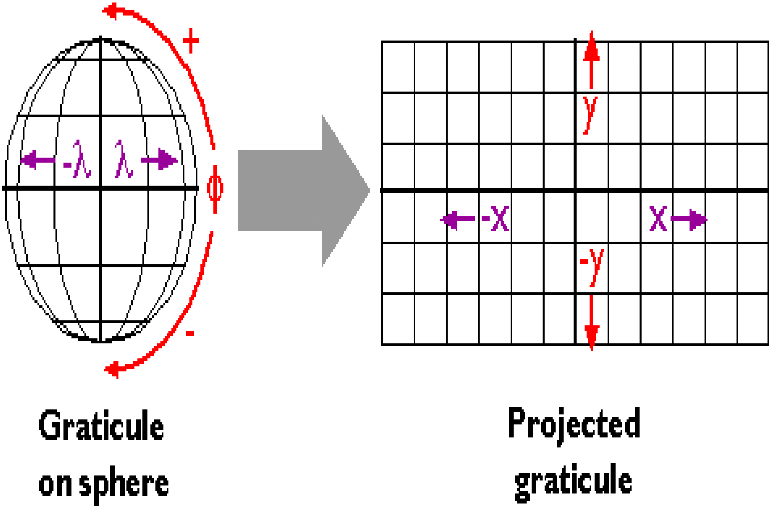
There are other ways to measure distance:

- Manhattan distance: functions on a fixed grid rather than shortest path
- Great circle distance: shortest path on a sphere, good and sometimes necessary for measuring large distances. Requires spherical geometry.
- Distance along a path: must follow roads, paths, sidewalks. Uses euclidean distance, but in segments. Can be calculated using GIS.

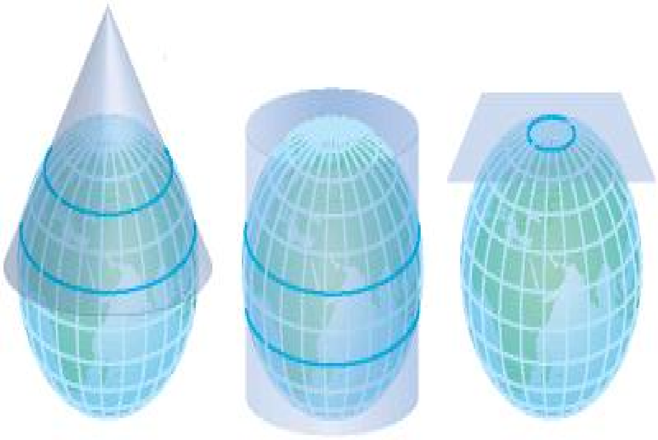
Geostatistical Data: Projections

- To represent (Euclidean) distances between spatial locations meaningfully, we need cartographic projections.
- The purpose of a projection is to convert geographic coordinates from the sphere (latitude, longitude) to a plane (x,y).
- Since the sphere cannot be directly flattened to a plane without distortion, and intermediate step is taken using a "developable surface" including a cone or cylinder.
- The projections themselves are complex differential equations, but thankfully all of the difficult mathematics has been done for us. In practice we choose a particular projection predefined for the local area or region of study.
- A local projection preserves distances. Examples: the UTM zone system, state plane system, albers equal area

Geostatistical Data: Projections



Geostatistical Data: Projections



Geostatistical Data: Projections

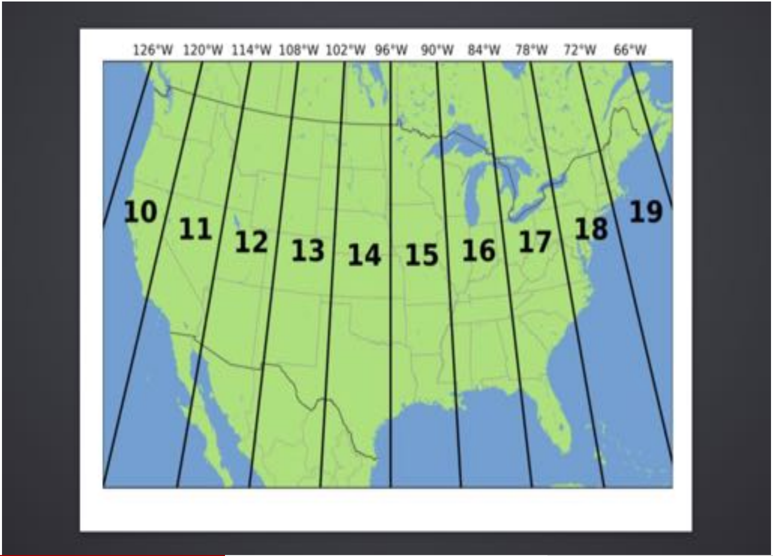
Common projections that we use are:

- Albers equal area (conic): must center at the correct location
- UTM (cylindrical): must pick the correct zone and hemisphere
- State Plane (conic or cylindrical): each state has its own system

A nice source of information about projections for R users can be found here: `www.nceas.ucsb.edu/~frazier/RSpatialGuides/OverviewCoordinateReferenceSystems.pdf`

Geostatistical Data: Projections

The UTM zones for the US



Geostatistical Data: Semivariogram

- The most widely used quantification of spatial autocorrelation is the **semivariogram**.
- It measures the similarity of values as a function of the distance between their locations.
- Traditional geostatisticians tend to favor the semivariogram over the covariogram/correlogram for historical reasons and because the empirical semivariogram is an unbiased estimator of the true semivariogram, while the covariogram is biased.

Geostatistical Data: Semivariogram

- $\text{Var}[Z(s+h) - Z(s)] = E[(Z(s+h) - Z(s))^2]$
- This is the expected squared difference between values, which generally increases as a function of the distance between the locations.
- $\text{Var}[Z(s+h) - Z(s)] = 2\gamma((s+h) - s) = 2\gamma(h)$
- $2\gamma(h)$ is the variogram and $\gamma(h)$ is the semivariogram

Geostatistical Data: Stationarity

Some additional properties of the semivariogram

- An assumption that is made in spatial analysis is that the spatial process under study repeats itself over the domain D. Such a spatial process is said to be stationary. For a stationary process the absolute coordinates at which we observe the process are unimportant. All that matters are the orientated distances between the points. In a stationary process if we translate the entire set of coordinates by a specific amount in a specified direction, the entire process remains the same.

Geostatistical Data: Stationarity

- It is useful to view spatial data as multivariate, despite it being the same measurement (i.e. $\text{PM}_{2.5}$ concentration) at multiple locations
- We have a joint probability density: $F(Z(\mathbf{s})) = P(Z(s_1) \leq z_1, Z(s_2) \leq z_2, ..., Z(s_n) \leq z_n)$
- Strong stationarity means that the joint density is invariant under translation: $P(Z(s_1) \leq z_1, Z(s_2) \leq z_2, ..., Z(s_n) \leq z_n) = P(Z(s_1 + h) \leq z_1, Z(s_2 + h) \leq z_2, ..., Z(s_n + h) \leq z_n)$

Geostatistical Data: Stationarity

- A weaker form of stationarity assumes that the moments (mean, variance) of the joint density are invariant.
- Called second-order stationarity
- $E[Z(s)] = \mu$
- $Cov(Z(s+h), Z(s)) = C(h)$
- $C(h)$ only depends on distance, h where C is a covariogram

Geostatistical Data: Stationarity

A third form of stationarity is intrinsic stationarity. This is the version we often use because it applies a technique of differencing of the spatial process to obtain stationarity. It derives from weak stationarity, and is what gives us the semivariogram.

- Differencing in what we previously described: Z(s+h)-Z(s)
- It is intrinsic if it has a constant mean and the variance of the differences at pairs of locations only depends on the distance h between locations
- These properties allow us to define the semivariogram (variogram)

$$\frac{1}{2}(Z(s+h) - Z(s)) = \gamma(h)$$

This is the preferred method (and thus instrinsic stationarity is the primiary type of stationary) for characterizing geostatistical spatial processes.

Geostatistical Data: Stationarity

Some additional properties of the semivariogram

- Recall: when the random (spatial) process is stationary it is a function of the spatial lag, or distance only ($\gamma(h)$).
- $\gamma(-h) = \gamma(h)$
- $\gamma(0) = 0$ since $Var(Z(s) - Z(s)) = 0$.
- the spatial process is **isotropic** if $\gamma(h) = \gamma(||h||)$
- the semivariogram and the covariance function are related by:

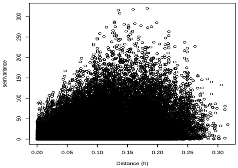
$$
\begin{aligned}
\gamma(h) &= \frac{1}{2}E[(Z(s+h) - Z(s))^2] \\
&= \frac{1}{2}E[((Z(s+h) - \mu) - (Z(s) - \mu))^2] \\
&= -E[(Z(s+h) - \mu)(Z(s) - \mu)] + \frac{1}{2}E[(Z(s+h) - \mu)^2] \\
&\quad + \frac{1}{2}E[(Z(s) - \mu)^2] \\
&= -C(h) + C(0)
\end{aligned}
$$

Geostatistical Data: Stationarity

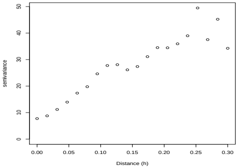
- Spatial processes have the assumption of stationarity, i.e. $E[Z(\mathbf{s})] = \mu$ for all $\mathbf{s} \in D$. This means the mean of the process does not depend on location.
- Stationarity also states that $Cov(Z(s_i), Z(s_j)) = C(s_i - s_j)$. This means that the covariance depends only on the difference between locations s_i and s_j and not on the locations themselves (stationarity). $C(\cdot)$ is the covariance function.

Geostatistical Data: Empirical Semivariograms

Plot the separation distance $||h||$ vs $\gamma(h)$ for all pairs of points, but this is difficult to interpret.

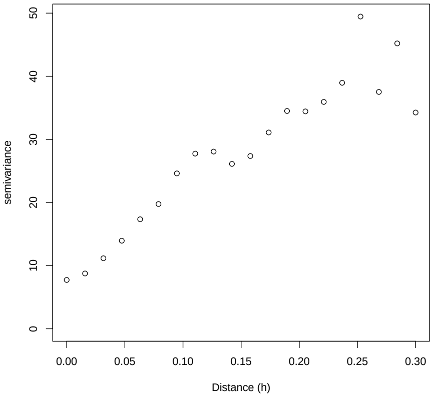


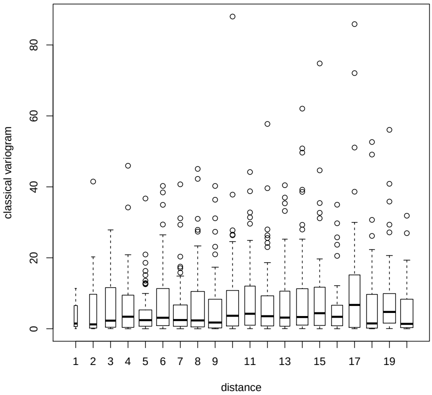
Instead an empirical estimate can be calculated by binning the distances:
$\hat{\gamma} = \frac{1}{2N(h)} \sum_{N(h)} [Z(s_i) - Z(s_j)]^2$



Geostatistical Data: Empirical Semivariograms

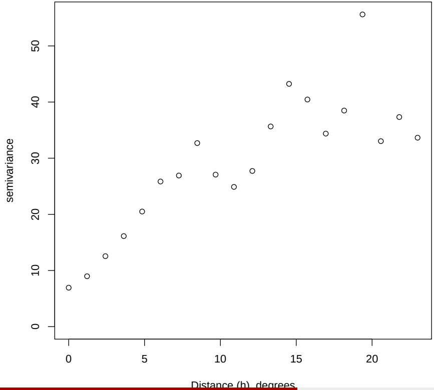
- Divide distance into K intervals $I_1 = (0, h_1), ..., I_K = (h_{K-1}, h_K)$
- $\hat{\gamma}(h_k) = \frac{1}{2N(h_k)} \sum_{N(h)} [Z(s_i) - Z(s_j)]^2$
- $N(h_k)$ is the set of pairs in the interval I_k

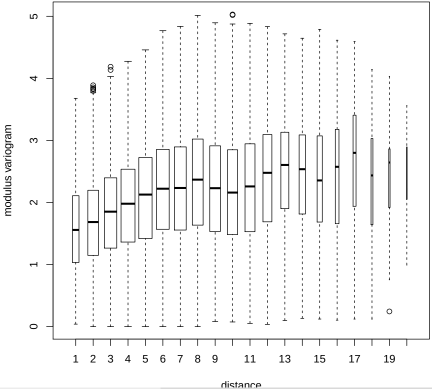




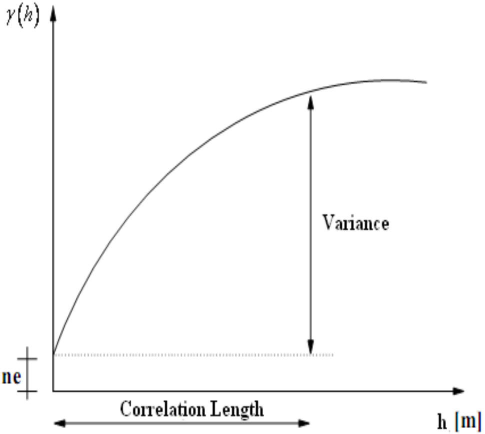
Geostatistical Data: Empirical Semivariograms

- There is a more robust estimate of the variogram by Cressie and Hawkins
- Less sensitive to outliers
- $\hat{\gamma}(h_k) = \frac{1}{N(h_k)} \sum_{N(h)} |Z(s_i) - Z(s_j)|^{1/2}$
- Again, $N(h_k)$ is the set of pairs in the interval I_k



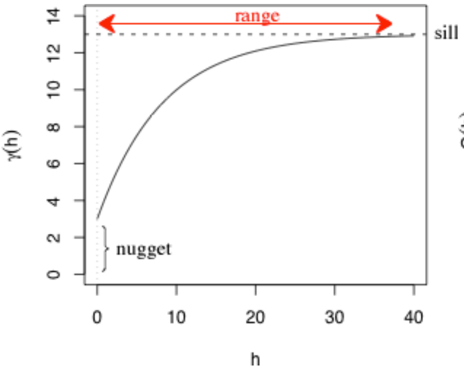


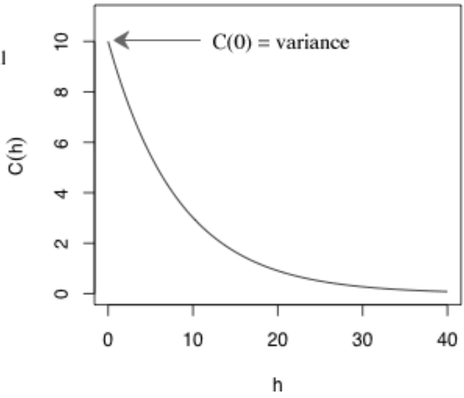
Geostatistical Data: Semivariogram Interpretation



- Observations that are close together are more alike than those far apart: increasing variance in pairwise difference with increasing h means decreasing autocorrelation.

Geostatistical Data: Semivariogram Interpretation





Geostatistical Data: Semivariogram Interpretation

- Strength of spatial structure is based on where the semivariogram reaches an asymptote. This distance is called the **range**, ρ. Beyond this distance, it is assumed that there is no autocorrelation.
- The semivariance where the asymptote is reached is the **sill**, σ^2.
- The discontinuity at the origin is called the **nugget**, τ^2.
- If there is a nugget, be careful to interpret the sill as the value after subtracting the nugget (the 'effective' sill).
- Recall that if the process is not stationary $C(h)$ doesn't exist.

Geostatistical Data: Theoretical Semivariograms

We want to fit a theoretical model to our "empirical" semivariogram to describe the shape of the spatial process.
Common parametric semivariogram functions:

- Exponential
- Spherical
- Gaussian
- Matern

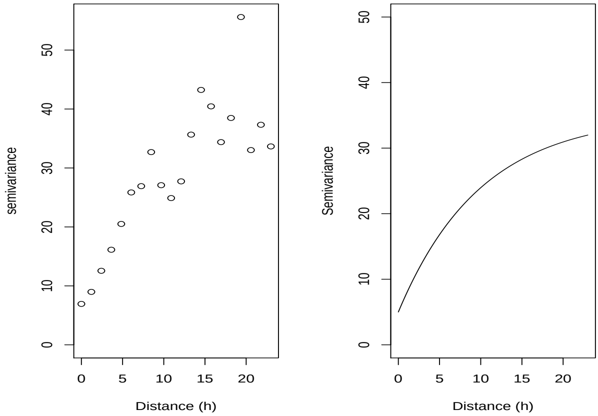
Geostatistical Data: Theoretical Semivariograms

We want to fit a theoretical model to our "empirical" semivariogram to describe the shape of the spatial process.
Common parametric semivariogram functions:

- Exponential
- Spherical
- Gaussian
- Matern
- Linear (beware!)

Exponential:

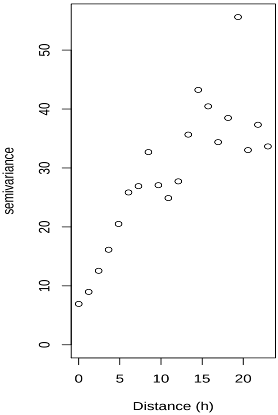
$$\gamma(h) = \tau^2 + \sigma^2(1 - exp(-h/\rho))$$

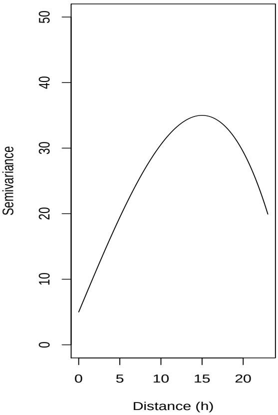


Geostatistical Data: Empirical vs Theoretical

Spherical:

$$\gamma(h) = \tau^2 + \sigma^2(3/2(h/\rho) - 1/2(h/\rho)^3)$$

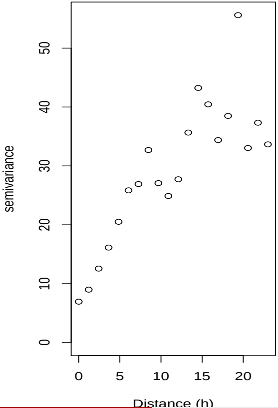


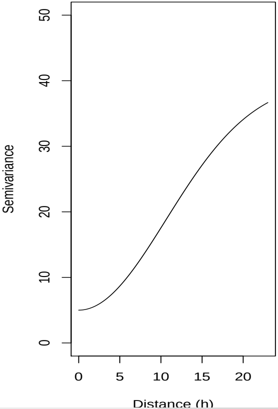


Geostatistical Data: Empirical vs Theoretical

Gaussian:

$$\gamma(h) = \tau^2 + \sigma^2(1 - exp(-\frac{h^2}{\rho^2}))$$





Geostatistical Data: Fitting a Semivariogram

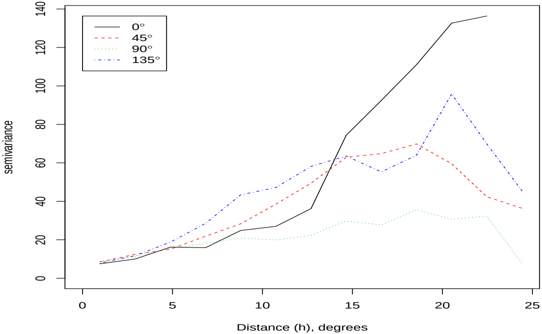
Eyeballing the semivariogram is useful for exploratory purposes and to find the approximate shape of the spatial process, but we would rather find a valid theoretical semivariogram function that reflects the empirical semivariogram.
Ordinary Least Squares: find the parameters $\theta = (\tau^2, \sigma^2, \rho)$ that minimize the squared vertical distance between the empirical and theoretical semivariograms.

$$(\hat{\gamma}(h) - \gamma(h;\theta))^T(\hat{\gamma}(h) - \gamma(h;\theta))$$

where $\hat{\gamma}(h)$ is the empirical semivariogram and $\gamma(h;\theta)$ is the theoretical semivariogram with parameters θ
We will cover this in the next lecture.

Geostatistical Data: Anisotropy

Isotropy means that the semivariance depends only on the distance between points, not direction.
Anisotropy means the semivariance also depends on direction as well as distance.
We can examine anisotropy with a **directional semivariogram**.



More on anisotropy next week.