Spatial Data Analysis

Week 3: Geostatistics I

Meredith Franklin

Department of Statistical Sciences and School of the Environment

September 19th, 2025

Geostatistical Data

In today's lecture we'll cover:

- Empirical semivariograms
- Stationarity, anisotropy
- Theoretical semivariograms
- Fitting semivariogram models
- Covariance functions
- Introduction to kriging

Geostatistical Data: Description

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

Examples:

- Housing prices in a metropolitan area.
- Field observations such as soil samples, water samples, air pollution, noise, weather (environmental data measured at point locations).
- Ecological data such as animal or plant species, habitats.

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

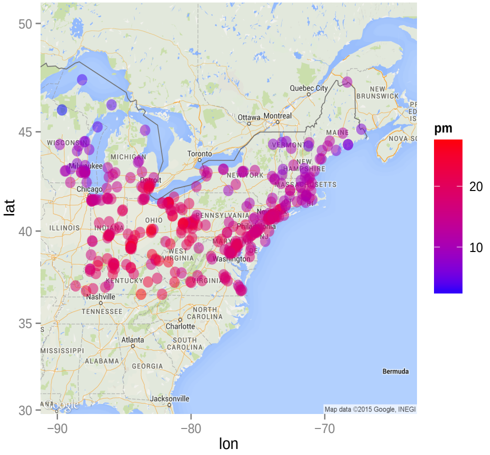
$$Z(\mathbf{s}) : \mathbf{s} \in D$$

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

A random field (stochastic field) is a mathematical model used to represent spatially (or temporally) varying data. It generalizes the concept of random variables (which describe uncertainty in one dimension) to multiple dimensions, such as space (or time).

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 $Z(s)$ (monthly $PM_{2.5}$ concentration).



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 $PM_{2.5}$ 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 observed values at the spatial locations $Z(\mathbf{s})$ as one realization of the random process.

Geostatistical Data: Analytical Goals

Goals of spatial statistics applied to point referenced data

- Visualization of points on a map to look at distribution. Add color 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.

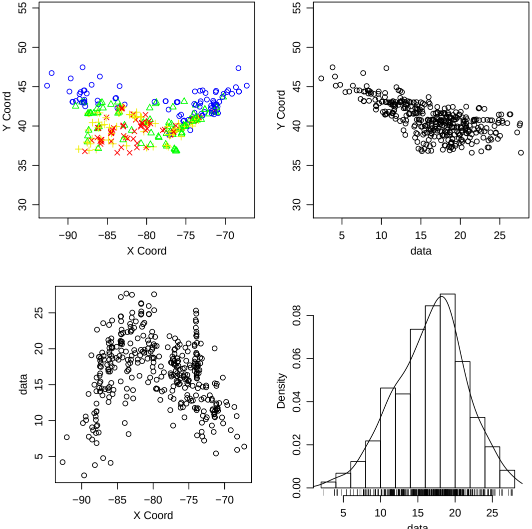
Geostatistical Data: Exploratory Analysis

Exploratory Data Analysis

- Exploratory analysis is critical in statistics and data science. We want to know what are the spread and distribution of the data, outliers, and in spatial statistics outlying locations.
- Might need to transform the data 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.

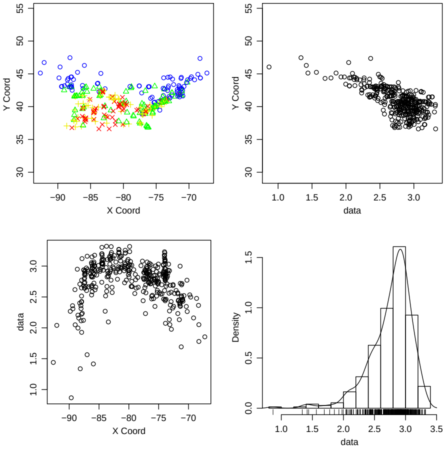
Geostatistical Data: Exploratory Analysis

Using the geoR package, 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, the <u>untransformed</u> data appear closer to normality (less skewed).

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 X and Y

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

Spatial Covariance

- In the spatial setting, **covariance is a function of distance**.
- Covariance and correlation for random processes are often called autocovariance and autocorrelation.
- 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.

Definition of spatial covariance and correlation

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

Where distance $h = |s_i - s_j|$
Spatial correlation, $\rho(h)$ can be defined as:

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

Variogram

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

Variogram and 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

Variogram and Semivariogram

$$2\gamma(h) = E[(Z(s+h) - Z(s))^2]$$
$$2\gamma(h) = E[Z(s)^2] + E[Z(s+h)^2] - 2E[Z(s)Z(s+h)]$$

where $E[Z(s)^2] = E[Z(x+h)^2] = \sigma^2$ (variance at a location, also $C(0)$), and $E[Z(s)Z(s+h)] = C(h)$ (covariance between values separated by distance h). Substituting:

$$2\gamma(h) = \sigma^2 + \sigma^2 - 2C(h)$$
$$2\gamma(h) = 2\sigma^2 - 2C(h)$$
$$\gamma(h) = C(0) - C(h)$$

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.
- There is strong stationarity and weak (second-order) stationarity.

Strong Stationarity

- It is useful to view spatial data as multivariate, despite it being the same measurement (i.e. $PM_{2.5}$ 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: $P(Z(s_1) \leq z_1, Z(s_2) \leq z_2, ..., Z(s_n) \leq z_n) = P(Z(s_1 + h) \leq z_1, Z(s_2 + h) \leq z_2, ..., Z(s_n + h) \leq z_n)$.

Second-Order Stationarity

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

Second-Order Stationarity

- With 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 $\text{Cov}(Z(s_i), Z(s_j)) = C(s_i - s_j)$. 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.

Second-Order (Intrinsic) Stationarity

Weak or second-order stationarity is also called 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.

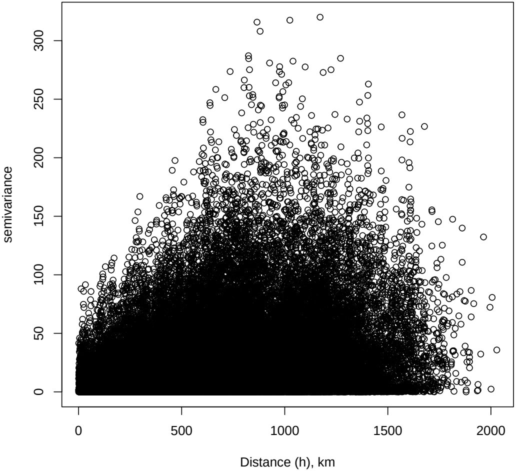
- Differencing in what we previously described: $Z(s+h) - Z(s)$
- It is intrinsic if it has a constant mean and the variance of the differences at pairs of locations only depends on the distance h between locations
- These properties allow us to define the semivariogram (variogram)

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

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

Empirical Semivariograms

We construct a semivariogram by plotting the separation distance h on the x-axis vs $\gamma(h)$ on the y-axis for all pairs of points.



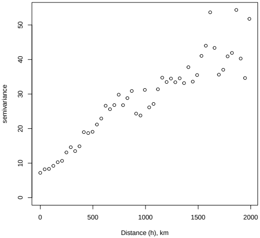
Empirical Semivariograms

Since this is hard to interpret, instead an empirical estimate can be calculated by binning the distances:

$$\hat{\gamma} = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [Z(s_i) - Z(s_j)]^2$$

Since s_i and s_j are separated by distance h it is sometimes written:

$$\hat{\gamma} = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [Z(s_i) - Z(s_i + h)]^2$$

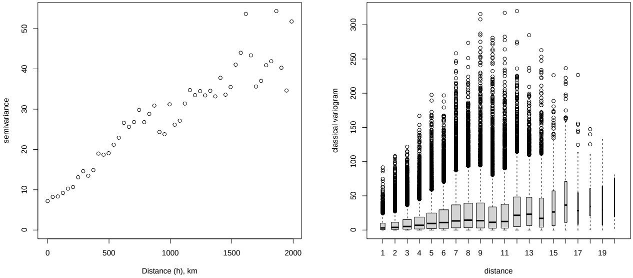


Empirical Semivariograms

Binning the distances involves dividing them into K intervals $\mathrm{I}_1 = (0, h_1), \ldots, I_K = (h_{K-1}, h_K)$. With this we can more specifically state:

$$\hat{\gamma}(h_k) = \frac{1}{2N(h_k)} \sum_{i=1}^{N(h)} [Z(s_i) - Z(s_j)]^2$$

Where $N(h_k)$ is the set of pairs in the interval I_k

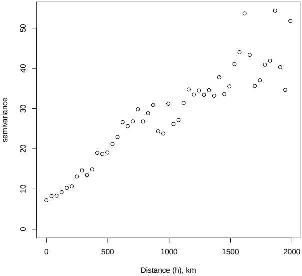


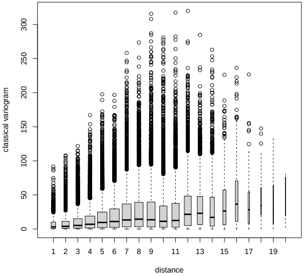
Empirical Semivariograms

- There is a more robust estimate of the semivariogram developed by Cressie and Hawkins that is less sensitive to outliers.

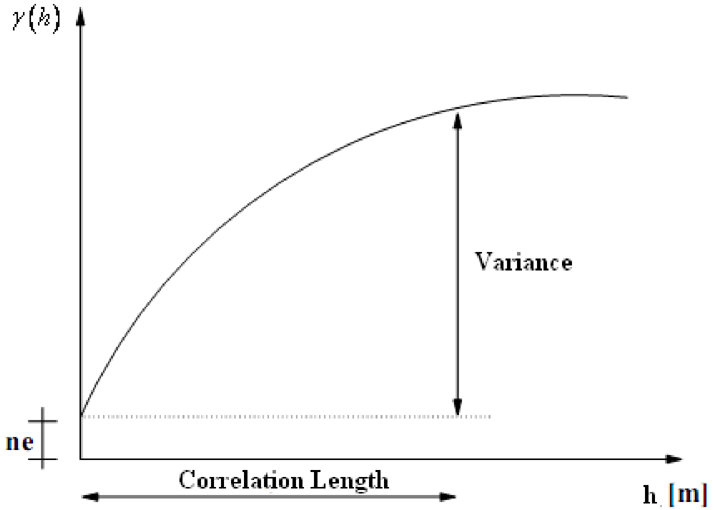
$$\gamma^*(h_k) = \frac{\left(\frac{1}{N(h_k)} \sum_{(i,j)\in N(h_k)} \left|Z(s_i) - Z(s_j)\right|^{1/2}\right)^4}{0.457 + \frac{0.494}{N(h_k)}}$$

- Again, $N(h_k)$ is the set of pairs in the interval I_k



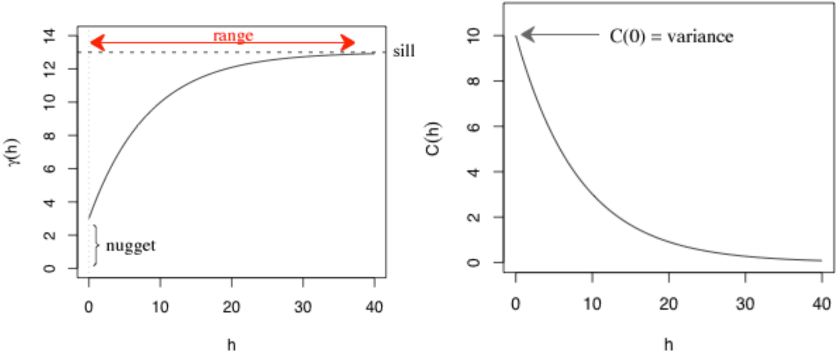


Semivariogram Interpretation

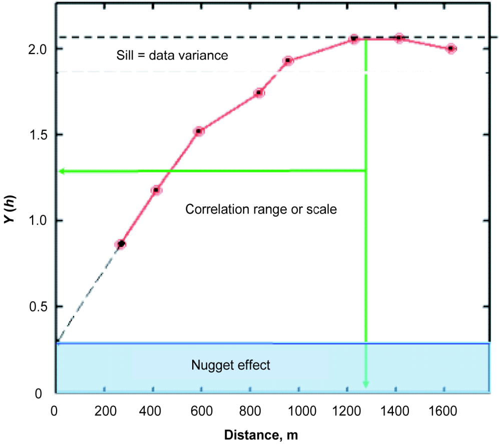


- Observations that are close together are more alike than those far apart: increasing variance in pairwise difference with increasing h means decreasing autocorrelation.

Semivariogram Interpretation



Another example of a semivariogram

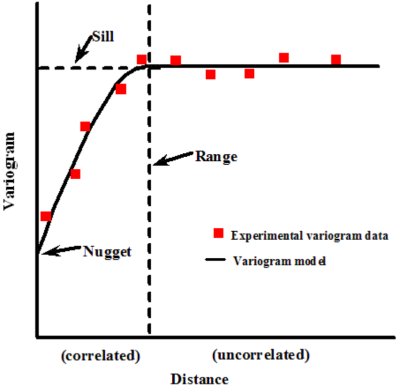


Semivariogram Interpretation

- Strength of spatial structure is based on where the semivariogram reaches an asymptote. This distance is called the **range**, ϕ. 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.

Theoretical Semivariograms

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

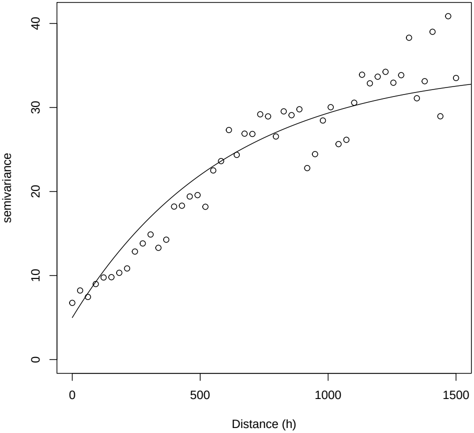


Common parametric semivariogram functions include: Exponential, Spherical, Gaussian, Matern, Linear (beware! Valid only for small ranges)

Theoretical: Exponential

Exponential:

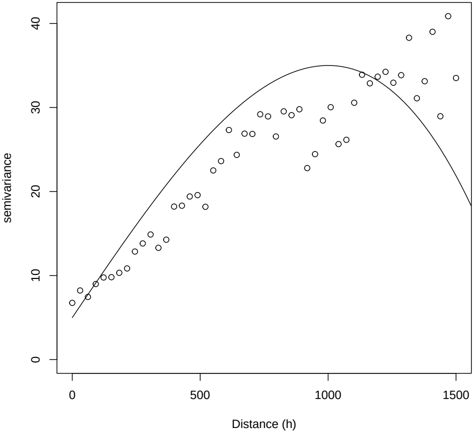
$$\gamma(h) = \tau^2 + \sigma^2(1 - exp(-h/\phi))$$



Theoretical: Spherical

Spherical:

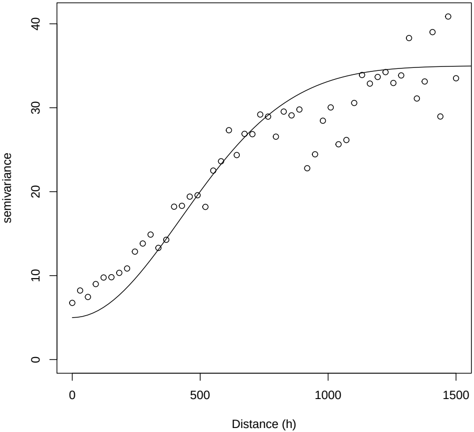
$$\gamma(h) = \tau^2 + \sigma^2(3/2(h/\phi) - 1/2(h/\phi)^3)$$



Theoretical: Gaussian

Gaussian:

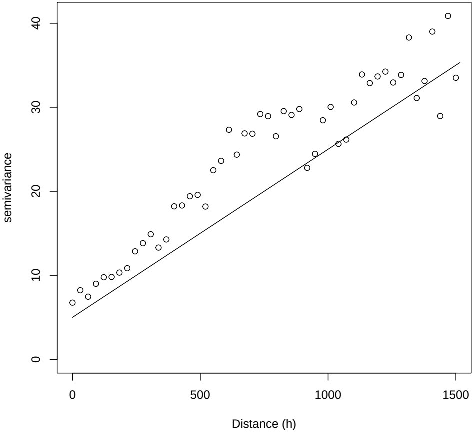
$$\gamma(h) = \tau^2 + \sigma^2(1 - exp(-\frac{h^2}{\phi^2}))$$



Theoretical: Linear

Linear:

$$\gamma(h) = \tau^2 + \sigma^2 h$$



Theoretical Semivariogram Properties

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:

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

Semivariogram Properties

Another way of writing the previous equations (in terms of variogram):

$$
\begin{aligned}
2\gamma(h) &= \mathrm{Var}[Z(s+h) - Z(s)] \\
&= \mathrm{Var}[Z(s+h)] + \mathrm{Var}[Z(s)] - 2\ \mathrm{Cov}[Z(s+h), Z(s)] \\
&= C(0) + C(0) - 2C(h) \\
&= 2\,[C(0) - C(h)]
\end{aligned}
$$

Thus, $\gamma(h) = C(0) - C(h)$. We see that if $C(h)$ exists, then we can get $\gamma(h)$, but can we get $C(h)$ from $\gamma(h)$? This requires the assumptions of stationarity.

Semivariogram and Covariance

- Under stationarity and ergodicity $C(h) \to 0$ as $||h|| \to \infty$ (i.e. covariance goes to 0 as distance goes to infinity where $||h||$ denotes the length of the h vector).
- If we take the limit on both sides of $\gamma(h) = C(0) - C(h)$, get $\lim_{h \to \infty} \gamma(h) = C(0)$
- However, the limit may not exist, for example in the linear semivariogram:

$$\gamma(h) = \tau^2 + \sigma^2 h \text{ if } h > 0; 0 \text{ otherwise}$$
$$\text{as } h \to \infty \text{ then } \gamma(h) \to \infty$$

- The linear semivariogram is not a second-order stationary process and $C(h)$ does not exist.

Ergodicity

- In geostatistics, we only observe **one realization** of a spatial process (e.g., one pollution map).
- A process is **ergodic** if spatial averages from that realization approximate the true expectations of the underlying random field.

Formal idea

For a spatial process $Z(\mathbf{s})$,

$$\lim_{|D|\to\infty} \frac{1}{|D|} \int_D Z(\mathbf{s})\, d\mathbf{s} = \mathbb{E}[Z(\mathbf{s})]$$

- **Why important?** Ergodicity justifies using **empirical semivariograms** (based on averages over space) to estimate theoretical semivariograms (defined as expectations).

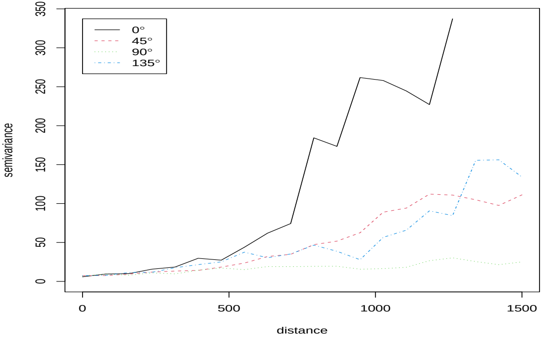
Stationarity and Anisotropy

What is a non-stationary process?

- The basic idea is that the parameters of a semivariogram model, i.e. the nugget, range, sill vary spatially.
- Anisotropy may be the issue.
- Tackling anisotropy and addressing general spatial trends before modeling the semivariance help deal with non-stationarity.
- Recall that semivariogram/covariance functions are only valid for stationary processes.

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**.

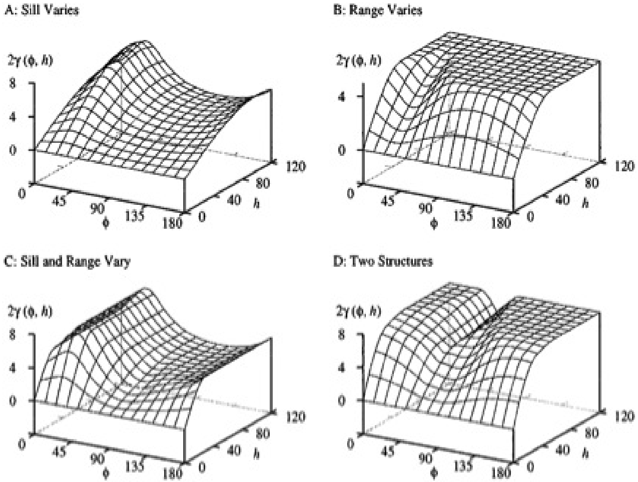


Anisotropy

Anisotropy

- There are two types of anisotropy: Geometric and Zonal
- Geometric: directional semivariograms have the same shape and sill, but different ranges. Sometimes called range anisotropy.
- To make semivariograms isotropic, adapt our known isotropic semivariograms using elliptical geometry.
- Rotate the coordinate axes so they are aligned with the major and minor axes of the ellipse.
- Zonal isotropy: when sill changes with direction but the range remains constant. Sometimes called sill anisotropy.
- See Eriksson and Siska (2000) for more information.

Anisotropy



Eriksson, M and P.P Siska (2000)

Anisotropy

What do we do about anisotropy if it is detected in our data?

- First try taking out linear or quadratic trends in x,y then look at directional semivariogram of residuals.
- Determine whether you have geometric or zonal isotropy. Geometric easier to deal with. Rather than isotropic spherical contours, apply elliptical contours in direction of anisotropy (spatial range different in different directions).

Anisotropy

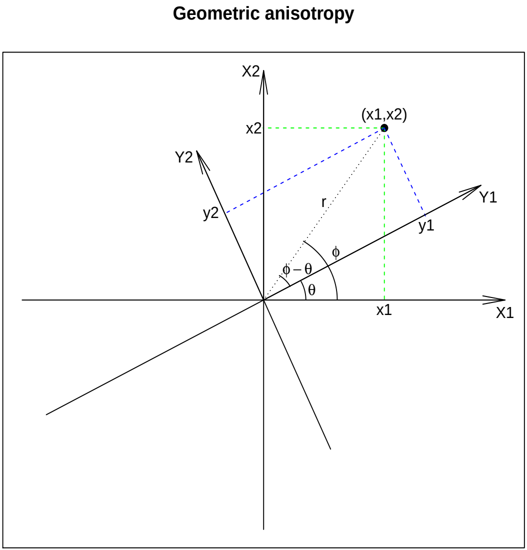
$$C(h) = C(s_i - s_j) = C([(s_i - s_j)^{'} B(s_i - s_j)]^{1/2})$$

C($\cdot$) is a valid isotropic covariance function, and B is a symmetric positive definite matrix that characterizes the elliptical contours. For example the isotropic and geometric anisotropic versions of the spherical semivariogram are:

$$C(h) = \sigma^2 \exp(-\phi^2 ||h||^2) = \sigma^2 \exp(-\phi^2 h^{'} h)$$
$$C(h) = \sigma^2 \exp(-\phi^2 h^{'} Bh)$$

Can apply equations found in Eriksson and Siska (2000).

Anisotropy



Semivariogram Models

- Eyeballing the semivariogram function is useful for exploratory purposes and to find the approximate shape of the spatial process.
- However, we would rather find a valid theoretical semivariogram function that reflects the empirical semivariogram.
- We choose from our set of valid theoretical semivariograms and see how well the function fits to our data.
- We can't just pick any curve that looks to fit our data because the semivariogram model must be **conditionally negative definite** to ensure that results aren't off (i.e. the covariances of multiple points are inconsistent with each other, or could have negative variance for weighted averages).

Conditionally Negative Definite Functions

- For a semivariogram $\gamma(h)$ to be **valid**, it must be **conditionally negative definite**.
- Formal definition: for any points $s_1, \ldots, s_n$ and real numbers $a_1, \ldots, a_n$ with $\sum_{i=1}^{n} a_i = 0$,

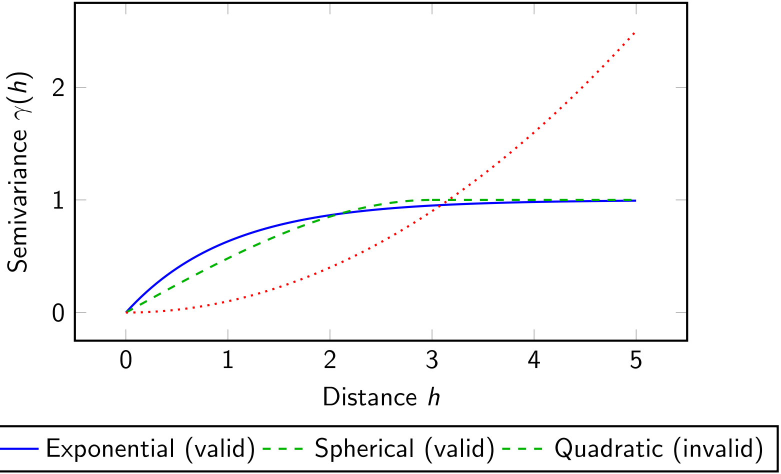
$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \, \gamma(s_i - s_j) \; \leq \; 0$$

- Why? Because covariance functions must be positive definite. Since $\gamma(h) = C(0) - C(h)$, this implies $\gamma(h)$ must satisfy conditional negative definiteness.

Intuition

- Ensures semivariograms lead to valid covariance matrices.
- Prevents "nonsense" results like negative variances in kriging.
- Valid models: Exponential, Spherical, Gaussian, Matérn, etc.
- Invalid example: $\gamma(h) = h^2$ (not conditionally negative definite).

Valid vs Invalid Semivariogram Models



Valid semivariogram models are **conditionally negative definite**, ensuring covariance matrices are valid. Not every curve that "looks reasonable" (e.g. quadratic) is valid.

Fitting a Semivariogram Model: Overview

- We can fit the theoretical semivariogram function to the data in a variety of ways: ordinary least squares (OLS), weighted least squares (WLS), maximum likelihood (ML), and restricted maximum likelihood (REML).
- Main objective is to estimate the parameters $\theta = (\tau^2, \sigma^2, \phi)$ that best fit the theoretical semivariogram to the empirical semivariogram.
- In Zimmerman and Zimmerman, A comparison of spatial semivariogram estimators and corresponding ordinary kriging predictors, Technometrics, 1991: 33(1)77-91 it was found that ML/REML is generally the best procedure to use, but (approximate) WLS very good compared other methods which are subject to erratic behavior in some situations.
- WLS is robust and does not require any distributional assumptions, so it is a good choice for semivariogram estimation.

Fitting a Semivariogram Model: OLS (1)

In ordinary least squares (OLS), we estimate parameters by minimizing the sum of squared residuals:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \big(Y_i - f(\mathbf{X}_i; \boldsymbol{\theta})\big)^2$$

- Y_i = observed values (response variable)
- $f(\mathbf{X}_i; \boldsymbol{\theta})$ = model prediction
- $\boldsymbol{\theta}$ = vector of parameters to estimate

This generic regression idea extends to semivariogram fitting.

Fitting a Semivariogram Model: OLS (2)

Consider the sample semivariogram values as observations, and fit a model as a function of distance h.

- Parameters $\boldsymbol{\theta} = (\tau^2, \sigma^2, \phi)$ minimize:

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

  which is equivalent to

$$SSE = \sum_h \left(\hat{\gamma}(h) - \gamma(h; \boldsymbol{\theta})\right)^2$$

- $\hat{\gamma}(h) =$ empirical semivariogram
- $\gamma(h; \boldsymbol{\theta}) =$ theoretical semivariogram

Fitting a Semivariogram Model: OLS (3)

OLS relies on assumptions:

- Independence (i.i.d.) of observations $\Rightarrow$ violated since each pair contributes to multiple bins
- Equal variance of bins (homoskedasticity) $\Rightarrow$ violated since bins with more pairs have lower variance

Solution: use Generalized Least Squares (GLS) or Weighted Least Squares (WLS).

Fitting a Semivariogram Model: GLS (1)

We use the binned empirical semivariogram:

$$\hat{\gamma}(h_k) = \frac{1}{2\,|N(h_k)|} \sum_{(i,j)\in N(h_k)} \left[Z(s_i) - Z(s_j)\right]^2$$

- Relationship between $\hat{\gamma}(h)$ and h is nonlinear.
- Use **generalized least squares (GLS)**, solved numerically.
- Objective: minimize

$$SSE = \sum_{k=1}^{K} \left[\hat{\gamma}(h_k) - \gamma(h_k; \boldsymbol{\theta})\right]^2,$$

where K is the number of bins.

Fitting a Semivariogram Model: GLS (2)

In matrix form, GLS minimizes:

$$(\hat{\gamma}(h) - \gamma(h; \boldsymbol{\theta}))^T \mathbf{V}^{-1} (\hat{\gamma}(h) - \gamma(h; \boldsymbol{\theta}))$$

- $\mathbf{V}$ = covariance matrix of $\hat{\gamma}(h)$
- Accounts for correlation among bins
- Problem: $\mathbf{V}$ depends on unknown $\boldsymbol{\theta}$ and is computationally expensive
- Practical solution: use **approximation** $\Rightarrow$ Weighted Least Squares (WLS)

Fitting a Semivariogram Model: WLS

- If $\mathbf{V}^{-1} = \mathbf{I}$, we recover OLS.
- If $\mathbf{V}^{-1} = \text{diag}\{\text{Var}[\hat{\gamma}(h_1)], \ldots, \text{Var}[\hat{\gamma}(h_K)]\}$, we obtain WLS.
- Approximate variance:

$$\text{Var}[\hat{\gamma}(h_j)] \;\approx\; \frac{2[\hat{\gamma}(h_j)]^2}{N(h_j)}$$

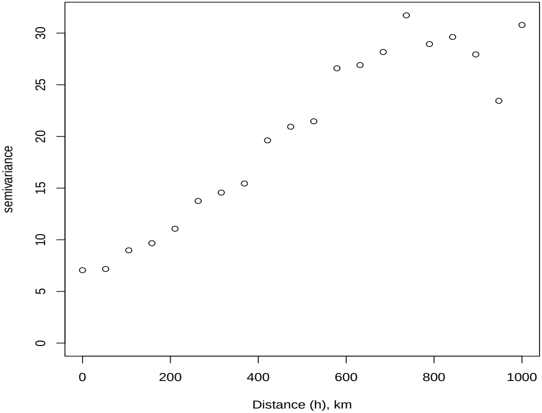
- Weighted SSE:

$$WSSE = \frac{1}{2} \sum_{j=1}^{K} \frac{N(h_j)}{\hat{\gamma}(h_j)} \left[\hat{\gamma}(h_j) - \gamma(h_j; \boldsymbol{\theta})\right]^2$$

- WLS gives more weight to bins with more pairs (more precise).

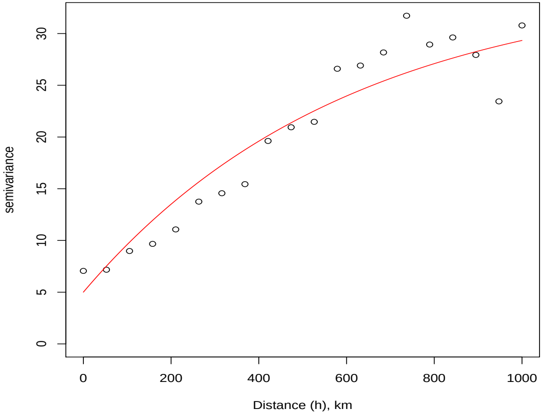
Fitting a Semivariogram Model

The binned empirical semivariogram using robust estimator (Cressie-Hawkins), 20 bins, projected coordinates, and maximum distance 1000 km.



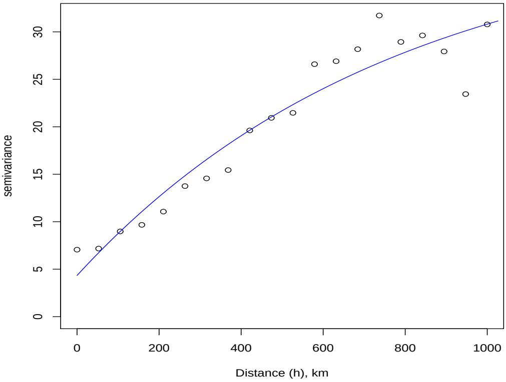
Fitting a Semivariogram Model

The binned empirical semivariogram with an eyeballed theoretical exponential semivariogram that looks to fit: `curve(5+30*(1-exp(-x/600)))`



Fitting a Semivariogram Model

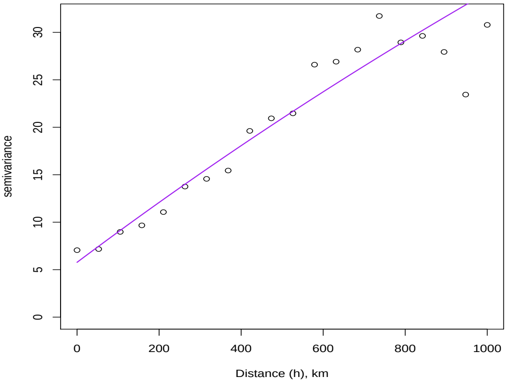
Result of fitting an exponential semivariogram by OLS to empirical semivariogram. Using `geoR's variofit()` **with weights="equal"**



Estimated parameters: $\hat{\sigma}^2 = 36.574$, $\hat{\phi} = 777.148$, $\hat{\tau}^2 = 4.339$; SSE: 120.603

Fitting a Semivariogram Model

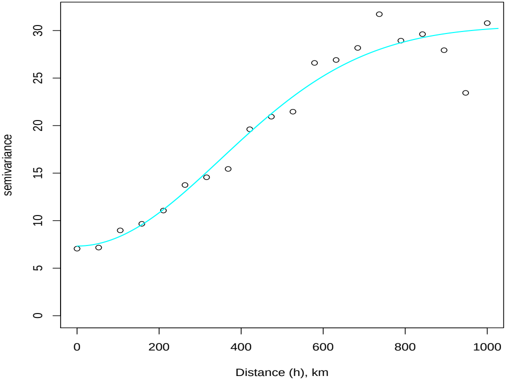
Result of fitting an <u>exponential</u> semivariogram by WLS to empirical semivariogram. Using `geoR's variofit()` with weights="cressie"



Estimated parameters: $\hat{\sigma}^2 = 121.6040$, $\hat{\phi} = 3755.1280$, $\hat{\tau}^2 = 5.7735$; SSE: 592.767
Note problem with range estimate!

Fitting a Semivariogram Model

Result of fitting a <u>Gaussian</u> semivariogram by WLS to empirical semivariogram. Using `geoR's variofit()` with weights="cressie"



Estimated parameters: $\hat{\sigma}^2 = 23.2096$, $\hat{\phi} = 494.8422$, $\hat{\tau}^2 = 7.3303$; SSE: 277.4442
Much better estimates!

Fitting a Semivariogram Model: Practical Issues

- Extensive literature exists, but fitting remains somewhat **arbitrary** and not fully rigorous statistically.
- Objective function (deviation between empirical and model semivariogram) is based on **pseudo-data**.
- Fitting is essentially **curve fitting** and is sensitive to:
  - binning scheme (number and width of bins),
  - choice of maximum distance.
- Advantage: calculations are fast, even with large datasets.
- Semivariogram models are not tied to a specific probability model for the data $\Rightarrow$ can be more robust to assumption violations.
- Compare SSE across different **models within the same method** (e.g., exponential vs spherical under WLS). Do **not** compare SSE across different fitting methods (e.g., OLS vs WLS).

Fitting a Covariance Model

- Recall the link: $C(h) = C(0) - \gamma(h)$, where $C(0) = \sigma^2$.
- The **statistical approach** is to fit a covariance function by maximum likelihood (ML).
- ML is widely used in statistics for parameter estimation.
- Example: in linear regression with normal, uncorrelated errors, least squares $=$ ML.
- ML requires a probability model (likelihood) for the data.
- For spatial data: assume a multivariate Gaussian process with second-order stationarity.

Fitting a Covariance Model: ML

- Gaussian process:

$$Z \sim \mathrm{MVN}\big(\mu, \boldsymbol{\Sigma}(\boldsymbol{\theta})\big)$$

- Spatial regression form:

$$E[Z(s)] = \mu + \epsilon(s) = X\beta + \epsilon(s), \quad \epsilon(s) \sim N(0, \boldsymbol{\Sigma}(\boldsymbol{\theta})).$$

- Log-likelihood:

$$\ell(\beta, \boldsymbol{\theta}; Z) = -\tfrac{1}{2} \ln |\boldsymbol{\Sigma}(\boldsymbol{\theta})| - \tfrac{1}{2}(Z - X\beta)^T \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}(Z - X\beta).$$

- Restricted maximum likelihood (REML): maximizes the likelihood of residuals (data with mean removed), correcting for degrees of freedom.
- REML reduces bias in covariance parameter estimates, especially when the mean is unknown or spatially varying.

Fitting a Covariance Model: ML/REML

- Goal: choose parameters $\boldsymbol{\theta}$ that maximize the likelihood of the observed data.
- Data are fixed, parameters are unknown.
- No closed-form solution $\Rightarrow$ optimization is numerical.
- Computation becomes intensive for large datasets ($n > 200$–500).

Choosing a Model: ML/REML

- Compare models using **Akaike's Information Criterion (AIC)**:

$$\mathrm{AIC} = 2k - 2\ln(\hat{L}),$$

  where k = number of parameters and $\hat{L}$ = maximized likelihood.
- AIC penalizes models with more parameters.
- Can be used to compare **non-nested** models.
- Models with smaller AIC are preferred.

Using Variogram and Covariance Models

When using a semivariogram:

- Fit theoretical semivariogram to empirical data to estimate nugget, sill, range.
- Use fitted semivariogram for prediction (kriging).
- Requires second-order stationarity.
- Must convert $\gamma(h)$ into a covariance $C(h)$.

When using covariance ML/REML:

- Directly estimate covariance function.
- No conversion from semivariogram required.

Kriging: Intro/Preview

- Kriging = spatial prediction at unobserved locations.
- Based on fitted covariance function and spatial regression:

$$E[Z(s)] = X\beta + \epsilon(s).$$

- Objective: estimate $Z(s_0)$ at unsampled location s_0 from observed values $Z(s_1), \ldots, Z(s_n)$.

Kriging Recipe

1. Choose a parametric semivariogram or covariance model.
2. Estimate parameters (via variogram fitting or ML/REML).
3. Make predictions and uncertainty estimates given parameters.

- Kriging predictions are weighted averages of observations.
- Weights depend on spatial correlation structure (distance + covariance/variogram).

Kriging Predictor

- Kriging predictor at s_0:

$$\hat{Z}(s_0) = \sum_{i=1}^{n} \lambda_i Z(s_i)$$

- Weights λ_i chosen to minimize mean squared error:

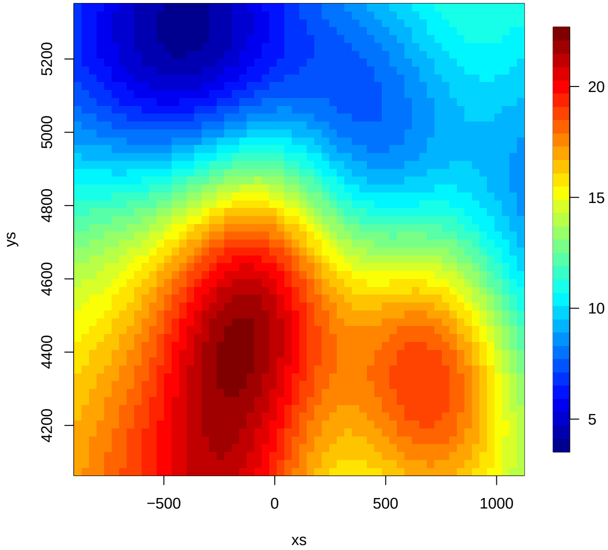
$$E\left[(\hat{Z}(s_0) - Z(s_0))^2\right].$$

- Best predictor is the conditional mean:

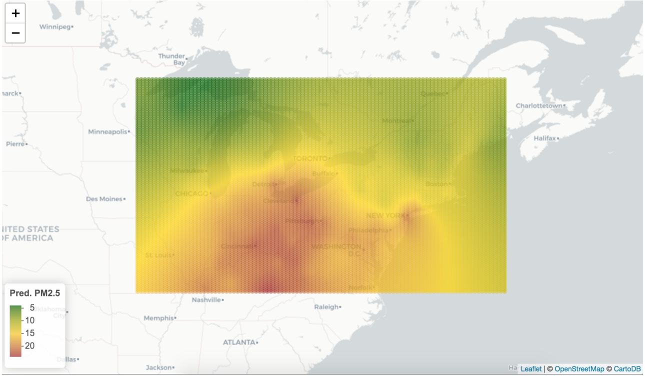
$$E\left[Z(s_0) \,\middle|\, Z\right].$$

- Requires knowledge (or estimates) of the covariance function.

Kriging Result



Kriging Result



Summary: Semivariogram vs Covariance Fitting

Semivariogram Fitting	Covariance Fitting (ML/REML)
Empirical $\hat{\gamma}(h)$ estimated from data, then fit with a theoretical model	Fit covariance function $C(h)$ directly using likelihood
Methods: OLS, GLS, WLS	Methods: ML, REML
Fast to compute, even with large datasets	Computationally intensive (matrix inversion, determinant)
Relies on binning choices (number, width, max distance)	No binning required
More heuristic: curve fitting to pseudo-data	Statistically rigorous: full probabilistic model
Robust, less sensitive to distributional assumptions	Assumes Gaussian process and second-order stationarity
Parameters: nugget, sill, range estimated indirectly	Parameters estimated directly via likelihood

Workflow: Semivariogram Fitting

Spatial data $Z(s_1), \ldots, Z(s_n)$

↓

Compute empirical semivariogram $\hat{\gamma}(h)$

↓

Fit theoretical semivariogram model (OLS / GLS / WLS)

↓

Convert to covariance $C(h)$ and use in kriging

Workflow: Covariance Fitting (ML/REML)

Spatial data $Z(s_1), \ldots, Z(s_n)$

$\downarrow$

Fit covariance function $C(h; \theta)$ directly by ML / REML

$\downarrow$

Use fitted covariance $C(h)$ in kriging