

Introduction to spatial statistics

Abhi Datta

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, frequentist and Bayesian estimation and spatial prediction (kriging)
- Large data – computing challenges, efficient alternatives, spatial machine learning
- Areal data – disease mapping
- *Point-pattern data – cluster/hotspot detection
- *Spatially varying relationships, change-of-support problems

More about the course

- Materials available on <https://github.com/abhirupdatta/advanced-spatial-statistics-2021>
- 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)

Types of spatial data

- 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

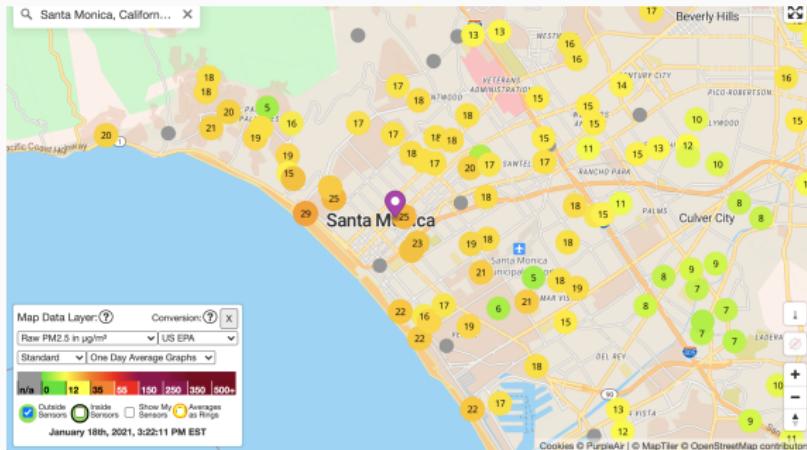


Figure: PM_{2.5} readings in μgm^{-3} from PurpleAir low-cost sensors in Santa Monica, CA on Jan 18, 2021.¹

¹Source: <https://www.purpleair.com/>

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

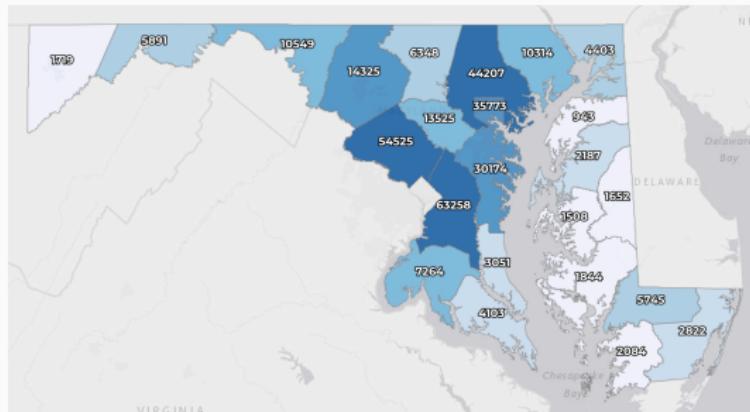


Figure: COVID cases for Maryland counties on Jan 18, 2021¹

¹Source: <https://coronavirus.maryland.gov/datasets/>

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.

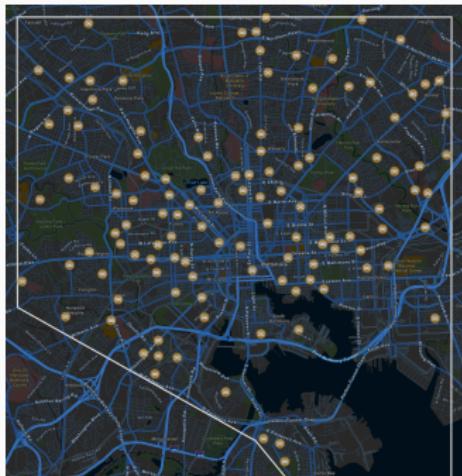


Figure: Locations of auto-theft in Baltimore city in Dec 2020 - Jan 2021¹

¹Source: <https://arcgisportal.baltimorepolice.org/publiccrimemap/>

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

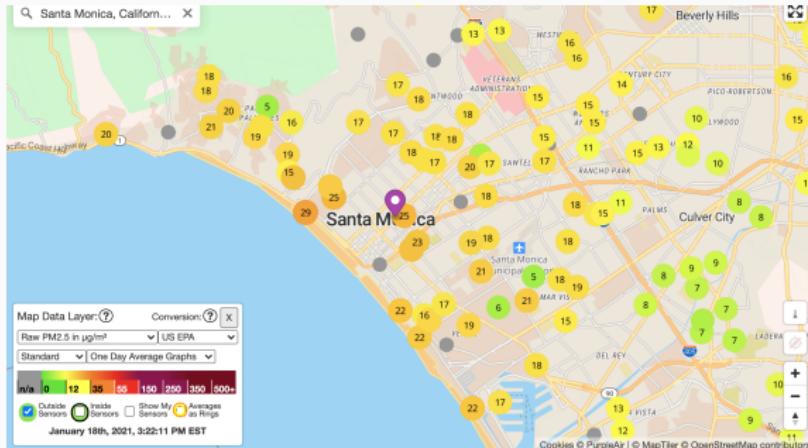


Figure: PM_{2.5} readings in $\mu\text{g}/\text{m}^{-3}$ from PurpleAir low-cost sensors in Santa Monica, CA on Jan 18, 2021.²

²Source: <https://www.purpleair.com/>

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 $Y(s_1), Y(s_2), \dots, Y(s_n)$ at sites s_1, s_2, \dots, 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

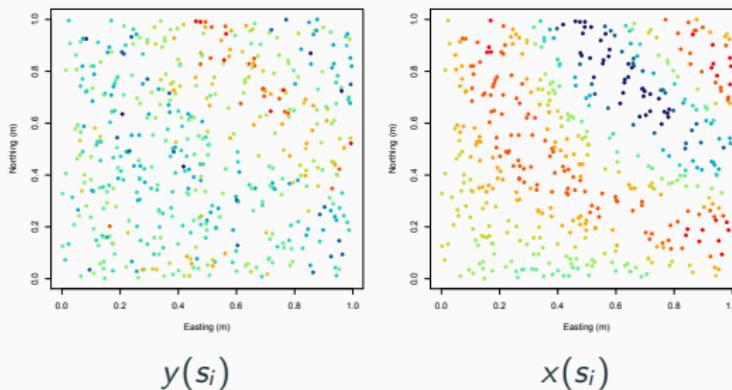


Figure: Response and covariate data for Dataset 1

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

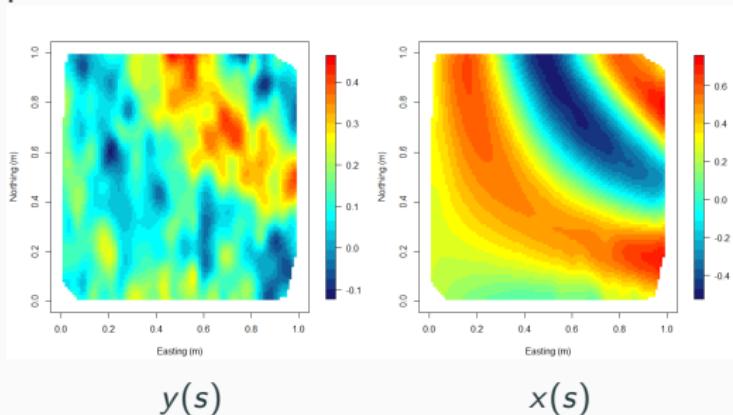


Figure: Response and covariate surface plots for Dataset 1

What's so special about spatial?

- Linear regression model: $y(s_i) = x(s_i)'\beta + \epsilon(s_i)$
- $\epsilon(s_i)$ are iid $N(0, \tau^2)$ errors
- $y = (y(s_1), y(s_2), \dots, y(s_n))'$; $X = (x(s_1)', x(s_2)', \dots, x(s_n)')$
- 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

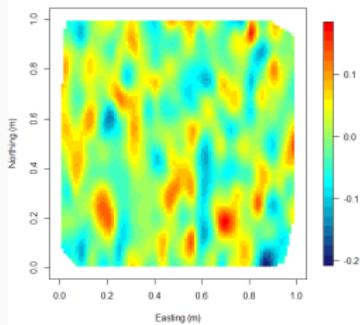


Figure: Residual plot for Dataset 1 after linear regression on $x(s)$

Residual plots

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

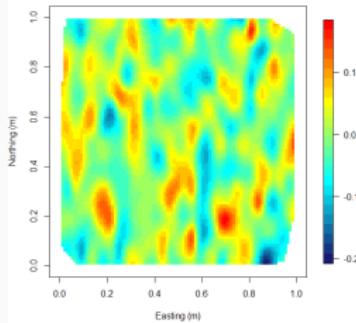
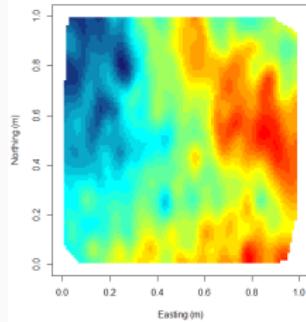


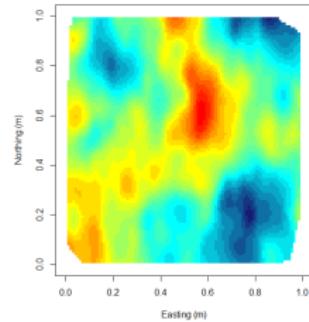
Figure: Residual plot for Dataset 1 after linear regression on $x(s)$

- 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?

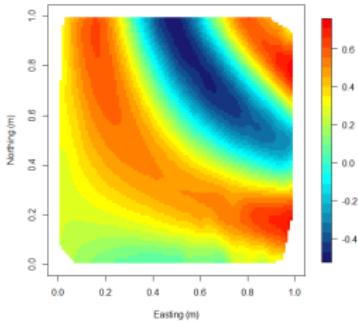
Two more datasets



Dataset 2: $y(s)$



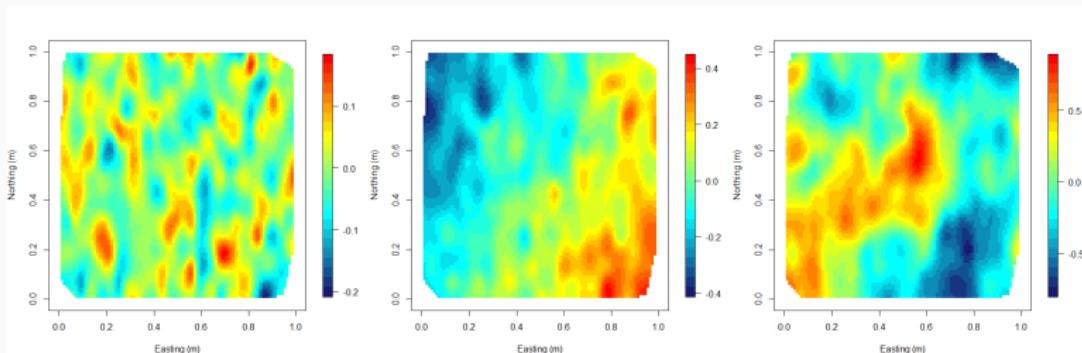
Dataset 3: $y(s)$



Same $x(s)$

Residual plots

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



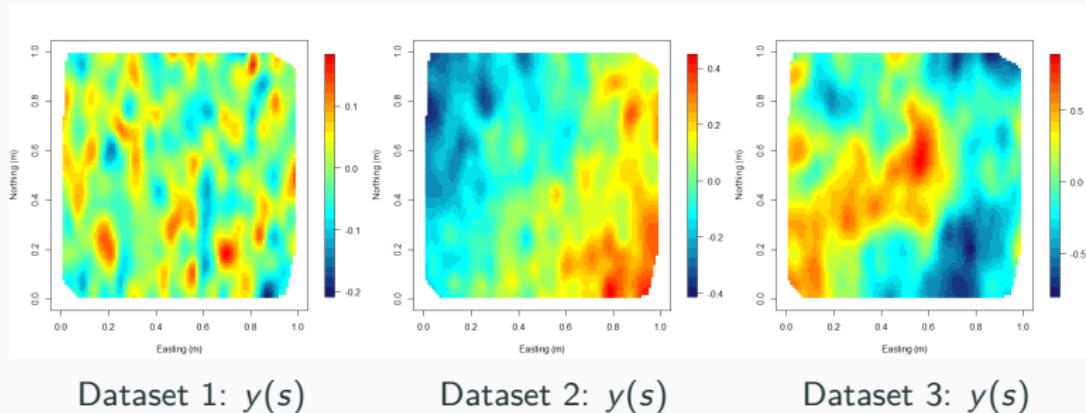
Dataset 1: $y(s)$

Dataset 2: $y(s)$

Dataset 3: $y(s)$

Residual plots

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



- 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 ?

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 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:

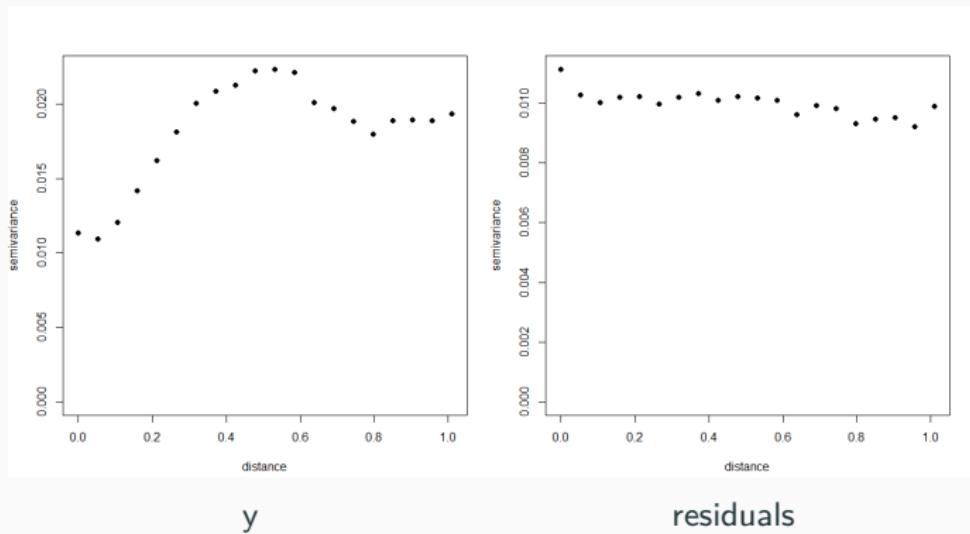
$$N(t_k) = \{(s_i, s_j) : \|s_i - s_j\| \in I_k\}, k = 1, \dots, K.$$

- **Empirical semivariogram:**

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

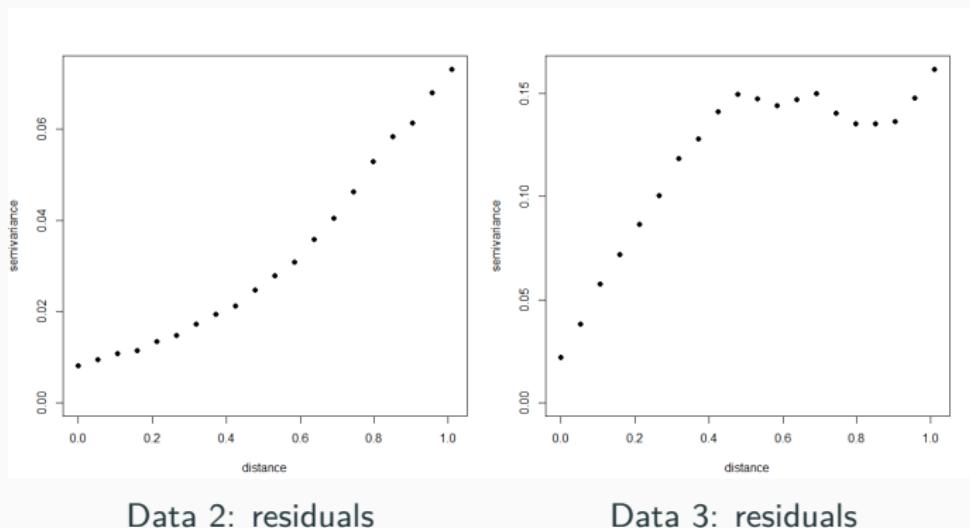
- For spatial data, the $\gamma(t_k)$ is expected to roughly increase with t_k
- A flat semivariogram would suggest little spatial variation

Empirical semivariogram: Data 1



- *variog* command in the *geoR* package in R calculates empirical semivariograms
- Residuals display little spatial variation

Empirical semivariograms: Data 2 and 3



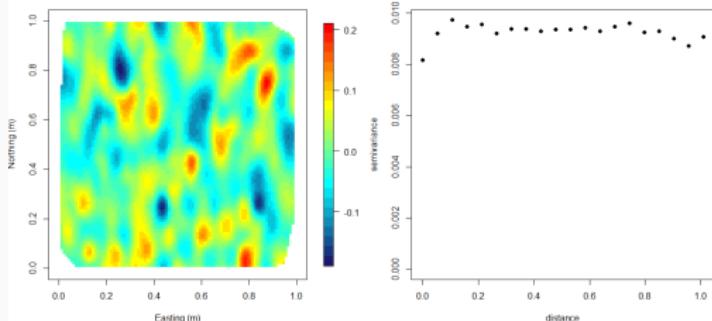
- Semivariograms of the residuals point to spatial variation

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:
 $y(\mathbf{s}_i) = \beta_0 + x(\mathbf{s}_i)\beta_1 + s_{ix}\beta_2 + s_{iy}\beta_3 + \epsilon(\mathbf{s}_i)$

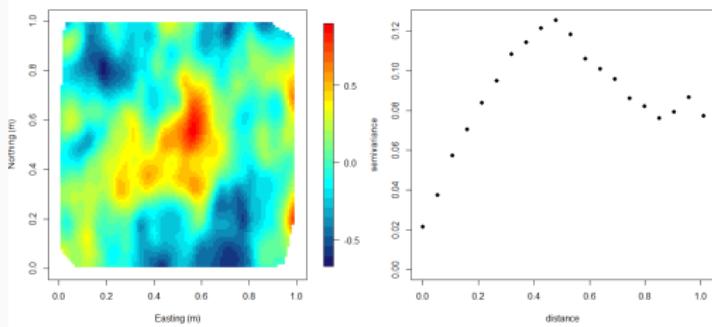


New residuals for data 2 Empirical semivariogram

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

Using the locations

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



Residuals for data 3 Empirical semivariogram

- Dataset 3 still exhibits strong spatial correlation

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 $s \in D$
- Since we want to predict at any location over the entire domain D , this choice will amount to choosing a **surface or function** $w(s)$
- How to choose a **smooth** function $w(\cdot)$?

General idea

- Model: $y(s) = x(s)'\beta + w(s) + \epsilon(s)$
- Optimize $\|y - X\beta - w\|_2^2 + P(w)$ where
 $y = (y(s_1), \dots, y(s_n))'$, X and w are similarly defined,
 $P(w)$ is some '**roughness**' penalty (**regularization term**)
shrinking towards smooth estimates of the function $w(\cdot)$
- Many approaches are related to this general approach: basis
function expansions, penalized regression splines, Gaussian
Processes

General idea

- Model: $y(s) = x(s)'\beta + w(s) + \epsilon(s)$
- Optimize $\|y - X\beta - w\|_2^2 + P(w)$ where
 $y = (y(s_1), \dots, y(s_n))'$, X and w are similarly defined,
 $P(w)$ is some '**roughness**' penalty (**regularization term**)
shrinking towards smooth estimates of the function $w(\cdot)$
- Many approaches are related to this general approach: basis function expansions, penalized regression splines, Gaussian Processes
- E.g., Modeling $w(\cdot)$ as a natural cubic spline $B(\cdot)\gamma$ will lead to the optimization of the form: $\|y - X\beta - B\gamma\|_2^2 + \lambda\gamma'\Omega\gamma$

General idea

- Model: $y(s) = x(s)'\beta + w(s) + \epsilon(s)$
- Optimize $\|y - X\beta - w\|_2^2 + P(w)$ where
 $y = (y(s_1), \dots, y(s_n))'$, X and w are similarly defined,
 $P(w)$ is some '**roughness**' penalty (**regularization term**)
shrinking towards smooth estimates of the function $w(\cdot)$
- Many approaches are related to this general approach: basis function expansions, penalized regression splines, Gaussian Processes
- E.g., Modeling $w(\cdot)$ as a natural cubic spline $B(\cdot)\gamma$ will lead to the optimization of the form: $\|y - X\beta - B\gamma\|_2^2 + \lambda\gamma'\Omega\gamma$
- In this course, we will focus on **Gaussian Processes**

Gaussian Processes (GPs)

- One popular approach to model $w(s)$ is via Gaussian Processes (GP)
- GP is a distribution over the class of functions, whose support is the space of ‘suitably smooth’ functions
- The collection of random variables $\{w(s) | s \in D\}$ is a GP if
 - it is a valid stochastic process
 - all finite dimensional densities $\{w(s_1), \dots, w(s_n)\}$ 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.
 $w = (w(s_1), \dots, w(s_n))' \sim N(m, C)$ where
 $m = (m(s_1), \dots, m(s_n))'$ and $C = C(s_i, s_j)$

Valid covariance functions and isotropy

- $C(\cdot, \cdot)$ needs to be **valid**. For all n and all $\{s_1, s_2, \dots, s_n\}$, the resulting covariance matrix $C(s_i, s_j)$ for $(w(s_1), w(s_2), \dots, w(s_n))$ 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 .

The Matérn covariance family

The stationary (isotropic) Matérn covariance function is specified by

$$C(\|h\|_2) = \sigma^2 \frac{2^{1-\nu} (\sqrt{2}\phi\|h\|_2)^\nu}{\Gamma(\nu)} \mathcal{K}_\nu(\sqrt{2}\phi\|h\|_2),$$

where \mathcal{K}_ν is the modified Bessel function of second kind.

- $\sigma^2 = \text{Cov}(w(s)) =$ Marginal variance for all s
- $\nu =$ Smoothness parameter; a GP $w(\cdot)$ with Matérn covariance with smoothness ν is ($\lceil \nu \rceil - 1$)-times (mean square) differentiable functions (Stein, 1999³)
- $\phi =$ Spatial decay (inverse scale) parameter

³Stein, M. L. (2012). Interpolation of spatial data: some theory for kriging. Springer Science & Business Media

The Matérn covariance family (special cases)

- $\nu = 1/2 \implies$ exponential covariance function;

$$C(\|h\|_2) = \sigma^2 \exp(-\phi\|h\|_2)$$

- $\nu = 3/2 \implies$ smoother than exponential;

$$C(\|h\|_2) = \sigma^2(1 + \phi\|h\|_2) \exp(-\phi\|h\|_2)$$

- $\nu = \infty \implies$ squared exponential (or Gaussian) covariance function (very smooth);

$$C(\|h\|_2) = \sigma^2 \exp(-\phi\|h\|_2^2)$$

Some traditional terminology

$w(\cdot) \sim GP(0, C)$ with marginal variance σ^2 and $\epsilon(s) \stackrel{\text{iid}}{\sim} N(0, \tau^2)$,
 $\implies w(\cdot) + \epsilon(\cdot) \sim GP(0, C^*)$ where $C^* = C + \delta$ with
 $\delta(s, s') = \tau^2 I(s = s')$.

- E.g. If C is the exponential covariance function, then
$$C^*(\|h\|) = \tau^2 I(h = 0) + \sigma^2 \exp(-\phi \|h\|)$$
- τ^2 (called the **nugget**) is the noise (measurement error) or “micro-scale variance”.
- Note **discontinuity** at 0 due to the nugget.
- The **partial sill** (σ^2) is the “**spatial effect variance**.”
- $\sigma^2 + \tau^2$ is the **sill**, i.e., the total variance.
- 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 $\log(0.05) \approx -3$.

Summary

- Geostatistics – Analysis of point-referenced spatial data
- Surface plots of data and residuals
- EDA with empirical semivariograms
- Basics of Gaussian Processes and covariance functions
- Modeling unknown surfaces with GP (to be contd.)