

# Spatial Statistics

## Introduction to Spatial Analysis

Xiaozhe Yin

Division of Biostatistics, University of Southern California

June 30, 2022

# Geostatistical Data

In this lecture we will cover:

- ▶ Mapping geostatistical data
- ▶ Exploratory tools for geostatistical data
- ▶ Definition of distance
- ▶ The semivariogram
- ▶ Stationarity
- ▶ Empirical semivariograms
- ▶ Theoretical semivariograms
- ▶ Kriging for interpolation
- ▶ Non-parametric smoothing for interpolation

# 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

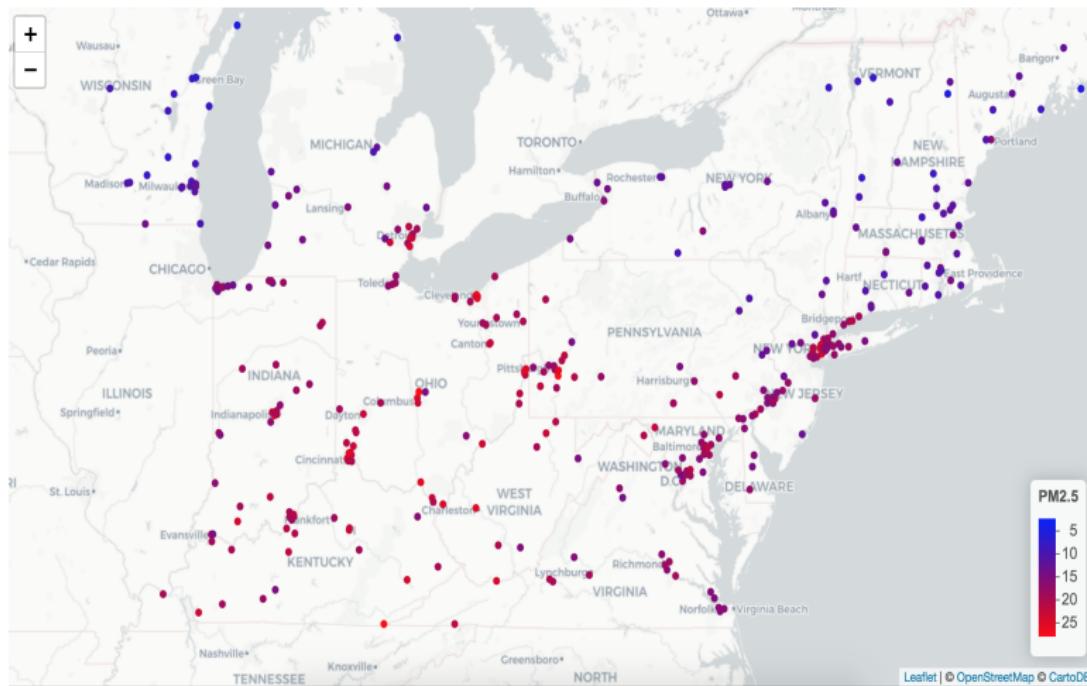
**The common thread that links the data is a random process (also called stochastic process or random field)**

$$Z(s) : s \in D$$

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

# Geostatistical Data: Example

Recall our example of PM<sub>2.5</sub> in the northeast US. Here we have each point representing a location (latitude, longitude), and an associated Z(s) which is monthly PM<sub>2.5</sub> concentrations.



# Geostatistical Data: Analytical Goals

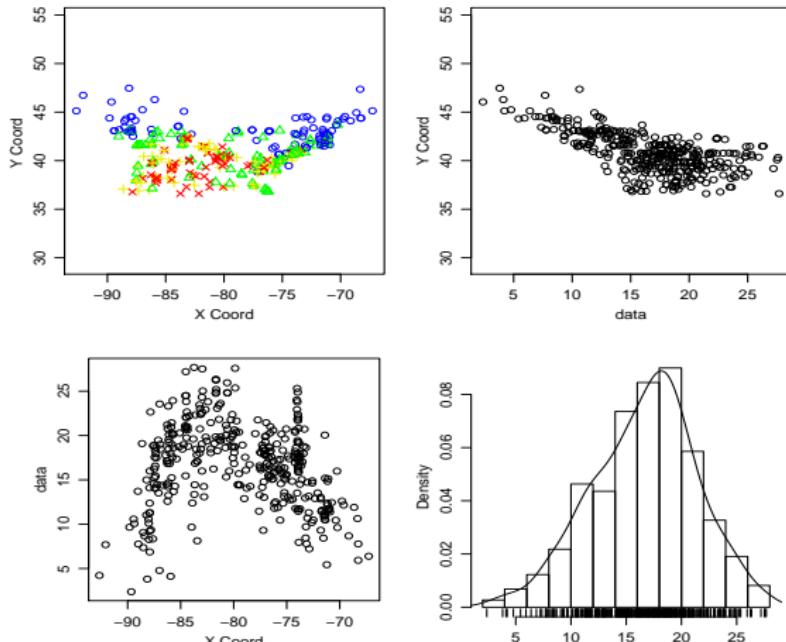
## Goals of spatial statistics applied to point referenced data

- ▶ 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.

# Geostatistical Data: Exploratory Analysis

Using the geoR library, we can do quick explorations of spatial data. Here we show the PM<sub>2.5</sub> concentrations, the PM<sub>2.5</sub> trend in each x (longitude) and y (latitude) directions, and a histogram of the PM<sub>2.5</sub> concentrations.

Create an `as.geodata()` object, and plot it.



# 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  $s_1, s_2, \dots, s_n$  in  $D$  (e.g. GPS locations of PM<sub>2.5</sub> 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 key.
- ▶ How is distance defined?
- ▶ Irregular spacing leads to few (usually one) pair of points for a given 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 ( $d(s_i, s_j) \leq d(s_i) + d(s_j)$ )
- ▶ 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: 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. PM<sub>2.5</sub> concentration) at multiple locations
- ▶ We have a joint probability density:  
$$F(Z(\mathbf{s})) = P(Z(s_1) \leq z_1, Z(s_2) \leq z_2, \dots, Z(s_n) \leq z_n)$$
- ▶ Strong stationarity means that the joint density is invariant under translation:  
$$P(Z(s_1) \leq z_1, Z(s_2) \leq z_2, \dots, Z(s_n) \leq z_n) = P(Z(s_1 + h) \leq z_1, Z(s_2 + h) \leq z_2, \dots, Z(s_n + h) \leq z_n)$$

# 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 intrinsic stationarity is the primary 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:

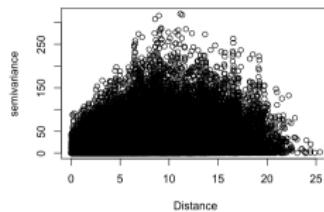
$$\begin{aligned}\gamma(h) &= \frac{1}{2}E[(Z(s+h) - Z(s))^2] \\ &= \frac{1}{2}E[((Z(s+h) - \mu) - (Z(s) - \mu))^2] \\ &= -E[(Z(s+h) - \mu)(Z(s) - \mu)] + \frac{1}{2}E[(Z(s+h) - \mu)^2] \\ &\quad + \frac{1}{2}E[(Z(s) - \mu)^2] \\ &= -C(h) + C(0)\end{aligned}$$

# Geostatistical Data: Stationarity

- ▶ Spatial processes have the assumption of stationarity, i.e.  $E[Z(s)] = \mu$  for all  $s \in D$ . 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.

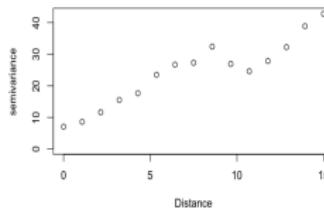
# Geostatistical Data: Empirical Semivariograms

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



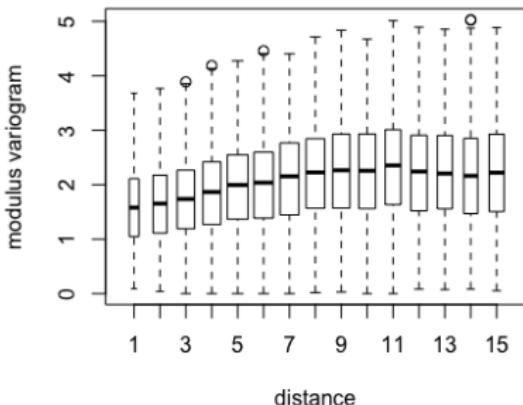
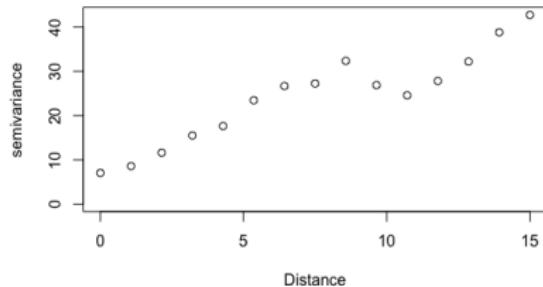
Instead an empirical estimate can be calculated by binning the distances:

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



# Geostatistical Data: Empirical Semivariograms

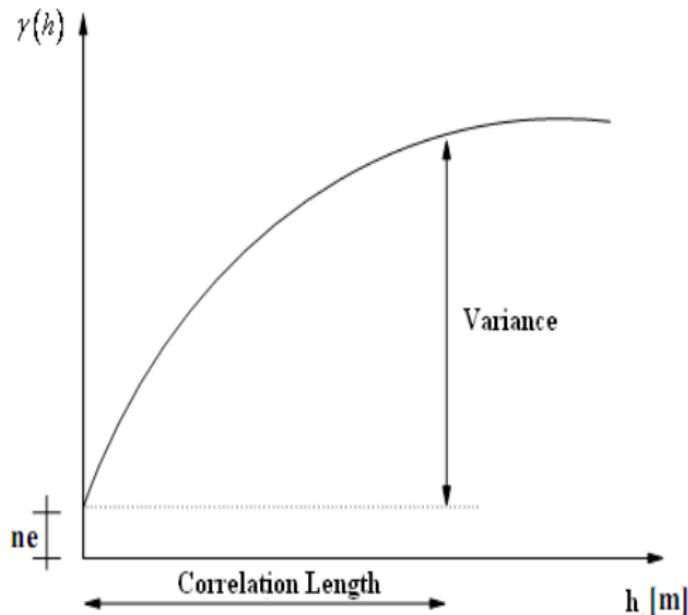
- ▶ Divide distance into K intervals  $I_1 = (0, h_1), \dots, I_K = (h_{K-1}, h_K)$
- ▶  $\hat{\gamma}(h_k) = \frac{1}{2N(h_k)} \sum_{N(h)} [Z(s_i) - Z(s_j)]^2$
- ▶  $N(h_k)$  is the set of pairs in the interval  $I_k$



## Geostatistical Data: Empirical Semivariograms

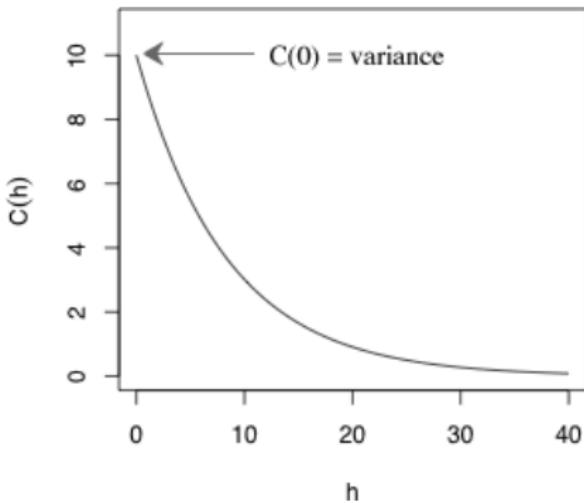
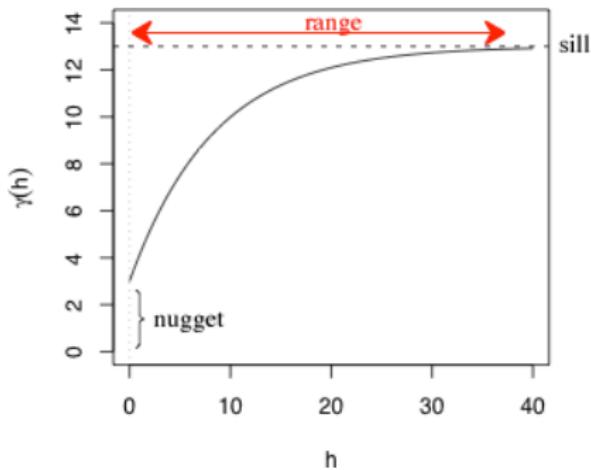
- ▶ There is a more robust estimate of the variogram by Cressie and Hawkins
- ▶ Less sensitive to outliers
- ▶  $\hat{\gamma}(h_k) = \frac{1}{N(h_k)} \sum_{N(h)} |Z(s_i) - Z(s_j)|^{1/2}$
- ▶ Again,  $N(h_k)$  is the set of pairs in the interval  $I_k$

# 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**,  $\rho$ . Beyond this distance, it is assumed that there is no autocorrelation.
- ▶ The semivariance where the asymptote is reached is the **sill**,  $\sigma^2$ .
- ▶ The discontinuity at the origin is called the **nugget**,  $\tau^2$ .
- ▶ 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.

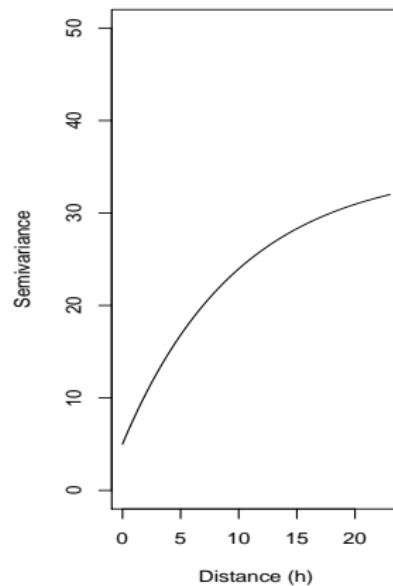
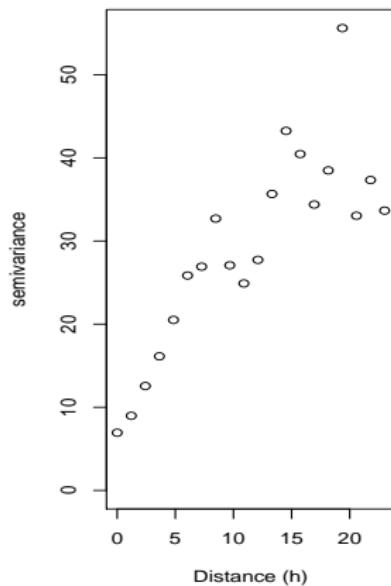
Common parametric semivariogram functions:

- ▶ Exponential
- ▶ Spherical
- ▶ Gaussian
- ▶ Matern

# Geostatistical Data: Empirical vs Theoretical

Exponential:

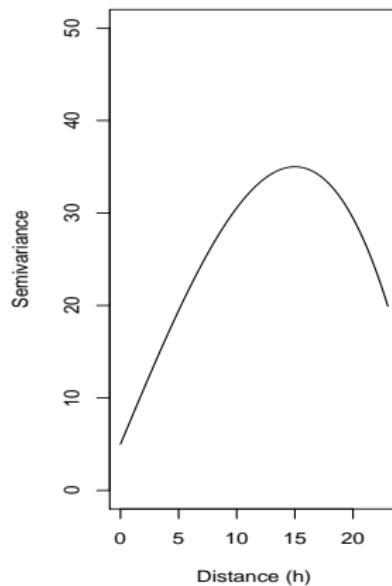
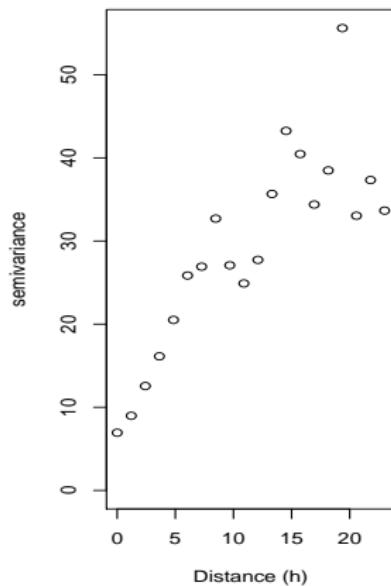
$$\gamma(h) = \tau^2 + \sigma^2(1 - \exp(-h/\rho))$$



# Geostatistical Data: Empirical vs Theoretical

Spherical:

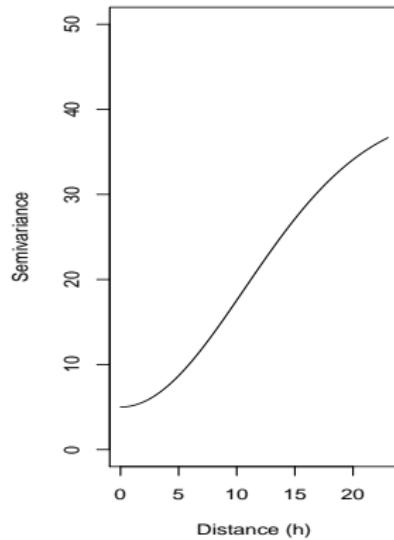
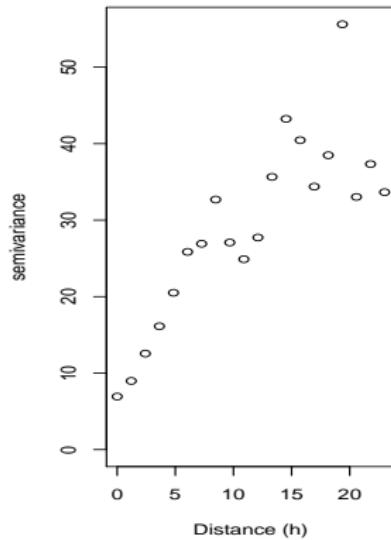
$$\gamma(h) = \tau^2 + \sigma^2(3/2(h/\rho) - 1/2(h/\rho)^3)$$



# Geostatistical Data: Empirical vs Theoretical

Gaussian:

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



## 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  $\theta$

# Geostatistical Data: Fitting a Semivariogram Model

We use the binned semivariogram,

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

- ▶ The relationship between  $\hat{\gamma}(h)$  and  $h$  is nonlinear (semivariogram model is not a linear function)
- ▶ Use generalized least squares, solved numerically
- ▶ Minimize SSE  $\sum_j^K [\hat{\gamma}(h_j) - \gamma(h_j)]^2$
- ▶ K bins from our empirical semivariogram

# Geostatistical Data: Fitting a Semivariogram Model

In Generalized Least Squares (GLS) we introduce the covariance matrix,  $V$

$$(\hat{\gamma}(h) - \gamma(h, \theta))^T V(h, \theta)^{-1} (\hat{\gamma}(h) - \gamma(h, \theta))$$

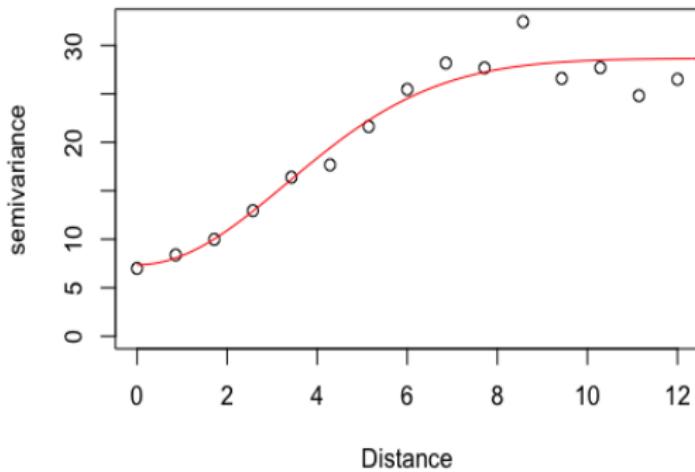
- ▶ Correlation among bins is accounted for with  $V(h, \theta)^{-1}$
- ▶ Difficult to calculate this since  $\theta$  are unknown, computationally intensive
- ▶ Use approximation and weighted least squares which accounts for unequal variance of bins (Cressie 1985)
- ▶ WLS still does not account for correlation, but is better than OLS as it gives more weight to bins having more data

# Geostatistical Data: Fitting a Semivariogram Model

- ▶  $V(h; \theta)^{-1} = I$  gives the OLS equation
- ▶ Taking  $V(h; \theta)^{-1} = \text{diag } Var[\hat{\gamma}(h_1)], \dots, Var[\hat{\gamma}(h_K)]$  gives a weighted least squares estimator
- ▶  $Var[\hat{\gamma}(h_j)] \approx 2[\hat{\gamma}(h_j)]^2/N(h_j)$
- ▶ now we minimize WSSE  $\frac{1}{2} \sum_j^K \frac{N(h_j)}{\hat{\gamma}(h_j)} [\hat{\gamma}(h_j) - \gamma(h_j)]^2$

# Geostatistical Data: Fitting a Semivariogram Model

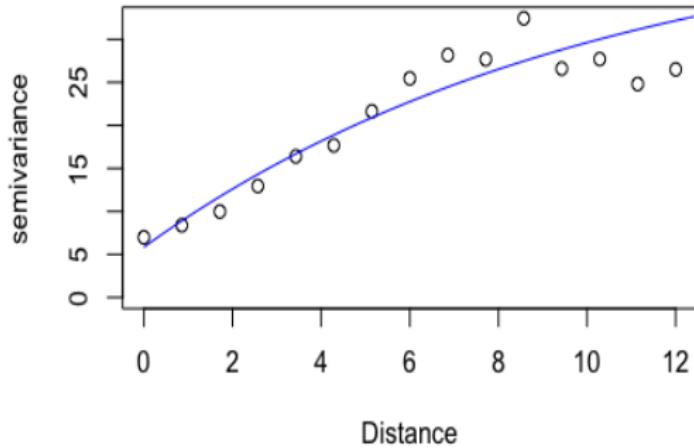
Result of fitting a Gaussian semivariogram by WLS to binned empirical semivariogram. In R use `variofit()` with `weights="cressie"`



Estimated parameters:  $\hat{\sigma}^2 = 21.2972$ ,  $\hat{\phi} = 4.6864$ ,  $\hat{\tau}^2 = 7.3589$ ; SSE: 253.5

# Geostatistical Data: Fitting a Semivariogram Model

Result of fitting an exponential semivariogram by WLS to binned empirical semivariogram. In R use `variofit()` with `weights="cressie"`



Estimated parameters:  $\hat{\sigma}^2 = 38.2043$ ,  $\hat{\phi} = 10.2684$ ,  $\hat{\tau}^2 = 5.8442$ ; SSE: 479.8

## Geostatistical Data: Fitting a Semivariogram Model

- ▶ There is extensive literature on fitting semivariograms, but approach is somewhat arbitrary and unsatisfying statistically.
- ▶ The objective that is minimized (deviation of empirical semivariogram values from semivariogram model) is based on pseudo-data.
- ▶ Fitting is basically just curve fitting and is sensitive to the binning and maximum distance chosen.
- ▶ Calculations based on semivariograms are fast, even with many observations.
- ▶ Semivariogram modeling is not based on a particular probability model for the data, so it may be more resistant to violations of assumptions.
- ▶ Compare SSE different models fit by one method (i.e. exponential vs spherical WLS). Don't compare SSE from WLS and OLS.

# Geostatistical Data: Fitting a Covariance Model

## Fitting covariance models by maximum likelihood

- ▶ The more standard statistical approach is to fit a covariance model by maximum likelihood (ML)
- ▶ ML is the most common approach in statistics to fit models by estimating parameter values
- ▶ Recall, in linear regression with normal (uncorrelated) errors, least squares is the same as maximum likelihood estimation.
- ▶ ML requires specification of a probability model (the likelihood) for the data
- ▶ Likelihoods involve unknown parameters that must be estimated from data
- ▶ Our spatial data must follow a multivariate Gaussian distribution and have second-order stationarity (Covariance exists)

# Geostatistical Data: Fitting a Covariance Model

## Fitting covariance models by maximum likelihood

- ▶ Gaussian process: MVN, mean =  $\mu$ , Cov =  $\sum(\theta)$
- ▶ Can think of this as spatial regression:  $E(Z(s)) = \mu + \epsilon(s) = X\beta + \epsilon(s)$
- ▶ Where  $\epsilon(s) \sim N(0, \sum(\theta))$
- ▶ The log-likelihood function has the form:  
$$L(\beta, \theta; Z) = -\frac{1}{2} \log |\sum(\theta)| - \frac{1}{2}(Z - X\beta)^T \sum(\theta)^{-1}(Z - X\beta)$$
- ▶ Restricted maximum likelihood (REML) is an alternative and is based on maximizing the likelihood when the data are differences

# Geostatistical Data: Fitting a Covariance Model

## Fitting covariance models by maximum likelihood

- ▶ The goal is to maximize the probability of the data relative to different parameter values
- ▶ The parameter values are treated as unknown and the data as fixed, and the parameter values that give the highest likelihood are chosen
- ▶ ML/REML is done in this case by numerical methods (there is no closed form solution) and can be intensive for large datasets (more than a few hundred observations)

# Geostatistical Data: Choosing a Model

## Choosing among models fit by ML

- ▶ The traditional way is to use Akaike's Information Criterion, which in its general form minimizes:
- ▶  $AIC = 2 \log(\text{maximized likelihood}) + 2(\text{number of parameters})$
- ▶ AIC can be used for non-nested models. It compares the likelihoods of different models and penalizes models with more parameters:
- ▶ Models with smaller AIC are favoured

# Geostatistical Data: Kriging

- ▶ Kriging is the spatial prediction of our process at unobserved locations
- ▶ Based on the fitted covariance function and the spatial regression model  
$$E(Z(s)) = \mu + \epsilon(s) = X\beta + \epsilon(s)$$
- ▶ Objective: To estimate the value of  $Z(s)$  at one or more unsampled locations in our region  $D$  based on our observed samples  $z(s_1), z(s_2), \dots, z(s_n)$

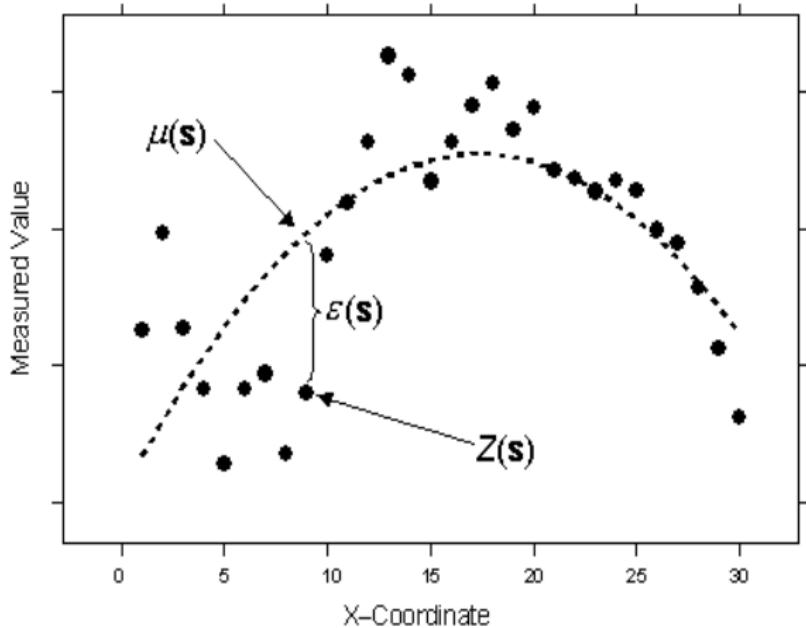
# Geostatistical Data: Kriging

- ▶ The basic kriging recipe:
  1. Choose a parametric model for the semivariogram or covariance function
  2. Estimate the semivariogram/covariance parameters.
  3. Make predictions and uncertainty estimates given the parameter estimates.
- ▶ The kriging predictions are weighted averages of the observations. The covariance/semivariogram indicates the strength of spatial association and determines the weighting.
- ▶ The issue is how heavily to weight the observations based on distance from the location.

# Geostatistical Data: Kriging

- ▶ The kriging predictor at new location  $s_0$  is  $\hat{Z}(s_0) = \sum \lambda_i Z(s_i)$
- ▶ Goal: to minimize squared error loss  $E[(\hat{Z}(s_0) - Z(s_0))^2]$
- ▶ The best prediction of this is the conditional mean:  $E(Z(s_0)|Z)$  which is the expected value of what you don't know given what you do know.
- ▶ This calculation assumes you know the covariance function (or have estimated it).
- ▶ In R we use the `krig.conv()` function in the `geoR` library.

# Geostatistical Data: Kriging

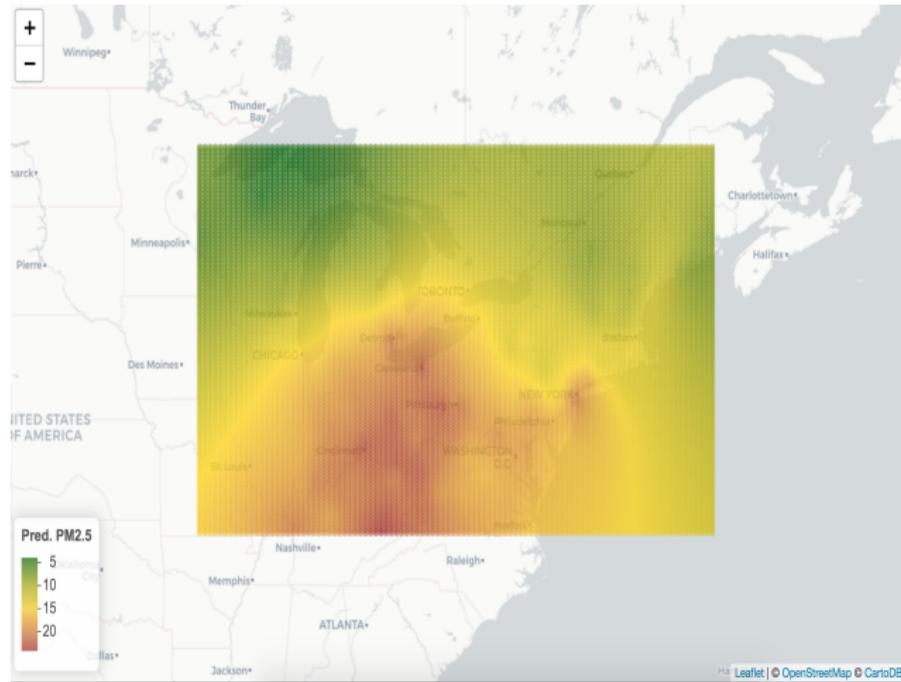


# Geostatistical Data: Kriging Types

There are different kriging types for different assumptions and analytical goals.  
The types are:

- ▶ Simple Kriging: assumes a constant known mean  $\mu = c$ . This type is not often used because for unbiasedness constraint to be applicable in kriging equations, we must estimate the expected value.
- ▶ Ordinary Kriging: assumes a constant unknown mean (mean needs to be estimated)  $Z(s) = \mu + \epsilon(s)$ .
- ▶ Universal Kriging: assumes a trend in  $x$  and  $y$ , and may include other spatially varying covariates  $Z(s) = \mu(s) + \epsilon(s)$  where  $\mu(s) = \sum_{k=1}^p \beta_k x_k(s_i)$ .

# Geostatistical Data: Ordinary Kriging Result



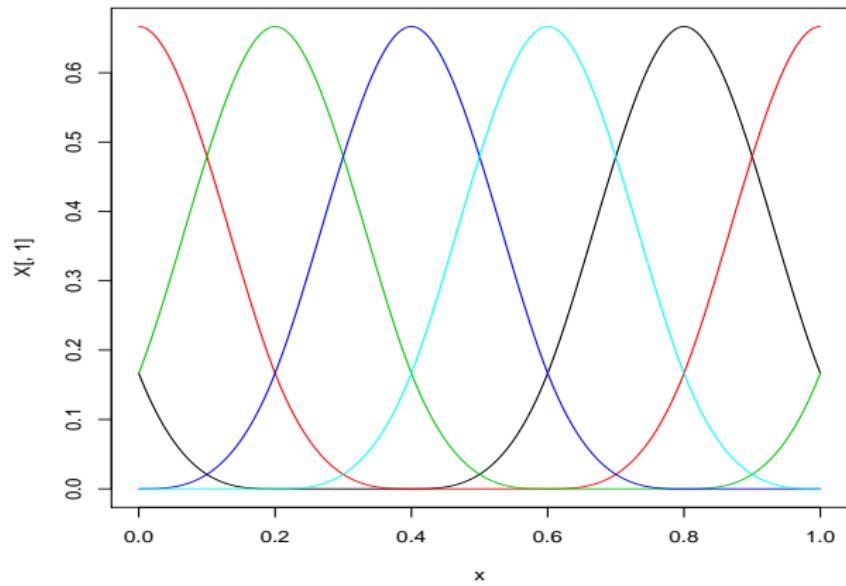
# Geostatistical Data: Smoothing

- ▶ Here we focus on modeling the spatial component in the mean part of a regression model rather than in the covariance
- ▶ In general,  $Z(s) = f(s) + \epsilon$  where errors are not correlated
- ▶  $f(s)$  is a "smooth" function described by a basis function, and consists of non-linear terms
- ▶ This is considered statistical smoothing because we can get an estimate of the prediction errors

# Geostatistical Data: Splines

- ▶ In general, splines are curves that are formed by combining pieces of a polynomial.
- ▶ There are several types of splines including natural, cubic, and b-splines (the b stands for basis).
- ▶  $f(t_i) = \sum_{j=1}^4 t^j \beta_j$
- ▶ B-spline curves are made up of polynomial pieces and are defined by a set of knots
- ▶ Choosing the number of knots defines how smooth (few) or wiggly (many) your functions

# Geostatistical Data: Splines



## Geostatistical Data: Splines

- ▶ Smoothing splines with penalty allows us to estimate where to put the knots by penalizing the wigginess of the function
- ▶ Minimize the function  $\sum_i (y_i - f(t_i))^2 + \lambda \int f''(t)^2 dt$
- ▶ Here,  $\lambda$  is a penalty parameter that controls how much to penalize wiggly functions
- ▶ Tradeoff between the goodness of fit (the sum of squares) and the wigginess of the function (the integral)
- ▶ Start by putting a knot at every data point, then penalize
- ▶ It is an optimization problem where we minimize:

$$\sum_i (y_i - B_i^T \beta)^2 + \lambda \beta^T S \beta$$

- ▶ the matrix  $S$  is constructed by using the spline basis we chose,  $B$  is the basis matrix

## Geostatistical Data: 2-D Splines

- ▶ Thin plate splines are smoothing splines in 2-d
- ▶ Extend the 1-d case to:

$$\sum_i (z_i - g(s_1, s_2))^2 + \iint g''(s_1, s_2)^2 ds_1 ds_2$$

- ▶ The thin plate spline can be implemented as an extension to the linear model in what's called a generalized additive model (GAM).

## Geostatistical Data: 2-D Splines

- ▶ Can use the spline model to predict as we did with kriging.
- ▶ Also useful as we can generate standard errors for our predictions.
- ▶ Implemented in R using the `gam()` and `predict.gam()` functions in the `mgcv` library.

# Geostatistical Data: 2-D Splines

