Introduction to spatial statistics

Abhi Datta

January 26, 2018

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

Course Outline

- Introduction – types of spatial data, exploratory data analysis
- Modeling univariate point referenced data – Gaussian Processes (GP), spatial regression, estimation, spatial prediction (kriging)
- Bayesian modeling – Metropolis Hastings, Gibbs sampler
- Large data – computing challenges, efficient alternatives
- Areal data – disease mapping

More about the course

- Materials available on https:
  //github.com/abhirupdatta/spatial-statistics-2018

- Texts for reference:

  - (Main) Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014), Hierarchical Modeling and Analysis for Spatial Data, Boca Raton, FL: Chapman and Hall/CRC, 2nd ed (BCG)

  - Cressie, N. A. C. and Wikle, C. K. (2011), Statistics for spatio-temporal data, Hoboken, NJ: Wiley, Wiley Series in Probability and Statistics

What is spatial data?

- Any data with some geographical information

What is spatial data?

- Any data with some geographical information

- Common sources of spatial data: climatology, forestry, ecology, environmental health, disease epidemiology, real estate marketing etc
  - have many important predictors and response variables
  - are often presented as maps

What is spatial data?

- Any data with some geographical information

- Common sources of spatial data: climatology, forestry, ecology, environmental health, disease epidemiology, real estate marketing etc
  - have many important predictors and response variables
  - are often presented as maps

- Other examples where spatial need not refer to space on earth:
  - Neuroimaging (data for each voxel in the brain)
  - Genetics (position along a chromosome)

- Three broad categories

Types of spatial data

- Point-referenced data
  - Each observation is associated with a location (point)
  - Data represents a sample from a continuous spatial domain
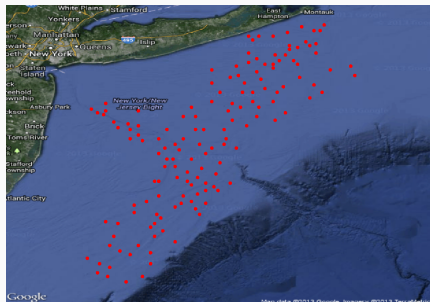 - Also referred to as *geocoded* or *geostatistical* data

Figure: Locations of scallops abundance data

Types of spatial data

- Areal data
  - Each observation is associated with a region like state, county etc.
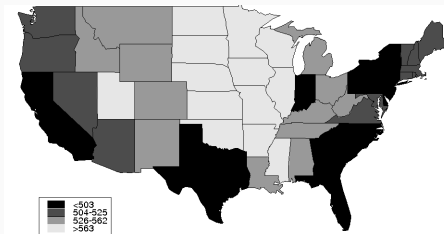 - Usually a result of aggregating point level data

Figure: SAT scores across the 48 contiguous states in the US

Types of spatial data

- Point pattern data
  - The locations are viewed as "random"
  - Need not have variables at locations, just the pattern of points
  - Interest in the pattern of occurrences of an event like disease incidence, species distribution, crimes etc.
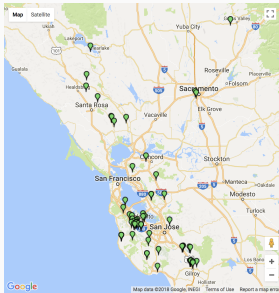

Figure: Locations of confirmed mountain lion sightings in the Bay area since 2004

Geostatistics – Analysis of point-referenced spatial data

- Each observation is associated with a location (point)
- Data represents a sample from a continuous spatial domain
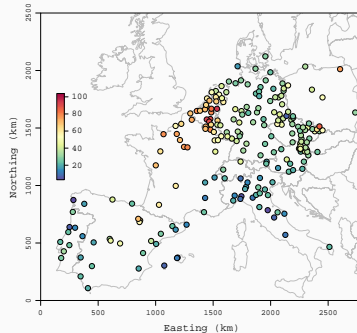- Also referred to as geocoded or geostatistical data

Figure: Pollutant levels in Europe in March, 2009

Point level modeling

- Point-level modeling refers to modeling of point-referenced data collected at locations referenced by coordinates (e.g., lat-long, Easting-Northing).
  Euclidean space.
- Example: Data about pollution levels $Y(s_1), Y(s_2), \ldots, Y(s_n)$ at sites $s_1, s_2, \ldots, s_n$
- Conceptually: Pollutant level exists at all possible sites
- We can learn about $Y(s)$ for any s in the region based on this data !
- Key to achieve this is exploiting structured dependence

Exploratory data analysis (EDA): Plotting the data

- At each s_i we observe the response $y(s_i)$ and a $p \times 1$ vector of covariates $x(s_i)'$
- Goals: Identify association between y and x, predict $y(s)$ at any arbitrary s

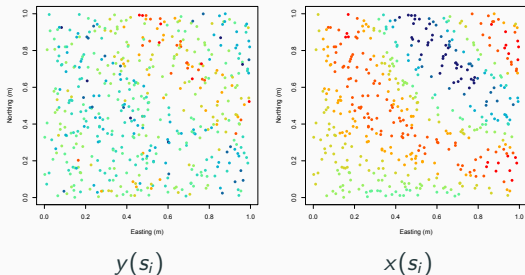$y(s_i)$ $\qquad\qquad\qquad\qquad\qquad$ $x(s_i)$

Figure: Response and covariate data for Dataset 1

7

Exploratory data analysis (EDA): Plotting the data

- At each s_i we observe the response $y(s_i)$ and a $p \times 1$ vector of covariates $x(s_i)'$
- Goals: Identify association between y and x, predict $y(s)$ at any arbitrary s
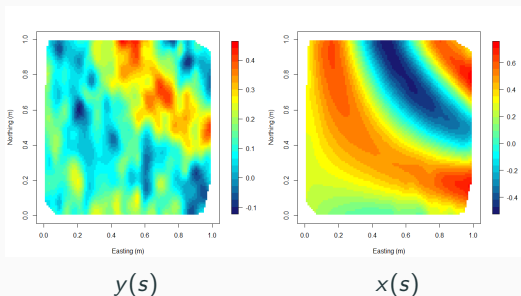- Surface plots of the data often helps to better understand spatial patterns

$y(s)$ $x(s)$

Figure: Response and covariate surface plots for Dataset 1

- Linear regression model: $y(s_i) = x(s_i)'\beta + \epsilon(s_i)$
- $\epsilon(s_i)$ are iid $N(0, \tau^2)$ errors
- $y = (y(s_1), y(s_2), \ldots, y(s_n))'$; $X = (x(s_1)', x(s_2)', \ldots, x(s_n)')'$
- Inference: $\hat{\beta} = (X'X)^{-1}X'Y \sim N(\beta, \tau^2(X'X)^{-1})$
- Prediction at new location s_0: $\widehat{y(s_0)} = x(s_0)'\hat{\beta}$
- Although the data is spatial, this is an ordinary linear regression model

Residual plots

- Surface plots of the residuals $(y(s) - \widehat{y(s)})$ help to identify any spatial patterns left unexplained by the covariates
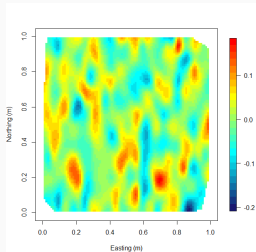

Figure: Residual plot for Dataset 1 after linear regression on $x(s)$

Residual plots

- Surface plots of the residuals $(y(s) - \widehat{y(s)})$ help to identify any spatial patterns left unexplained by the covariates
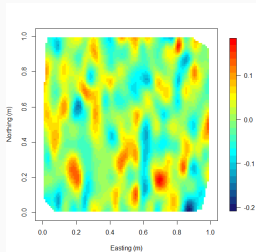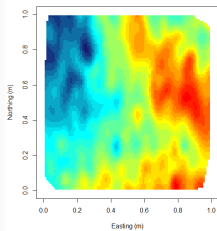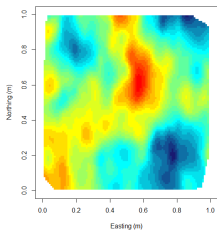

Figure: Residual plot for Dataset 1 after linear regression on $x(s)$

- No evident spatial pattern in plot of the residuals
- The covariate $x(s)$ seem to explain all spatial variation in $y(s)$
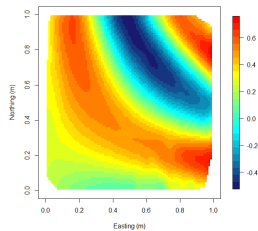- Does a non-spatial regression model always suffice?
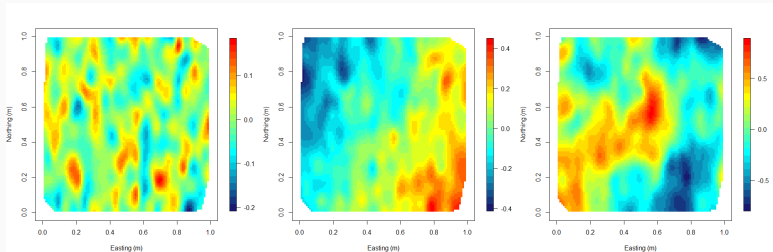
Dataset 2: $y(s)$ Dataset 3: $y(s)$ Same $x(s)$

Residual plots

- Linear regression: $y(\mathbf{s}_i) = \beta_0 + x(\mathbf{s}_i)\beta_1 + \epsilon(\mathbf{s}_i)$

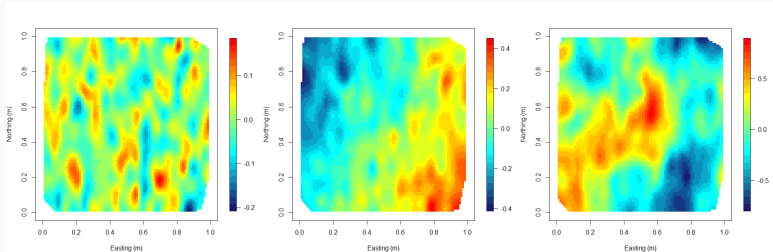

Dataset 1: $y(s)$ Dataset 2: $y(s)$ Dataset 3: $y(s)$

Residual plots

- Linear regression: $y(\mathbf{s}_i) = \beta_0 + x(\mathbf{s}_i)\beta_1 + \epsilon(\mathbf{s}_i)$

Dataset 1: $y(s)$ Dataset 2: $y(s)$ Dataset 3: $y(s)$

- Strong residual spatial pattern in datasets 2 and 3
- The covariate $x(\mathbf{s})$ does not explain all spatial variation in $y(\mathbf{s})$

More EDA

- Besides eyeballing residual surfaces, how to do more formal EDA to identify spatial pattern ?

- Besides eyeballing residual surfaces, how to do more formal EDA to identify spatial pattern ?

First law of geography

"Everything is related to everything else, but near things are more related than distant things." – Waldo Tobler

More EDA

- Besides eyeballing residual surfaces, how to do more formal EDA to identify spatial pattern ?

First law of geography

"Everything is related to everything else, but near things are more related than distant things." – Waldo Tobler

- In general pairwise squared differences of the data should have higher values if the locations are far apart
- In other words: $(Y(s + h) - Y(s))^2$ should be roughly increasing with $||h||$ will imply a spatial correlation
- Can this be formalized to identify spatial pattern?

Empirical semivariogram

- **Binning:** Make intervals $I_1 = (0, m_1)$, $I_2 = (m_1, m_2)$, and so forth, up to $I_K = (m_{K-1}, m_K)$. Representing each interval by its midpoint t_k, we define:
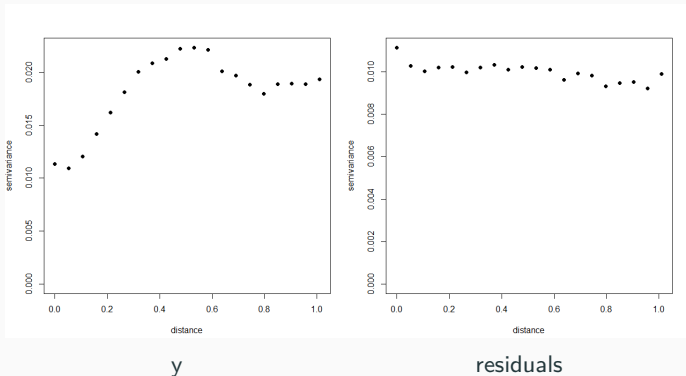
$$N(t_k) = \{(s_i, s_j) : \|s_i - s_j\| \in I_k\}, k = 1, \ldots, K.$$

- **Empirical semivariogram:**

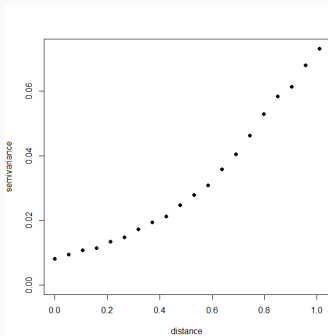$$\gamma(t_k) = \frac{1}{2|N(t_k)|} \sum_{s_i, s_j \in N(t_k)} (Y(s_i) - Y(s_j))^2$$

- For spatial data, the $\gamma(t_k)$ is expected to roughly increase with t_k

- A flat semivariogram would suggest little spatial variation

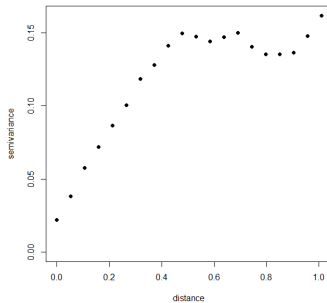# Empirical semivariogram: Data 1

y   residuals

- *variog* command in the *geoR* package in R calculates
  empirical semivariograms
- Residuals display little spatial variation
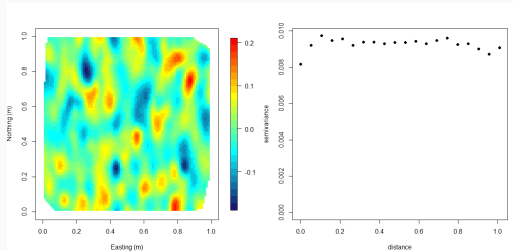
Data 2: residuals   Data 3: residuals

- Semivariograms of the residuals point to spatial variation

Using the locations

- When covariates does not explain all variation, one needs to leverage the information from the locations

Using the locations

- When covariates does not explain all variation, one needs to leverage the information from the locations
- Linear regression with the co-ordinates added as regressors:
$$y(\mathbf{s}_i) = \beta_0 + x(\mathbf{s}_i)\beta_1 + s_{ix}\beta_2 + s_{iy}\beta_3 + \epsilon(\mathbf{s}_i)$$

New residuals for data 2 Empirical semivariogram

- The linear model for the co-ordinates explains most of the spatial variation in dataset 2

16

- Linear regression with the co-ordinates added as regressors:
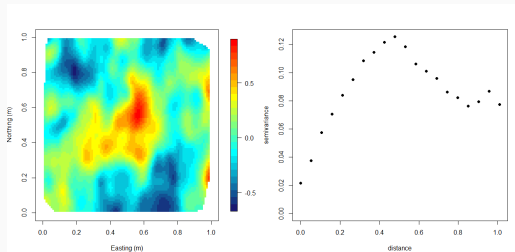 $$y(\mathbf{s}_i) = \beta_0 + x(\mathbf{s}_i)\beta_1 + s_{ix}\beta_2 + s_{iy}\beta_3 + \epsilon(\mathbf{s}_i)$$

Residuals for data 3   Empirical semivariogram

- Dataset 3 still exhibits strong spatial correlation

- Linear model for the co-ordinates often does not suffice

- More general model: $y(\mathbf{s}_i) = \beta_0 + x(\mathbf{s}_i)\beta_1 + w(\mathbf{s}_i) + \epsilon(\mathbf{s}_i)$

- How to choose the function $w(\cdot)$?

- Since we want to predict at any location over the entire domain, this choice will amount to choosing a surface $w(\mathbf{s})$

- How to do this ?

Modeling with the locations

- When purely covariate based models does not suffice, one needs to leverage the information from locations

- General model using the locations:
  $y(s) = x(s)'\beta + w(s) + \epsilon(s)$ 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 $w(s)$

- How to do this ?

Gaussian Processes (GPs)

- One popular approach to model w(s) is via Gaussian Processes (GP)
- The collection of random variables $\{w(s) \mid s \in D\}$ 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
- A GP is completely characterized by a mean function $m(s)$ and a 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 $C = C(s_i, s_j)$

Valid covariance functions and isotropy

- $C(\cdot, \cdot)$ needs to be valid. For all n and all $\{s_1, s_2, ..., s_n\}$, the resulting covariance matrix $C(s_i, s_j)$ for $(w(s_1), w(s_2), ..., w(s_n))$ must be positive definite
- So, $C(\cdot, \cdot)$ needs to be a positive definite function
- Simplifying assumptions:
    - Stationarity: $C(s_1, s_2)$ only depends on $h = s_1 - s_2$ (and is denoted by $C(h)$)
    - Isotropic: $C(h) = C(||h||)$
    - Anisotropic: Stationary but not isotropic
- Isotropic models are popular because of their simplicity, interpretability, and because a number of relatively simple parametric forms are available as candidates for C.

Some common isotropic covariance functions

Model	Covariance function, $C(t) = C(\|h\|)$
Spherical	$C(t) = \begin{cases} 0 & \text{if } t \geq 1/\phi \\ \sigma^2 \left[1 - \frac{3}{2}\phi t + \frac{1}{2}(\phi t)^3\right] & \text{if } 0 < t \leq 1/\phi \\ \tau^2 + \sigma^2 & \text{otherwise} \end{cases}$
Exponential	$C(t) = \begin{cases} \sigma^2 \exp(-\phi t) & \text{if } t > 0 \\ \tau^2 + \sigma^2 & \text{otherwise} \end{cases}$
Powered exponential	$C(t) = \begin{cases} \sigma^2 \exp(-|\phi t|^p) & \text{if } t > 0 \\ \tau^2 + \sigma^2 & \text{otherwise} \end{cases}$
Matérn at $\nu = 3/2$	$C(t) = \begin{cases} \sigma^2 \left(1 + \phi t\right) \exp(-\phi t) & \text{if } t > 0 \\ \tau^2 + \sigma^2 & \text{otherwise} \end{cases}$

$$C(t) = \begin{cases} \tau^2 + \sigma^2 & \text{if } t = 0 \\ \sigma^2 \exp(-\phi t) & \text{if } t > 0 \end{cases}.$$

- We define the effective range, t_0, as the distance at which this correlation has dropped to only 0.05. Setting $\exp(-\phi t_0)$ equal to this value we obtain $t_0 \approx 3/\phi$, since $\log(0.05) \approx -3$.
- 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
- Note discontinuity at 0 due to the nugget. Intentional! To account for measurement error or micro-scale variability.

- Recall: Empirical semivariogram:
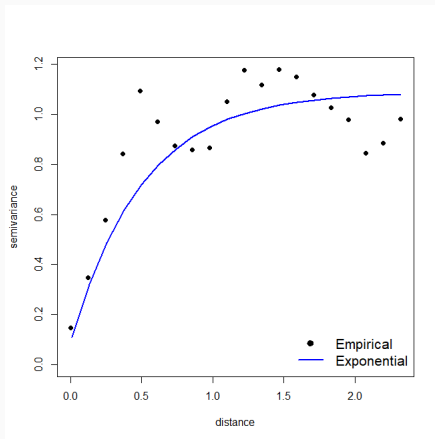 $\gamma(t_k) = \frac{1}{2|N(t_k)|} \sum_{s_i, s_j \in N(t_k)} (Y(s_i) - Y(s_j))^2$
- For any stationary GP,
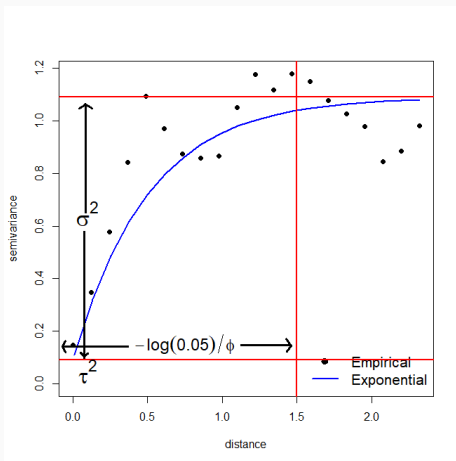 $E(Y(s+h) - Y(s))^2/2 = C(0) - C(h) = \gamma(h)$
- $\gamma(h)$ is the semivariogram corresponding to the covariance function $C(h)$
- Example: For exponential GP,
  $\gamma(t) = \begin{cases} \tau^2 + \sigma^2(1 - \exp(-\phi t)) & \text{if } t > 0 \\ 0 & \text{if } t = 0 \end{cases}$, where $t = ||h||$

- The Matèrn is a very versatile family:

$$
C(t) = \left\{
\begin{array}{ll}
\frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}(2\sqrt{\nu}t\phi)^\nu K_\nu(2\sqrt{(\nu)}t\phi) & \text{if } t > 0 \\
\tau^2 + \sigma^2 & \text{if } t = 0
\end{array}
\right.
$$

  K_ν is the modified Bessel function of order ν (computationally tractable)

- ν is a smoothness parameter controlling process smoothness. Remarkable!

- $\nu = 1/2$ gives the exponential covariance function

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 $m = E(w)$, $C = Cov(w)$, $c = Cov(w, w(s_0))$
  - Variance: $C(s_0, s_0) - c' C^{-1} c$
- The GP formulation gives the full predictive distribution of $w(s_0)|w$

Spatial linear model

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

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

- $y = (y(s_1), \ldots, y(s_n))' \sim N(X\beta, \sigma^2 R(\phi) + \tau^2 I)$
- We can obtain MLEs of parameters $\beta, \tau^2, \sigma^2, \phi$ based on the above model and use the estimates to krige at new locations
- In practice, the likelihood is often very flat with respect to the spatial covariance parameters and choice of initial values is important
- Initial values can be eyeballed from empirical semivariogram of the residuals from ordinary linear regression
- Estimated parameter values can be used for kriging

Model comparison

- For k total parameters and sample size n:
  - AIC: $2k - 2\log(l(y \mid \hat{\beta}, \hat{\theta}, \hat{\tau}^2))$
  - BIC: $\log(n)k - 2\log(l(y \mid \hat{\beta}, \hat{\theta}, \hat{\tau}^2))$

- Prediction based approaches using holdout data:
  - Root Mean Square Predictive Error (RMSPE):
    $\sqrt{\frac{1}{n_{out}} \sum_{i=1}^{n_{out}} (y_i - \hat{y}_i)^2}$

  - Coverage probability (CP): $\frac{1}{n_{out}} \sum_{i=1}^{n_{out}} I(y_i \in (\hat{y}_{i,0.025}, \hat{y}_{i,0.975}))$

  - Width of 95% confidence interval (CIW):
    $\frac{1}{n_{out}} \sum_{i=1}^{n_{out}} (\hat{y}_{i,0.975} - \hat{y}_{i,0.025})$

  - The last two approaches compares the distribution of y_i instead of comparing just their point predictions

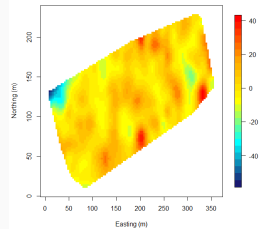# Western Experimental Forestry (WEF) data

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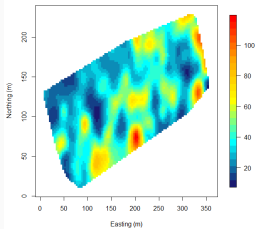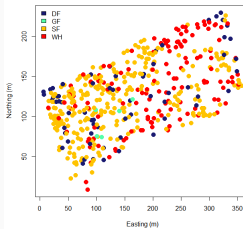
DBH   Species   Residuals

Western Experimental Forestry (WEF) data

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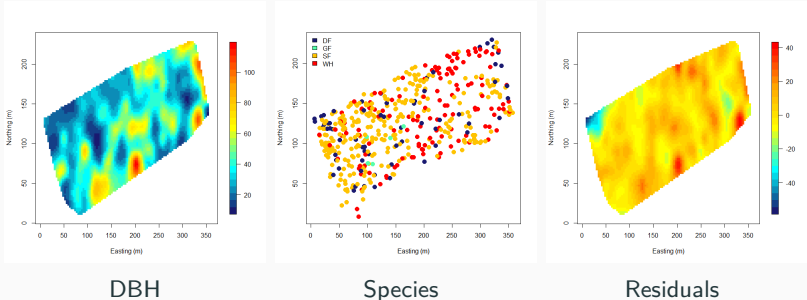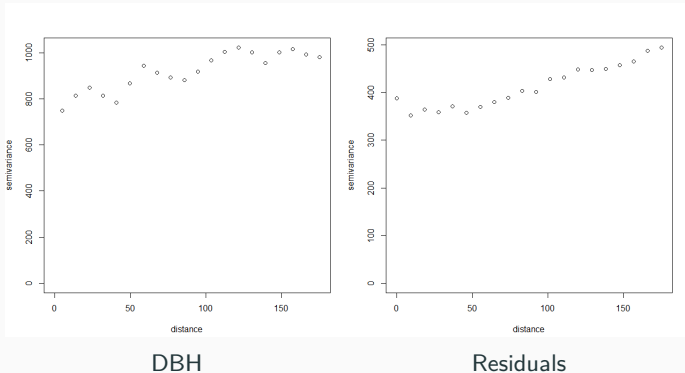- Covariate: Tree species (Categorical variable)

DBH   Species   Residuals

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

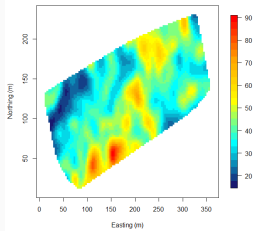## Empirical semivariograms

- Regression model: DBH $\sim$ Species

DBH   Residuals

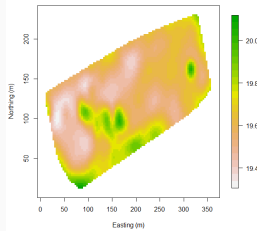- Semivariogram of the residuals confirm unexplained spatial variation

Model comparisons

Table: Model comparison

	Spatial	Non-spatial
AIC	4419	4465
BIC	4448	4486
RMSPE	18	21
CP	93	93
CIW	77	82

DBH Estimates

Standard errors

Summary

- Geostatistics – Analysis of point-referenced spatial data

- Surface plots of data and residuals

- EDA with empirical semivariograms

- Modeling unknown surfaces with Gaussian Processes

- Kriging: Predictions at new locations

- Spatial linear regression using Gaussian Processes