Introduction to Geostatistics

Sudipto Banerjee

September 03–05, 2017

Department of Biostatistics, Fielding School of Public Health, University of California, Los Angeles

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)

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

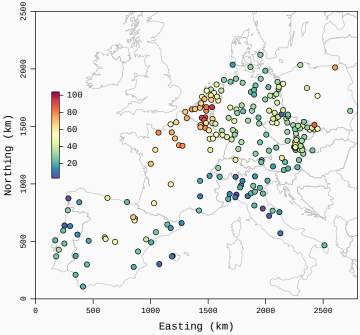


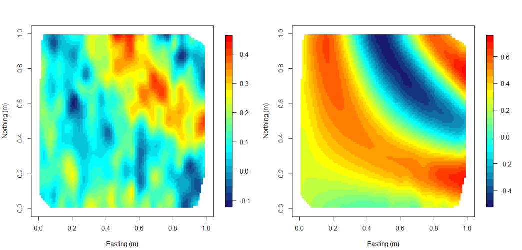
Figure: Pollutant levels in Europe in March, 2009

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).
- Data from a spatial process $\{Y(s) : s \in D\}$, D is a subset in Euclidean space.
- Example: $Y(s)$ is a pollutant level at site s
- Conceptually: Pollutant level exists at all possible sites
- Practically: Data will be a partial realization of a spatial process – observed at $\{s_1, \ldots, s_n\}$
- Statistical objectives: Inference about the process $Y(s)$; predict at new locations.
- Remarkable: Can learn about entire $Y(s)$ surface. The key: Structured dependence

Exploratory data analysis (EDA): Plotting the data

- A typical setup: Data observed at n locations $\{s_1, \ldots, s_n\}$
- At each s_i we observe the response $y(s_i)$ and a $p \times 1$ vector of covariates $x(s_i)^\top$
- Surface plots of the data often helps to understand spatial patterns



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

Figure: Response and covariate surface plots for Dataset 1

What's so special about spatial?

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

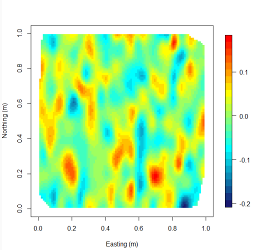


Figure: Residual plot for Dataset 1 after linear regression on $x(s)$

Residual plots

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

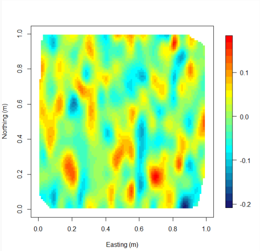
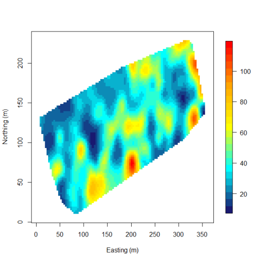


Figure: Residual plot for Dataset 1 after linear regression on $x(s)$

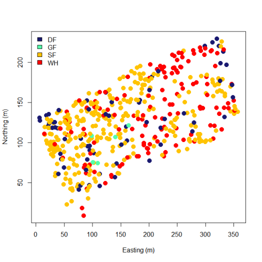
- No evident spatial pattern in plot of the residuals
- The covariate $x(s)$ seem to explain all spatial variation in $y(s)$
- Does a non-spatial regression model always suffice?

Western Experimental Forestry (WEF) data

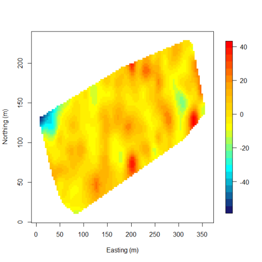
- Data consist of a census of all trees in a 10 ha. stand in Oregon
- Response of interest: Diameter at breast height (DBH)
- Covariate: Tree species (Categorical variable)



DBH



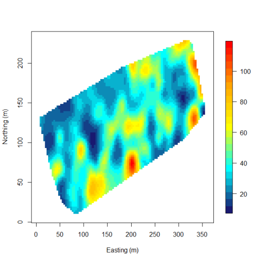
Species



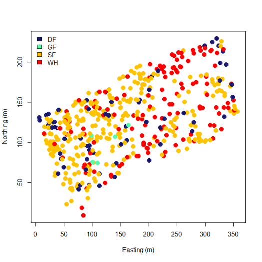
Residuals

Western Experimental Forestry (WEF) data

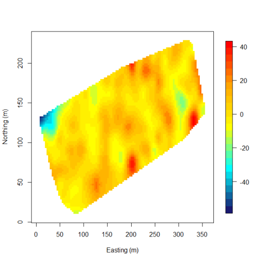
- Data consist of a census of all trees in a 10 ha. stand in Oregon
- Response of interest: Diameter at breast height (DBH)
- Covariate: Tree species (Categorical variable)



DBH



Species



Residuals

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

More EDA

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

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

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 $(Y(s+h) - Y(s))^2$ 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:

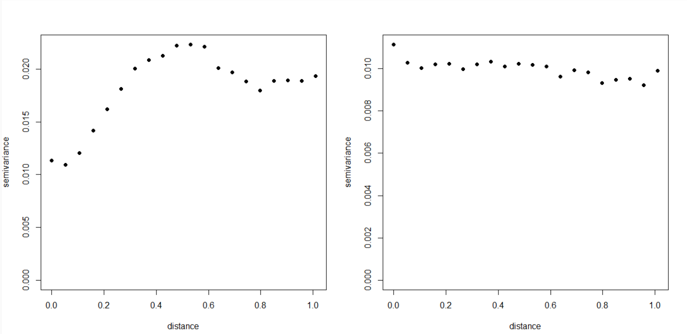
$$N(t_k) = \{(s_i, s_j) : \|s_i - s_j\| \in I_k\}, k = 1, \dots, K.$$

- 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 variogram: Data 1



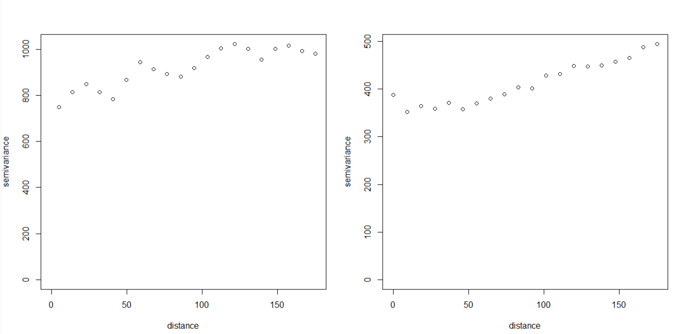
y

residuals

- Residuals display little spatial variation

Empirical variograms: WEF data

- Regression model: DBH $\sim$ Species



DBH Residuals

- Variogram of the residuals confirm unexplained spatial variation

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: $y(s) = x(s)^\top \beta + w(s) + \epsilon(s)$ 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.
  $w = (w(s_1), \ldots, w(s_n))^\top \sim N(m, C)$ where
  $m = (m(s_1), \ldots, m(s_n))^\top$ 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 $\nu = 3/2$	$C(t) = \begin{cases} \sigma^2 (1 + \phi t) \exp(-\phi t) & \text{if } t > 0 \\ \tau^2 + \sigma^2 & \text{otherwise} \end{cases}$

Notes on exponential model

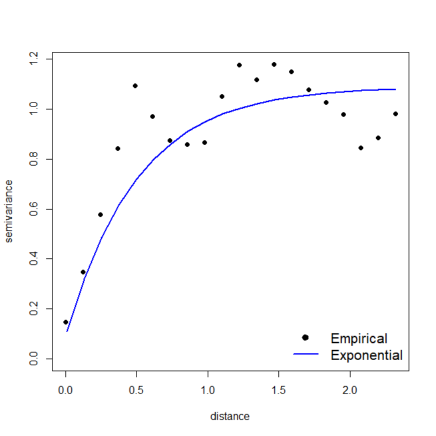
$$C(t) = \begin{cases} \tau^2 + \sigma^2 & \text{if } t = 0 \\ \sigma^2 \exp(-\phi t) & \text{if } t > 0 \end{cases} .$$

- 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.

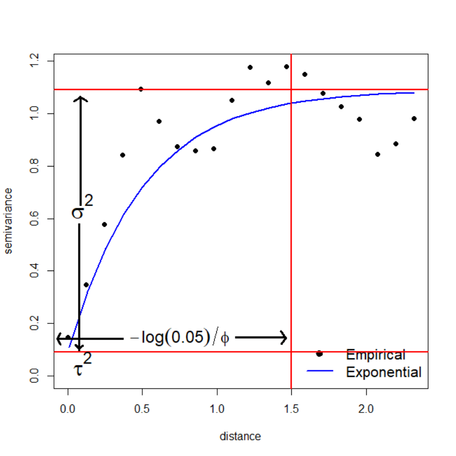
Covariance functions and semivariograms

- 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 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 $t = ||h||$

Covariance functions and semivariograms



Covariance functions and semivariograms



The Matèrn covariance function

- The Matèrn is a very versatile family:

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

  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 $w = (w(s_1), w(s_2), \ldots, w(s_n))^\top$, predict $w(s_0)$ for a new location s_0
- If $w(s)$ is modeled as a GP, then $(w(s_0), w(s_1), \ldots, w(s_n))^\top$ jointly follow multivariate normal distribution
- $w(s_0) \mid w$ follows a normal distribution with
  - Mean (kriging estimator): $m(s_0) + c^\top C^{-1}(w - m)$
  - where $m = E(w)$, $C = Cov(w)$, $c = Cov(w, w(s_0))$
  - Variance: $C(s_0, s_0) - c^\top C^{-1} c$
- The GP formulation gives the full predictive distribution $p(w(s_0) \mid w)$

Modeling with GPs

Spatial linear model

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

- $w(s)$ 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: $C(s+h, s) = \sigma^2 R(||h|| \,; \phi)$
- $w = (w(s_1), \ldots, w(s_n))^\top \sim N(0, \sigma^2 R(\phi))$ where $R(\phi) = \sigma^2(R(||s_i - s_j|| \,; \phi))$
- $y = (y(s_1), \ldots, y(s_n))^\top \sim N(X\beta, \sigma^2 R(\phi) + \tau^2 I)$

Parameter estimation

- $y = (y(s_1), \ldots, y(s_n))^\top \sim N(X\beta, \sigma^2 R(\phi) + \tau^2 I)$
- 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

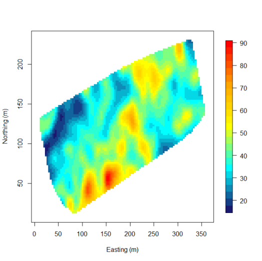
- For k total parameters and sample size n:
  - AIC: $2k - 2\log(l(y \mid \hat{\beta}, \hat{\theta}, \hat{\tau^2}))$
  - BIC: $\log(n)k - 2\log(l(y \mid \hat{\beta}, \hat{\theta}, \hat{\tau^2}))$
- Prediction based approaches using holdout data:
  - Root Mean Square Predictive Error (RMSPE): $\sqrt{\frac{1}{n_{out}} \sum_{i=1}^{n_{out}} (y_i - \hat{y}_i)^2}$
  - 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

Back to WEF data

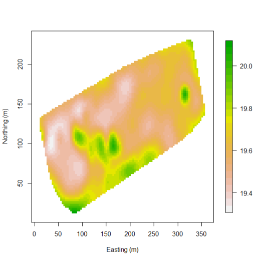
Table: Model comparison

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

WEF data: Kriged surfaces



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