Spatial Statistics
Geostatistics (Point Reference Data) Unit 3

PM569 Spatial Statistics

Lecture 4: September 18, 2015

Geostatistical Data

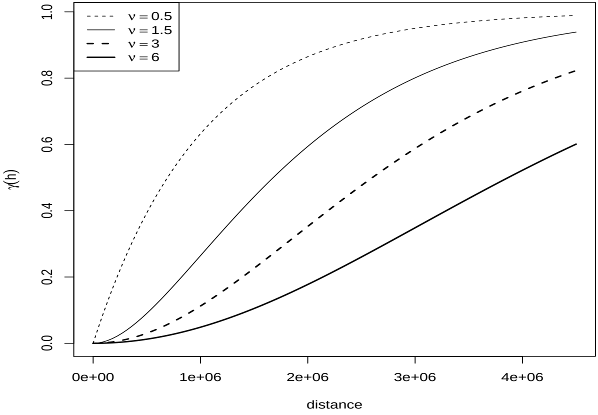
- Matern semivariogram/covariance function
- Kriging
- Spatial Regression
- Spatial prediction interpolation and smoothing methods

- Commonly used because of its flexibility
- $\gamma(h) = \tau^2 + \sigma^2[1 - \frac{(2\sqrt{\nu}h)^\nu}{2^{\nu-1}\rho^\nu\Gamma(\nu)} K_\nu(\frac{2\sqrt{\nu}h}{\rho})]$
- $C(h) = \sigma^2[\frac{(2\sqrt{\nu}h)^\nu}{2^{\nu-1}\rho^\nu\Gamma(\nu)} K_\nu(\frac{2\sqrt{\nu}h}{\rho})]$
- ν controls the smoothness, $\Gamma(\nu)$ is the gamma function, K_ν is the modified Bessel function of order ν

- Some properties:
- when $\nu = 0.5$ the Matern is equivalent to the exponential semivariogram/covariance
- when $\nu \to \infty$ the Matern is equivalent to the Gaussian semivariogram/covariance

Geostatistical Data

Matern semivariogram/covariance function



- Technique for spatial prediction
- Use estimated spatial process to predict values 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 $z(s_1), z(s_2), ..., z(s_n)$

- 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.

Best linear unbiased predictor (BLUP)

- The kriging predictor at any new location s_0 is the BLUP
- The prediction variance $E[\hat{Z}(s_0) - Z(s_0)]^2$ is minimized ("best")
- It is a linear prediction based on a weighted average of the observations ("linear")
- The expected value of the prediction $E[\hat{Z}(s_0)]$ is equal to the expected value at s_0 ("unbiased")
- Unbiased condition: $E[\hat{Z}(s_0)] = E[Z(s_0)]$ also implies that $\sum_i \lambda_i = 1$
- $\sum_i \lambda_i = 1$ ensures non-negative and finite variance

- Goal: to minimize squared error $E[(\hat{Z}(s_0) - Z(s_0))^2]$ subject to the unbiasedness constraint $\sum_{i=1}^{n} \lambda_i = 1$
- The problem boils down to finding the set of coefficients $\lambda_1, ..., \lambda_n$ that serve as the weights of our observations
- The Lagrange multiplier method is used for finding the minimum subject to this constraint
- This calculation assumes you know the moments, i.e. the semivariance/covariance function (or have estimated it) $\gamma(h)$ or $C(h)$ and the mean $\mu(s)$

- $\hat{Z}(s_0) = \sum_{i=1}^{N} \lambda_i Z(s_i)$
- $E[(\hat{Z}(s_0)] = \mu = E[Z(s_0)]$
- with constraint $\sum_{i=1}^{N} \lambda_i = 1$
- Minimize $E[(\hat{Z}(s_0) - Z(s_0))^2]$
- See handout

- Different kriging types for different assumptions and analytical goals
- Simple Kriging: assumes a known mean $\mu = c$
- 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 $Z(s) = \mu(s) + \epsilon(s)$

From the Lagrange multiplier approach, the kriging equations are $\sum_j \lambda_j C_{ij} + m = C_{i0}$ where C_{i0} and C_{ij} are evaluated based on our data and chosen covariance function

$$\begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_N \\ m \end{pmatrix} = \begin{bmatrix} C_{11} & C_{12} & \dots & C_{1N} & 1 \\ C_{21} & C_{22} & \dots & C_{2N} & 1 \\ \vdots & \vdots & \dots & \vdots & \vdots \\ C_{N1} & C_{N2} & \dots & C_{NN} & 1 \\ 1 & 1 & \dots & 1 & 0 \end{bmatrix}^{-1} \times \begin{pmatrix} C_{10} \\ C_{20} \\ \vdots \\ C_{N0} \\ 1 \end{pmatrix}$$

We have a linear kriging system of n+1 equations with n+1 unknowns

- Variance of the prediction $\hat{Z}(s_0)$ is important as a measure of uncertainty in our predictions
- From the Lagrange multiplier result, $Var(\hat{Z}(s_0)) = \sigma^2 - \lambda' C_{i0}$

- In addition to estimating the parameters of the spatially varying covariates, we have spatial correlation in the data that is not explained by the covariates.
- $Z(s) \sim MVN(X(s)\beta, \Sigma)$
- The term $X(s)\beta$ is referred to as the mean structure (often describing large scale variation or trend) and is distinguished from Σ which is the residual small-scale variation.
- The residual process adjusts the model for any residual spatial variation after accounting for the covariates.

- It is possible that you can account for all spatial correlation in your response with spatially varying covariates. More often than not there remains inherent spatial correlation that cannot be explained, so we need to estimate Σ.
- There can be spatial confounding if your covariates vary spatially at the same scale. Beware of this.

- Recall ordinary least squares, $\hat{\beta}_{OLS} = (X'X)^{-1}X'Y$
- Note: using $Y(s_i)$ here for regression versus $Z(s_i)$ in kriging equations, but they both represent the spatial process we wish to model.
- The variance $\hat{\sigma}^2_{OLS} = \frac{(Y-X\hat{\beta}_{OLS})'(Y-X\hat{\beta}_{OLS})}{N-p}$
- The OLS estimators of our regression parameters are unbiased (and the confidence intervals on the estimates are correct) when the model is correctly specified. Our covariates correctly specify the model and the previous assumptions are met.

- The variance of our our parameter estimates $Var(\hat{\beta}_{OLS}) = \hat{\sigma}^2_{OLS}(X'X)^{-1}$
- Our variance-covariance matrix is $\Sigma = \sigma^2\mathbf{I}$
- There no covariance between errors
- These are all incorrect if there is residual spatial correlation unaccounted for in covariates. Our estimates of β_{OLS} will be biased

- There is covariance between errors, i.e. $Cov(Y(s_i), Y(s_j))$
- Our variance-covariance matrix is $\Sigma = var(\epsilon) = \sigma^2\mathbf{V}$
- We need maximum likelihood methods to generate estimates $\hat{\beta}_{GLS}$
- $\hat{\beta}_{GLS} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y$
- $\hat{\sigma}^2_{GLS} = \frac{(Y-X\hat{\beta}_{GLS})'\Sigma^{-1}(Y-X\hat{\beta}_{GLS})}{N-p}$

- The variance of our our parameter estimates
  $\hat{V}ar(\hat{\beta}_{GLS}) = (X'\Sigma^{-1}X)^{-1}$
- Note, if we used OLS when there was spatial correlation, we would get the wrong estimates of uncertainty in our parameter estimates.
- The GLS estimator is the minimum variance linear unbiased estimator so it always has lower variance then the OLS estimator.

- We want to use GLS but the problem is we don't know Σ
- As before, we need to parameterize our spatial covariance, but we don't know the parameters
- We need to estimate βs and Σ

- There are two methods for approaching this problem:
  1. Iteratively Reweighted Generalized least squares
  2. Maximum Likelihood (or Restricted Maximum Likelihood)

Iteratively Reweighted Generalized least squares

- STEPS:
  1. Obtain a starting estimate of $\hat{\beta}$ from OLS
  2. Compute the residuals $r = Y - X\hat{\beta}$
  3. Estimate the semivariogram parameters based on r. This gives us $\Sigma(\hat{\theta})$
  4. Obtain a new estimate of $\hat{\beta}$ using GLS with $\Sigma(\hat{\theta})$
  5. Iterate steps 2-4 until the estimates of $\hat{\beta}$ are small (converge, $< 10^{-5}$

ML and REML

- The probability model for the data is $Y \sim MVN(X\beta, \Sigma)$
- Maximize likelihood
  $L(\beta, \Sigma) = -log|\Sigma|^{1/2} - (1/2)(Y - X\beta)^T\Sigma^{-1}(Y - X\beta)$
- Our covariance matrix $\Sigma_{ij} = Cov(s_i, s_j)$ is spatial as before, with parameters σ^2, τ^2, ϕ
- This cannot be solved analytically so numerical methods are implemented (e.g. Newton Raphson)

Universal Kriging

- Before, in ordinary kriging, we assumed we had a constant mean $E(Z(s)) = \mu$
- We develop our BLUP using the Lagrange multiplier approach as before
- Multiple constraint problem
- See handout for explanation

- Choice between detrending data and then using ordinary kriging versus implementing universal kriging
- Detrending means using GLS to estimate $\hat{\beta}_{GLS}$
- Then $\hat{Z}(s_0) = x(s_0)\hat{\beta}_{GLS} + \hat{\epsilon}(s_0)_{OK}$
- This should give the same predictions as universal kriging as long as the residuals are estimated using GLS and the covariance function for OK and UK is the same
- HOWEVER, the prediction errors are NOT the same.
- The kriging variance with detrended data does not reflect the uncertainty in estimating $\hat{\beta}_{GLS}$

Other Types of Kriging

- We have gone through the primary types of spatial prediction via the covariance function
  - Simple kriging (known mean), ordinary kriging (unknown mean, estimated), universal kriging (unknown trend, covariates estimated)
- Spatial prediction using the above techniques can be seen as point interpolation, taking a set of points as inputs and enabling prediction on a grid (raster)
- Spatial prediction may also be done at a specific point or set of points if desired
- There are additional types of kriging:
  - Cokriging
  - Indicator kriging

- Cokriging is a multivariate version of kriging
- Two interrelated variables measured at $\mathbf{s}$, harness the correlation between the two
- Often used when one variable of interest is measured more sparsely than another and it is desirable to predict the sparse variable
- Helps minimize the variance of the estimation error by exploiting the cross-correlation between variables because cross-correlated information contained in the secondary variable should help reduce the variance of the estimation errors

- Cross-variograms

$$2\nu_{1,2}(h) = Cov(Z_1(s+h) - Z_1(s), Z_2(s+h) - Z_2(s))$$
$$2\gamma_{1,2} = Var(Z_1(s+h) - Z_2(s))$$

- Note that the first equation relies on Z_1 and Z_2 being measured at the same locations s and $s+h$
- The second equation does not require observations at the same locations, but only at distances h

- Cokriging enables the use of observed values of two variables of interest to predict $Z_1(s_0)$ or $Z_2(s_0)$
- We must assume the following exist for Z_1 and Z_2

$$\begin{aligned} 2\gamma_1(s_i - s_j) \\ 2\gamma_2(s_g - s_h) \\ 2\gamma_{1,2}(s_i - s_g) \end{aligned}$$

  for all s_i, s_j, s_g, s_h in the domain D
- See handout on blackboard

Geostatistical Data

Indicator Kriging

- When we have point locations but the outcome is binary, but based on an exceedance, for example $P(Z(s_0) > |Z_1, ..., Z_N)$
- It is a transformation of the original (continuous)
- It is a way to make a probability map, whereby the probability is that of a certain value being exceeded

$$
\begin{aligned}
I(Z(s) > z) &= \quad 1 \quad \text{if} Z(s) > z \\
&= \quad 0 \quad \text{otherwise}
\end{aligned}
$$

- The probability can be estimated by kriging the indicator $I(Z(s_0) > z)$ from the indicator data $I(Z(s_1) > z, ..., I(Z(s_N) > z)$
- Giving the estimate $E(I(Z_0) > z|I(Z(s_i) > z)$ which is an estimate of $P(Z(s_0) > |Z_1, ..., Z_N)$

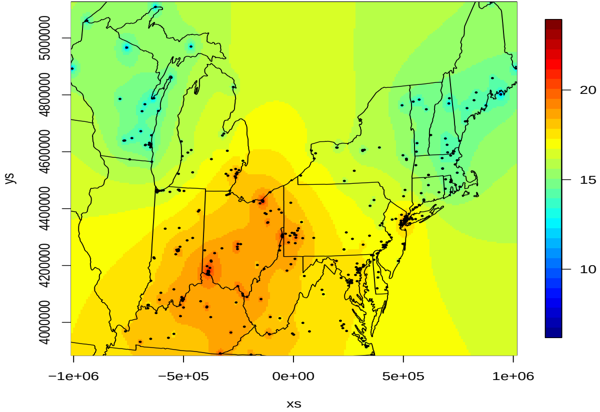
$$I_{OK}(s_0) = \sum_{i=1}^{N} \lambda_i I(Z(s_i) > z)$$

Other types of spatial prediction

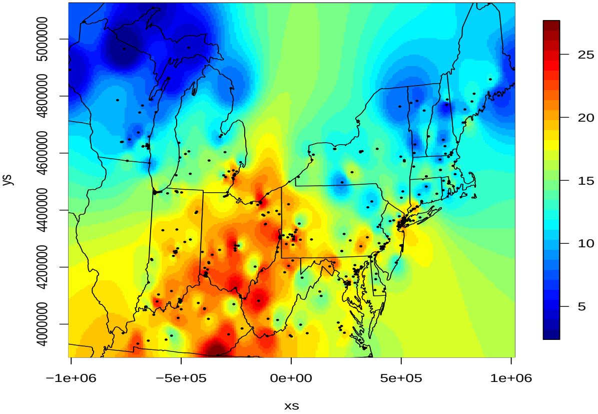
- Inverse-distance weighted interpolation
- Kernel smoothing (local weighted averaging)
- Loess smoothing (local polynomial fitting)
- Spline smoothing

- Inverse distance weighting
- Weight observations by the inverse of the distance between its location and the prediction location
- $\hat{z}_0 = \frac{\sum_i Z(s_i) d(s_i, s_0)^{-p}}{\sum_i d(s_i, s_0)^{-p}}$
- Larger p relates to a more localized weighting, and small p is smoother
- This is not considered statistical smoothing because there is no statistical estimation that gives standard errors for our predictions

p=1

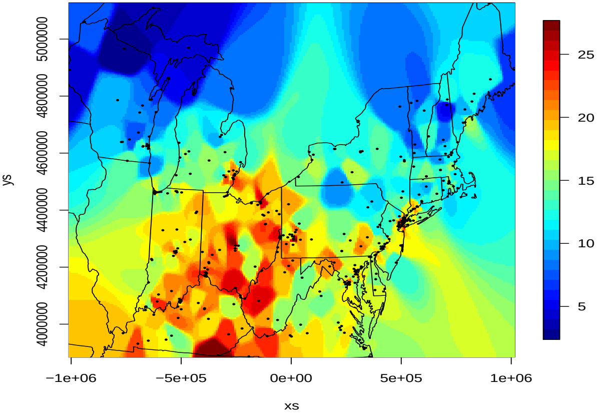

ys
5000000
4800000
4600000
4400000
4200000
4000000
−1e+06
−5e+05
0e+00
5e+05
1e+06
xs
20
15
10

p=3



ys
5000000
4800000
4600000
4400000
4200000
4000000
−1e+06
−5e+05
0e+00
5e+05
1e+06
xs
25
20
15
10
5

p=12



ys
5000000
4800000
4600000
4400000
4200000
4000000
−1e+06
−5e+05
0e+00
5e+05
1e+06
xs
25
20
15
10
5

- 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

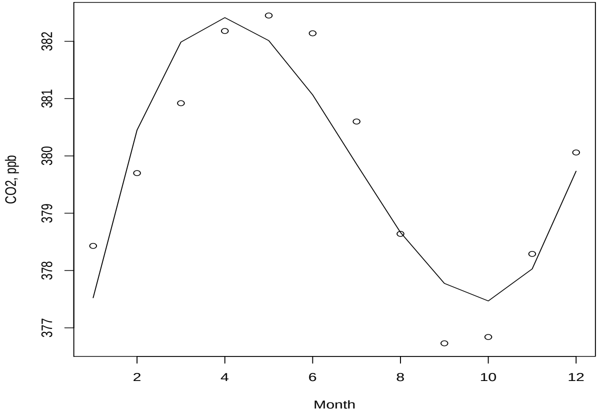
- In 1-d such as in time series, we can see how basis functions work
- Polynomial bases are a good way to illustrate what is going on
- Consider the regression model $y_i = \beta_0 + \beta_1 t_i + \beta_2 t_i^2 + \beta_3 t_i^3 + \epsilon_i$
- This is a function represented by four basis functions, so $f(t_i) = \sum_{j=1}^{4} t^j \beta_j$
- A basis is a set of functions that can be added together in a weighted fashion to form a more complicated function
- Here our weights are the regression coefficients β_j

- In general, a basis function is represented by $f_i = \sum b_j(t_i)\beta_j$

$$\begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \end{pmatrix} = \begin{bmatrix} 1 & b_1(t_1) & b_2(t_1) & b_3(t_1) & b_4(t_1) \\ 1 & b_1(t_2) & b_2(t_2) & b_3(t_2) & b_4(t_2) \\ 1 & b_1(t_3) & b_2(t_3) & b_3(t_3) & b_4(t_3) \\ 1 & b_1(t_4) & b_2(t_4) & b_3(t_4) & b_4(t_4) \end{bmatrix} \times \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix}$$

Geostatistical Data

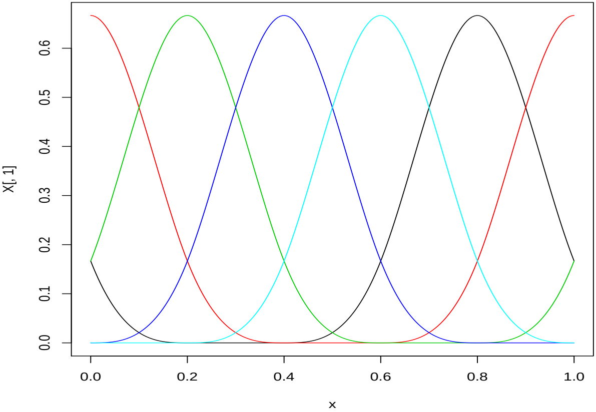
Smoothing: Polynomial Basis



Geostatistical Data

Smoothing: Regression Splines

- Regression splines fit smooth regression relationships using spline bases. There are several types of spline bases including b-splines, cubic splines
- $f(t_i) = \sum_{j=1}^{4} t^j \beta_j$
- 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


X[, 1]
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.0
0.2
0.4
0.6
0.8
1.0
x

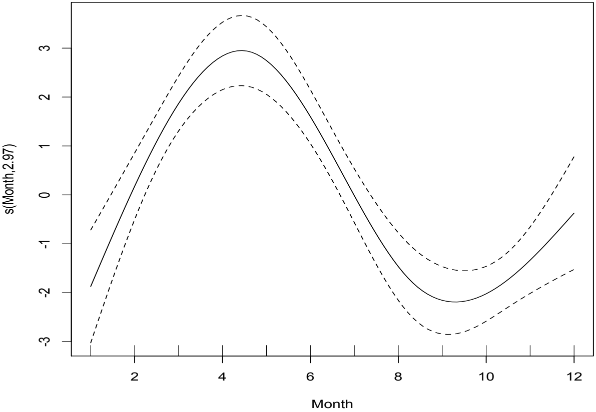
- Smoothing splines with penalty allows us to estimate where to put the knots by penalizing the wiggliness of the function
- Minimize the function $\sum_i (y_i - f(t_i))^2 + \lambda \int f''(t)^2 \mathrm{d}t$
- Here, λ is a penalty parameter that controls how much to penalize wiggly functions
- Tradeoff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

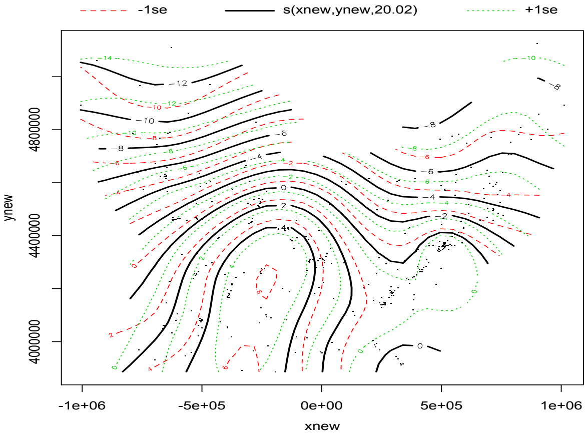
Smoothing: Regression Splines



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

$$\sum_i z_i - g(s_1, s_2))^2 + \iint g''(s_1, s_2)^2 \mathrm{d}s_1 \mathrm{d}s_2$$

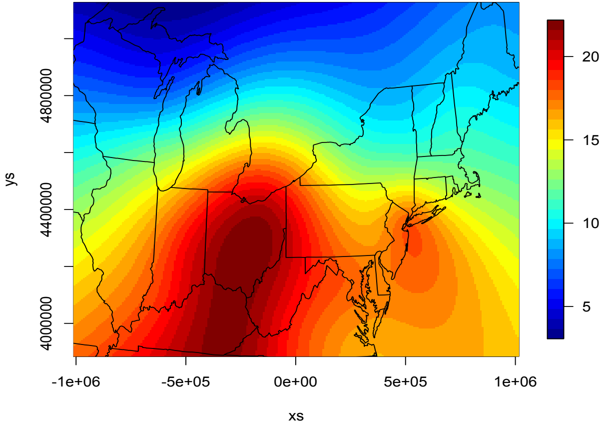
Smoothing: Thin Plate Splines



- 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

Smoothing: Thin Plate Splines



Geostatistical Data

Smoothing: Thin Plate Splines