Spatial Statistics

PM569 Lecture 3: Geostatistical Data 2

Meredith Franklin

Division of Biostatistics, University of Southern California

September 13th, 2019

Geostatistical Data

In today's lecture we'll cover:

- Review/recap of stationarity and the semivariogram
- More on anisotropy: types
- Fitting semivariogram model to data (theoretical semivariogram to empirical semivariogram)
- Covariance functions
- Introduction to kriging

Geostatistical Data: Stationarity

We assume that our spatial process has a mean $E[Z(s)] = \mu(s)$ and the variance of $Z(s)$ exists.

- **Strong (also called strict):** the joint density is invariant under translation:
- this is often too restrictive for spatial applications

$$\begin{aligned} &P(Z(s_1) \leq z_1, Z(s_2) \leq z_2, ..., Z(s_n) \leq (z_n) \\ = &P(Z(s_1 + h) \leq z_1, Z(s_2 + h) \leq z_2, ..., Z(s_n + h) \leq z_n) \end{aligned}$$

Geostatistical Data: Stationarity

- **Weak (also called second order):** the moments (mean, variance) of the joint density are invariant. It has a constant mean $E[Z(s)] = \mu$ and covariance for all s is dependent only on distance $Cov[Z(s+h), Z(s)] = C(h)$ and we are interested in characterizing the covariance function. At h=0, we get $Cov[Z(s+0), Z(s)] = C(0) = Var[Z(s)]$

Geostatistical Data: Stationarity

- **Intrinsic:** when the difference (i.e. $Z(s+h) - Z(s)$) is second order stationary. $E[Z(s+h) - Z(s)] = 0$ Then we can write $Var[Z(s+h) - Z(s)] = 2\gamma[(s+h) - s] = 2\gamma(h)$
- second order stationarity implies intrinsic stationarity but the reverse is not implied
- note the similarity between intrinsic stationarity and second order stationarity: intrinsic is defined in terms of the variogram and second order is defined in terms of the covariance function
- the variogram is a generalization of the covariance function and under second order stationarity the two functions are related

Geostatistical Data: Variogram and Covariance

Relationship between semivariogram and covariance:

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

- if C(h) exists then we can get $\gamma(h)$, but can we get C(h) from $\gamma(h)$?

Geostatistical Data: Variogram and Covariance

- If $C(h) \to 0$ as $|h| \to \infty$
- Covariance goes to 0 as distance goes to infinity
- if we take the limit on both sides of $\gamma(h) = C(0) - C(h)$, get $\lim_{h \to \infty} \gamma(h) = C(0)$
- but the limit may not exist, for example in the linear semivariogram case

Geostatistical Data: Variogram and Covariance

- Example: Linear semivariogram
- $\gamma(h) = \tau^2 + \sigma^2 h$ if $h > 0; 0$ otherwise
- as $h \to \infty$ then $\gamma(h) \to \infty$
- thus is is not a second-order stationary process and $C(h)$ does not exist

Geostatistical Data: Semivariograms

Recall we also defined the empirical binned semivariogram:

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

The Cressie-Hawkins robust binned semivariogram (this is how it is defined in R):

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \frac{\{\sum_{N(h)} |Z(s_i) - Z(s_j)|^{1/2}\}}{0.457 + 0.494/N(h)}$$

Geostatistical Data: Stationarity and Anisotropy

What is a non-stationary process?

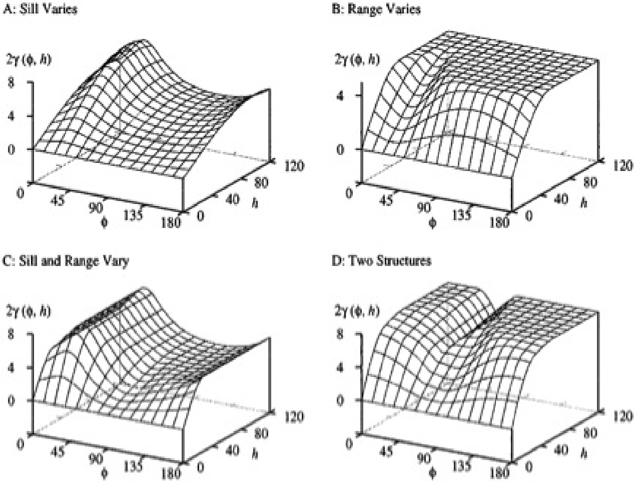
- the basic idea is that the parameters of a semivariogram model, i.e. the nugget, range, sill vary spatially.
- anisotropy may be the issue.
- tackling anisotropy and addressing general spatial trends before modeling the semivariance help deal with non-stationarity.
- recall that semivariogram/covariance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

Geostatistical Data: 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.

Geostatistical Data: Anisotropy


A: Sill Varies

B: Range Varies

C: Sill and Range Vary

D: Two Structures

Eriksson, M and P.P Siska (2000)

Geostatistical Data: 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).

Geostatistical Data: Anisotropy

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

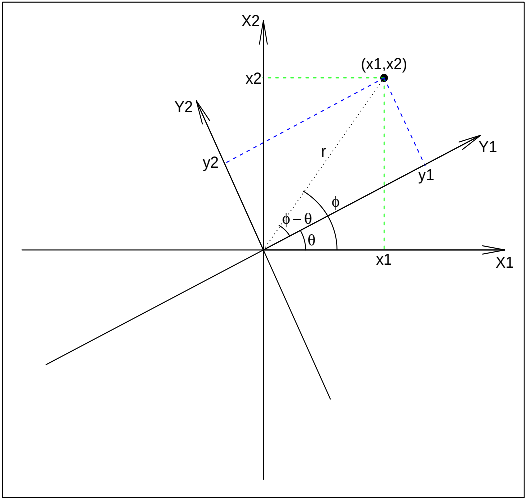
C(·) is a valid isotropic covariance function, and B is a symmetric positive definite matrix that characterizes the elliptical contours. For example the isotropic and geometric anisotropic versions of the spherical semivariogram are:

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

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

Geostatistical Data: Anisotropy

Geometric anisotropy



Geostatistical Data: Variogram and Covariance

Using the semivariogram/covariance function

- We want to apply the semivariogram function (covariance function) to make spatial predictions.
- Assumptions of prediction via kriging require that we have stationarity.
- We may look at the semivariogram but then need a covariance function, thus we must have intrinsic stationarity which is also a second-order stationary process.

We need to fit the theoretical semivariogram to our empirical semivariogram (data) to get the parameter estimates of our spatial function (nugget, sill, range).

Geostatistical Data: Fitting a Semivariogram Model

- 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.
- 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 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)

Geostatistical Data: Fitting a Semivariogram Model

- We can fit the theoretical semivariogram to the data in a variety of ways: ordinary least squares (OLS), weighted least squares (WLS), maximum likelihood (ML), and restricted maximum likelihood (REML).
- 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.

Geostatistical Data: Fitting a Semivariogram Model

In ordinary least squares (OLS), the objective is to find the set of parameters $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3)$ that minimize the sum of the squared residuals.

$$\sum_{i=1}^{n} (Y_i - f(\mathbf{X_i}, \boldsymbol{\theta}))^2$$

Where Y_i is the response variable and $f(\mathbf{X_i}, \boldsymbol{\theta})$ is a predicted value based on some model.

Geostatistical Data: Fitting a Semivariogram Model

OLS requires:

- independence, i.i.d. (but we have each observation entering into multiple bins)
- equal variance of bins, homoskedasticity (but there are different numbers of pairs in each bin and thus different variance in bins)

Solution: the usual adjustment to OLS when observations are correlated and heteroskedastic is generalized least squares (GLS)

Geostatistical Data: Fitting a Semivariogram Model

In the case of fitting a semivariogram model, we consider the sample semivariogram values as our observations and fit a model to them as a function of the distance, h.
That is, estimate the parameters $\boldsymbol{\theta} = (\tau^2, \sigma^2, \phi)$ that minimize the squared vertical distance between the empirical and theoretical semivariograms.

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

which is equivalent to

$$\sum_h (\hat{\gamma}(h) - \gamma(h, \theta))^2$$

where $\hat{\gamma}(h)$ is the empirical semivariogram and $\gamma(h, \theta)$ is the theoretical semivariogram with parameters $\boldsymbol{\theta}$

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
- the same applies to the C-H robust binned semivariogram (this version is preferred to deal with any outliers)

Geostatistical Data: Fitting a Semivariogram Model

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

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

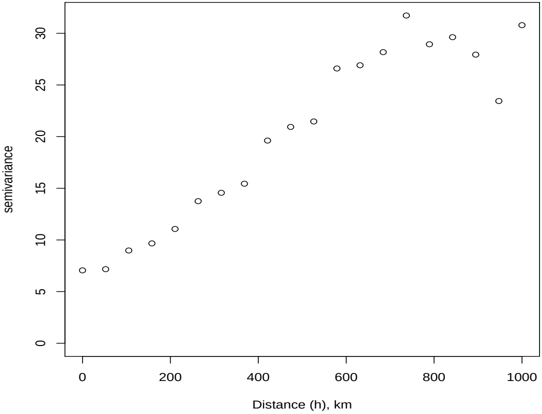
- Correlation among bins is accounted for with $V(h, \theta)^{-1}$
- Difficult to calculate this since θ are unknown, computationally intensive
- Use approximation and weighted least squares which accounts for unequal variance of bins (Cressie 1985)
- WLS still does not account for correlation, but is 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 $V(h;\theta)^{-1} =$ diag $Var[\hat{\gamma}(h_1)], ..., Var[\hat{\gamma}(h_K)]$ 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 $\frac{1}{2}\sum_j^K \frac{N(h_j)}{\hat{\gamma}(h_j)}[\hat{\gamma}(h_j) - \gamma(h_j)]^2$

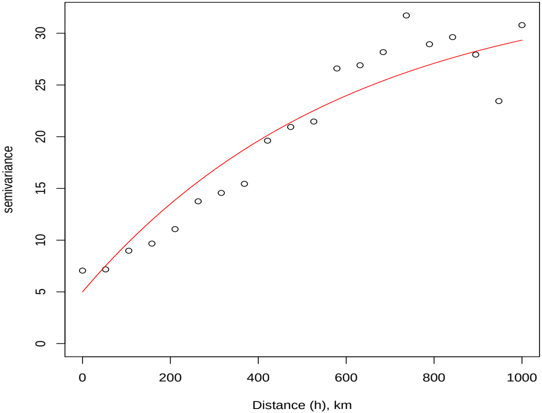
Geostatistical Data: Fitting a Semivariogram Model

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



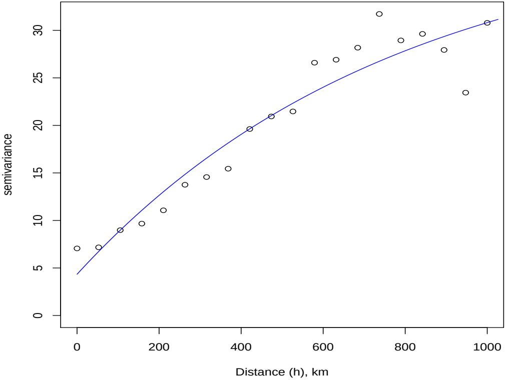
Geostatistical Data: Fitting a Semivariogram Model

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



Geostatistical Data: Fitting a Semivariogram Model

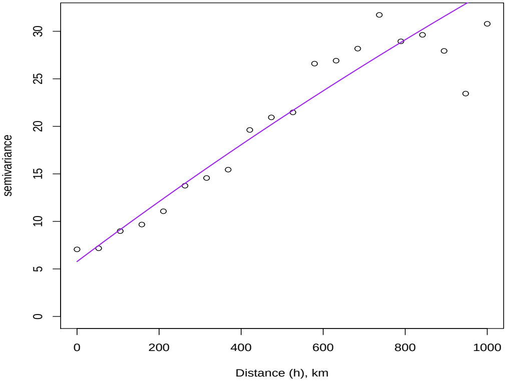
Result of fitting an exponential semivariogram by OLS to empirical semivariogram. In R use variofit() with weights="equal"



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

Geostatistical Data: Fitting a Semivariogram Model

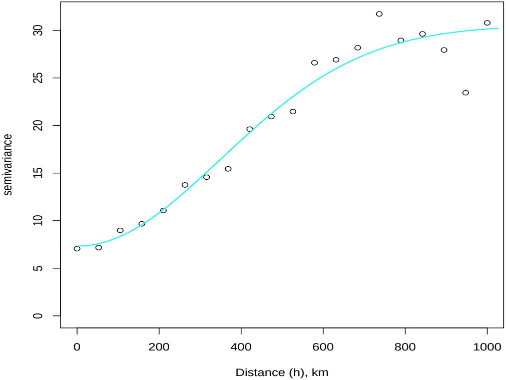
Result of fitting an exponential semivariogram by WLS to empirical semivariogram. In R use variofit() with weights="cressie"



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

Geostatistical Data: Fitting a Semivariogram Model

Result of fitting an Gaussian semivariogram by WLS to empirical semivariogram. In R use 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!

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 $= 2\log($maximized likelihood$) + 2($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 $z(s_1), z(s_2), ..., z(s_n)$

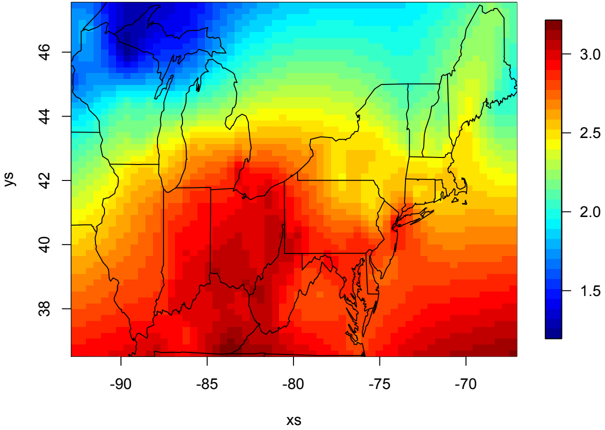
Geostatistical Data: Kriging

- 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).

Geostatistical Data: Kriging Result


ys
46
44
42
40
38
xs
-90
-85
-80
-75
-70
3.0
2.5
2.0
1.5