Analysis of univariate point referenced spatial data

Abhi Datta

February 5, 2018

Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland

Review of last lecture

- Types of spatial data – point referenced, areal, point pattern
- Exploratory data analysis with point referenced data
  - Surface plots of the response, covariates and residuals
  - Empirical variograms of the residuals
- When purely covariate based models does not suffice, one needs to leverage the information from locations
  - Simple choices like adding the co-ordinates as covariates in a linear regression
  - More general model: $y(s) = x(s)'\beta + w(s) + \epsilon(s)$ for all $s \in D$
- How to choose the function $w(\cdot)$?
- Since we want to predict at any location over the entire domain D, this choice will amount to choosing a surface $w(s)$
- We will do this using Gaussian Processes

Gaussian Processes (GPs)

- The collection of random variables $\{w(s) \,|\, s \in D\}$ is a GP if
  - it is a valid stochastic process
  - all finite dimensional densities $\{w(s_1), \ldots, w(s_n)\}$ follow multivariate Gaussian distribution
- Why GPs are attractive - only need a mean function $m(s)$ and a valid covariance function $C(\cdot, \cdot)$
- Advantage: Likelihood based inference.
  $w = (w(s_1), \ldots, w(s_n))' \sim N(m, C)$ where
  $m = (m(s_1), \ldots, m(s_n))'$ and $C = (C(s_i, s_j))$
- For the model $y(s) = x(s)'\beta + w(s) + \epsilon(s)$, $x(s)'\beta$ is modeling the mean. Hence, $m(s)$ is often chosen to be 0.

Valid covariance functions and isotropy

- $C(\cdot,\cdot)$ needs to be a positive definite function
- Simplifying assumptions:
    - Stationarity: $C(s_1, s_2) = Cov(w(s_1), w(s_2))$ only depends on $h = s_1 - s_2$ (and is denoted by $C(h)$)
    - Isotropic: $C(h) = C(||h||)$ (Simplest and most interpretable)
    - Anisotropic: Stationary but not isotropic
- Exponential covariance function: $C(h) = \sigma^2 \exp(-\phi||h||)$ is a popular choice for $C(\cdot,\cdot)$

Modeling with GPs

Spatial linear model

$$y(s) = x(s)'\beta + w(s) + \epsilon(s)$$

- $w(s)$ modeled as $GP(0, C(\cdot \,|\, \theta))$ (usually without a nugget)
- $\epsilon(s) \overset{\text{iid}}{\sim} N(0, \tau^2)$ is the measurement error
- $w = (w(s_1), \ldots, w(s_n))' \sim N(0, \sigma^2 R(\phi))$ where $R(\phi) = (exp(-\phi||s_i - s_j||))$
- $y = (y(s_1), \ldots, y(s_n))' \sim N(X\beta, \sigma^2 R(\phi) + \tau^2 I)$

Parameter estimation

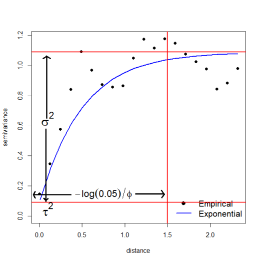
- $y = (y(s_1), \ldots, y(s_n))' \sim N(X\beta, \sigma^2 R(\phi) + \tau^2 I)$
- We can obtain maximum likelihood estimates (MLEs) of parameters $\beta, \tau^2, \sigma^2, \phi$ based on the above model
- In practice, the likelihood is often very flat with respect to the spatial covariance parameters and choice of initial values is important

Parameter estimation

- $\hat{\beta}_{init} = (X'X)^{-1}X'Y$ is often a good estimate (or initial estimate) for β
- Note that: $y(s) - x(s)'\hat{\beta}_{init} \approx w(s) + \epsilon(s)$
- If $w(s) \sim GP(0, C(\cdot,\cdot))$ and $\epsilon(s) \overset{\text{iid}}{\sim} N(0, \tau^2)$, then $w(s) + \epsilon(s) \sim GP(0, C_1(\cdot,\cdot))$ where $C_1(h) = C(h) + \tau^2 I(h = 0)$
- Initial values can be eyeballed from empirical semivariogram of the residuals $y(s) - x(s)'\hat{\beta}_{init}$

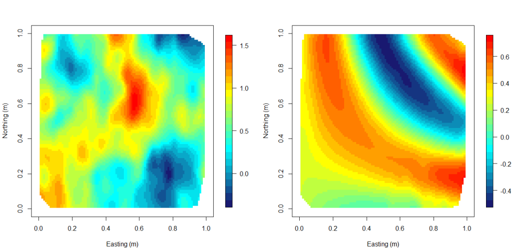
Covariance functions and semivariograms

- Recall: Empirical semivariogram: $\gamma(t_k) = \frac{1}{2|N(t_k)|} \sum_{s_i, s_j \in N(t_k)} (Y(s_i) - Y(s_j))^2$
- For any isotropic GP, $E(Y(s+h) - Y(s))^2/2 = C(0) - C(||h||) = \gamma(||h||)$
- $\gamma(||h||)$ is the semivariogram corresponding to the covariance function $C(||h||)$
- For exponential GP + measurement error, $\gamma(||h||) = \tau^2 I(||h|| > 0) + \sigma^2(1 - \exp(-\phi||h||))$



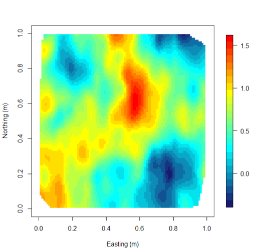
- Effective range $\approx 3/phi$, is the distance at which this correlation has dropped to only 0.05.
- The nugget τ^2 is often viewed as a "nonspatial effect variance"
- The partial sill (σ^2) is viewed as a "spatial effect variance."
- $\sigma^2 + \tau^2$ gives the maximum total variance often referred to as the sill

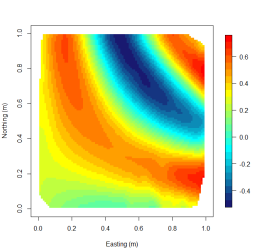
Dataset 3 from last lecture



Dataset 3: $y(s)$ Dataset 3: $x(s)$

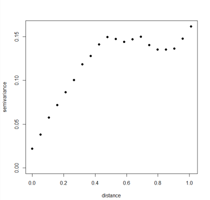
Dataset 3 from last lecture

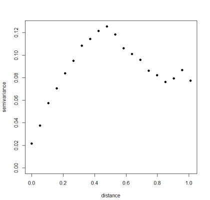




Dataset 3: $y(s)$ Dataset 3: $x(s)$

- Model 1: $y(s) = \beta_0 + \beta_1 x(s) + \epsilon(s)$
- Model 2: $y(s) = \beta_0 + \beta_1 x(s) + \beta_2 s_x + \beta_3 s_y + \epsilon(s)$





Residuals: Model 1 Residuals: Model 2

Modeling using Gaussian Process

- Model 3: $y(s) = \beta_0 + \beta_1 x(s) + w(s) + \epsilon(s)$
- $w(s) \sim GP(0, C(\cdot, \cdot))$, $C(s_i, s_j) = \sigma^2 \exp(-\phi||s_i - s_j||)$
- $\epsilon(s) \overset{\text{iid}}{\sim} N(0, \tau^2)$
- Parameters estimated using *likfit* function of *geoR* package
- Note: In *geoR* package, the ϕ is defined as the range, i.e., it is the reciprocal of our definition of ϕ

Model comparison

- $l(y \mid \beta, \theta, \tau^2)$ is the likelihood function where $\theta = (\sigma^2, \phi)'$
- For k total parameters and sample size n:
  - AIC: $2k - 2\log(l(y \mid \hat{\beta}, \hat{\theta}, \hat{\tau^2}))$
  - BIC: $\log(n)k - 2\log(l(y \mid \hat{\beta}, \hat{\theta}, \hat{\tau^2}))$

Model comparison

- $l(y \mid \beta, \theta, \tau^2)$ is the likelihood function where $\theta = (\sigma^2, \phi)'$
- For k total parameters and sample size n:
  - AIC: $2k - 2\log(l(y \mid \hat{\beta}, \hat{\theta}, \hat{\tau^2}))$
  - BIC: $\log(n)k - 2\log(l(y \mid \hat{\beta}, \hat{\theta}, \hat{\tau^2}))$

Table: Model comparison

	Model 1	Model 3
AIC	402	-208
BIC	415	-187

Conditional normal distribution

- Let $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix})$
- Then $X_1 \mid X_2 \sim N(\mu_{1|2}, \Sigma_{1|2})$
- $\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2)$ is the conditional mean
- $\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ is the conditional variance
- $\mu_{1|2}$ is the 'best' (minimum variance) predictor of X_1 based on X_2

Kriging: Spatial prediction at new locations

- Goal: Given observations $w = (w(s_1), w(s_2), \ldots, w(s_n))'$, predict $w(s_0)$ for a new location s_0
- If $w(s)$ is modeled as a GP, then $(w(s_0), w(s_1), \ldots, w(s_n))'$ jointly follow multivariate normal distribution
- $w(s_0) \,|\, w$ follows a normal distribution with
  - Mean (kriging estimator): $m(s_0) + c'C^{-1}(w - m)$
  - where $m = E(w) = (m(s_1), \ldots, m(s_n))'$,
    $C = Cov(w) = \sigma^2(C(s_i, s_j|\theta))$ and
    $c = Cov(w, w(s_0)) = (C(s_1, s_0|\theta), \ldots, C(s_n, s_0|\theta))'$
  - Variance: $C(s_0, s_0) - c'C^{-1}c$
- The GP formulation gives the full predictive distribution of $w(s_0)|w$

Kriging: Spatial prediction at new locations

- What is the kriging estimator when $s_0 = s_i$ for some i ?

Kriging: Spatial prediction at new locations

- What is the kriging estimator when $s_0 = s_i$ for some i ?
- $c = (C(s_1, s_i|\theta), \ldots, c(s_i, s_i|\theta), \ldots, c(s_n, s_i|\theta)'$
- $C = \begin{pmatrix} C(s_1, s_1|\theta) & \ldots & C(s_1, s_i|\theta) & \ldots & C(s_1, s_n|\theta) \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ C(s_n, s_1|\theta) & \ldots & C(s_n, s_i|\theta) & \ldots & C(s_n, s_n|\theta) \end{pmatrix}$
- c is the i^{th} column of C. Hence $C^{-1}c = (0, \ldots, 0, 1, 0, \ldots, 0)'$

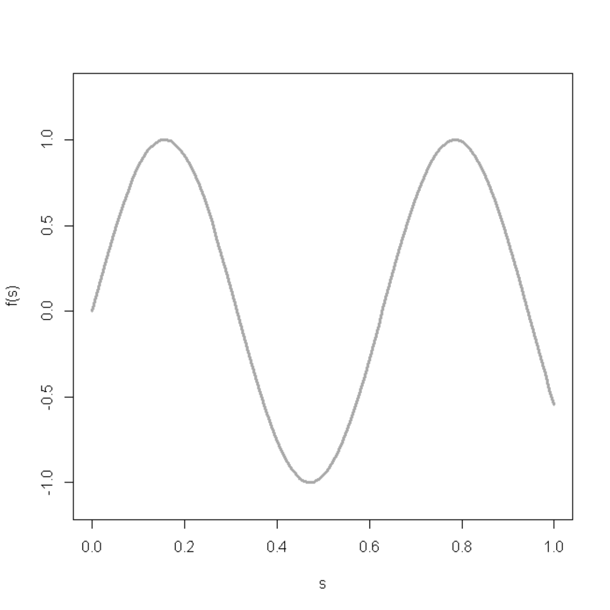
Kriging: Spatial prediction at new locations

- What is the kriging estimator when $s_0 = s_i$ for some i ?
- $c = (C(s_1, s_i|\theta), \ldots, c(s_i, s_i|\theta), \ldots, c(s_n, s_i|\theta)'$
- $C = \begin{pmatrix} C(s_1, s_1|\theta) & \ldots & C(s_1, s_i|\theta) & \ldots & C(s_1, s_n|\theta) \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ C(s_n, s_1|\theta) & \ldots & C(s_n, s_i|\theta) & \ldots & C(s_n, s_n|\theta) \end{pmatrix}$
- c is the i^{th} column of C. Hence $C^{-1}c = (0, \ldots, 0, 1, 0, \ldots, 0)'$
- Kriging mean: $m(s_i) + c'C^{-1}(w - m) = m(s_i) + w(s_i) - m(s_i)$
- Kriging variance: $C(s_i, s_i) - c'C^{-1}c = 0$

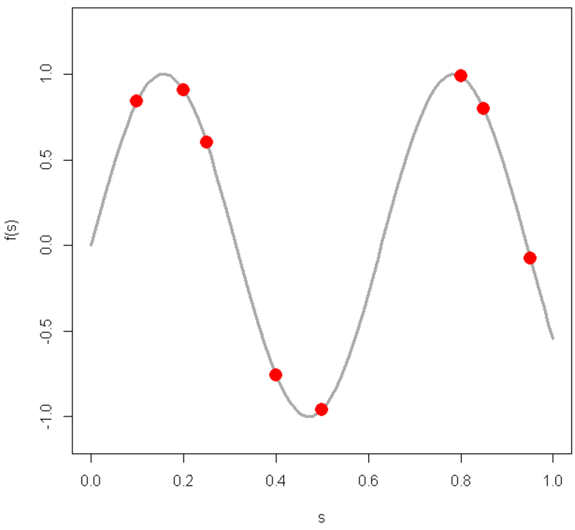
Kriging: Spatial prediction at new locations

- What is the kriging estimator when $s_0 = s_i$ for some i ?
- $c = (C(s_1, s_i|\theta), \ldots, c(s_i, s_i|\theta), \ldots, c(s_n, s_i|\theta)'$
- $C = \begin{pmatrix} C(s_1, s_1|\theta) & \ldots & C(s_1, s_i|\theta) & \ldots & C(s_1, s_n|\theta) \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ C(s_n, s_1|\theta) & \ldots & C(s_n, s_i|\theta) & \ldots & C(s_n, s_n|\theta) \end{pmatrix}$
- c is the i^{th} column of C. Hence $C^{-1}c = (0, \ldots, 0, 1, 0, \ldots, 0)'$
- Kriging mean: $m(s_i) + c'C^{-1}(w - m) = m(s_i) + w(s_i) - m(s_i)$
- Kriging variance: $C(s_i, s_i) - c'C^{-1}c = 0$
- Kriging predictions at the data locations are the observed values themselves with prediction variance equaling zero
- Kriging interpolates !

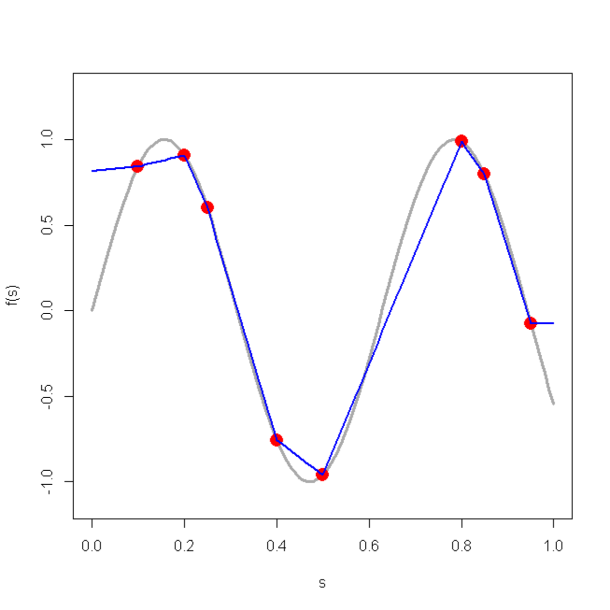
Kriging is an interpolator



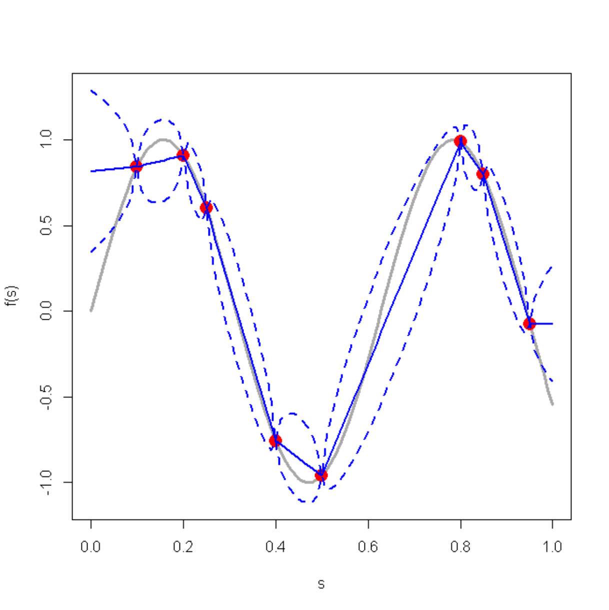
Kriging is an interpolator



Kriging is an interpolator



Kriging is an interpolator



Kriging: Spatial prediction at new locations

- What happens when s_0 is far away from all the s_i's ?

Kriging: Spatial prediction at new locations

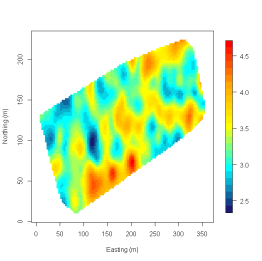
- What happens when s_0 is far away from all the s_i's ?
- $c = (C(s_1, s_0|\theta), \ldots, c(s_i, s_0|\theta), \ldots, c(s_n, s_0|\theta)' \approx (0, \ldots, 0)'$
- Kriging mean: $\approx m(s_0) =$ unconditional mean
- Kriging variance: $\approx C(s_0, s_0) =$ unconditional variance
- $w(s_0)$ is almost independent of the $w(s_i)$'s i.e., information on the process at far away locations does not help much

Model comparison using the predictions

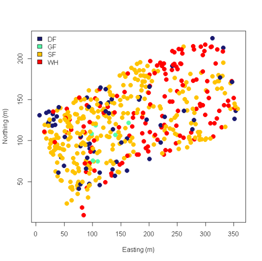
- Usually in spatial analysis data at some of the locations are held out for evaluating prediction performance
- Root Mean Square Predictive Error (RMSPE): $\sqrt{\frac{1}{n_{out}}\sum_{i=1}^{n_{out}}(y_i - \hat{y}_i)^2}$ where $\hat{y}_i$ are the kriging predictions
- Kriging also allows us to compute the q^{th} quantiles: $\hat{y}_{i,q} = \hat{y}_i + z_q\sqrt{(\hat{v}_i)}$ where $z_q =$ the q^{th} quantile of $N(0,1)$ and $\hat{v}_i =$ kriging variance
- Coverage probability (CP): $\frac{1}{n_{out}}\sum_{i=1}^{n_{out}} I(y_i \in (\hat{y}_{i,0.025}, \hat{y}_{i,0.975}))$
  - Ideally should be close to 95%
  - Otherwise we will have under or over coverage
- Width of 95% confidence interval (CIW): $\frac{1}{n_{out}}\sum_{i=1}^{n_{out}}(\hat{y}_{i,0.975} - \hat{y}_{i,0.025})$
- CP and CIW compare the distributions of y_i instead of comparing just their point predictions

Western Experimental Forestry (WEF) data

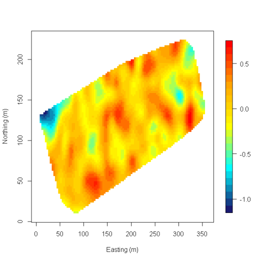
- Data consist of a census of all trees in a 10 ha. stand in Oregon
- Response of interest: log(Diameter at breast height), i.e., log(DBH)
- Covariate: Tree species (Categorical variable)



DBH



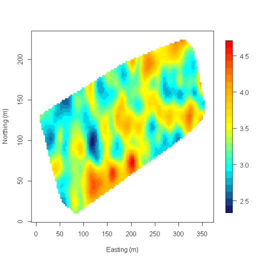
Species



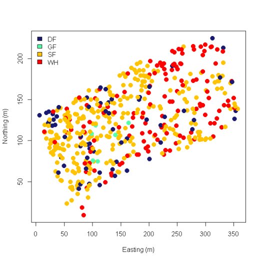
Residuals

Western Experimental Forestry (WEF) data

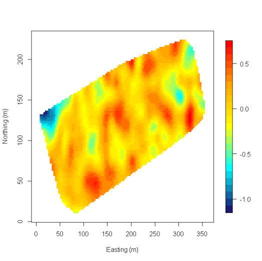
- Data consist of a census of all trees in a 10 ha. stand in Oregon
- Response of interest: log(Diameter at breast height), i.e., log(DBH)
- Covariate: Tree species (Categorical variable)



DBH



Species

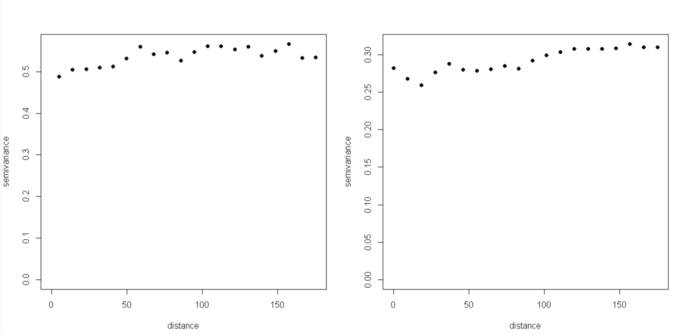


Residuals

- Local spatial patterns in the residual plot
- Simple regression on species seems to be not sufficient

Empirical semivariograms

- Regression model: $\log(DBH) \sim$ Species

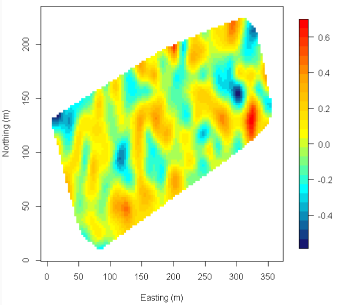


DBH Residuals

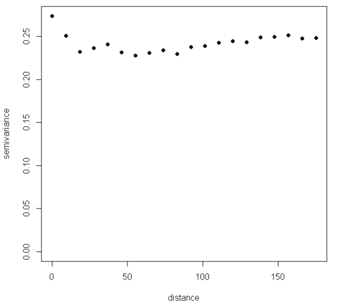
- Semivariogram of the residuals confirm unexplained spatial variation

Spatial model

- Regression model: $\log(DBH) \sim$ Species + Exponential GP



Residuals



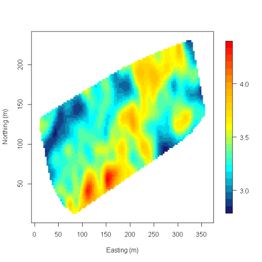
Variogram of residuals

Model comparisons

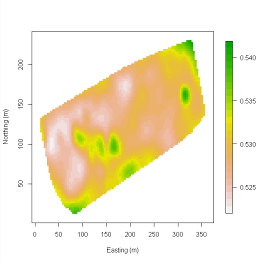
Table: Model comparison

	Spatial	Non-spatial
AIC	803	825
BIC	832	846
RMSPE	0.52	0.55
CP	97	97
CIW	2.07	2.15

WEF data: Kriged surfaces



DBH Estimates



Standard errors

Summary

- Spatial linear regression model for univariate point-referenced spatial data
- Modeling unknown surfaces with Gaussian Processes
- Kriging: Predictions at new locations
- Out of sample prediction
- Model comparison: AIC, BIC, RMSPE, CP, CIW
- Analysis in R