

# Introduction to Spatial Data

Abhi Datta

03.31.2017

# Course Outline

- Introduction – types of spatial data, goals of analysis
- Exploratory data analysis – plotting, variograms
- Modeling – Gaussian Processes (GP), spatial prediction (kriging)
- Estimation – variogram fitting, spatial regression and GLM
- Bayesian modeling – Metropolis Hastings, Gibbs sampler
- Large data – computing challenges, efficient alternatives

## More about the course

- **Evaluation** – presenting a paper on large scale spatial analysis
- Materials available on <https://github.com/abhirupdatta/spatial-course-2017>
- 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

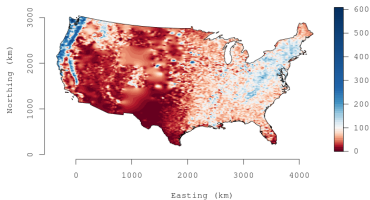
# What is spatial data?

# What is spatial data?

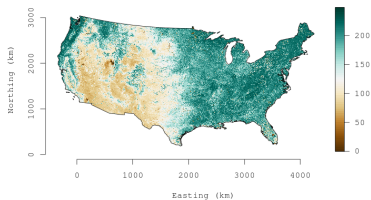
- Any data with some geographical information

# What is spatial data?

- Any data with some geographical information
- Example: US forest biomass data



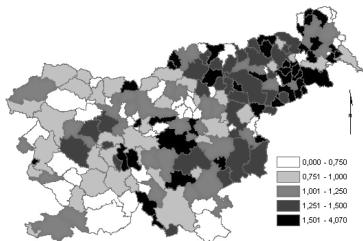
(a) US forest biomass data



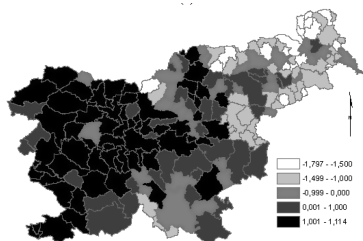
(b) NDVI (predictor)

# What is spatial data?

- Any data with some geographical information
- Example: Slovenia stomach cancer data



(a) Standardized cancer incidence



(b) Socioeconomic score (predictor)

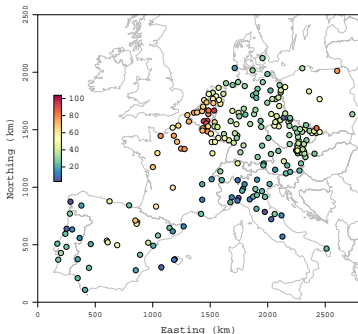
# What is spatial data?

- 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 the space need not be the space on earth:
  - Neuroimaging (data for each voxel in the brain)
  - Genetics (position along a chromosome)

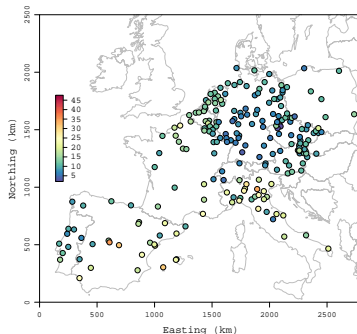


# Spatio-temporal Data

- Data from multiple timepoints at some or each of the locations
- Example: European air pollution data



(a) PM<sub>10</sub> levels in March, 2009



(b) PM<sub>10</sub> levels in June, 2009

# 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

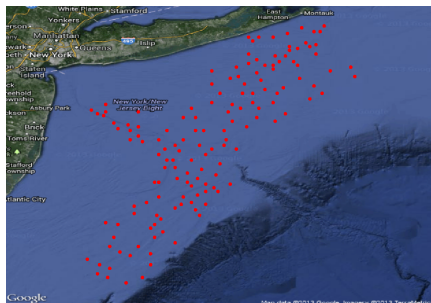


Figure: Locations of scallops abundance data

# 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
  - The spatial information is represented in terms of a graph depicting the relative orientation of the regions

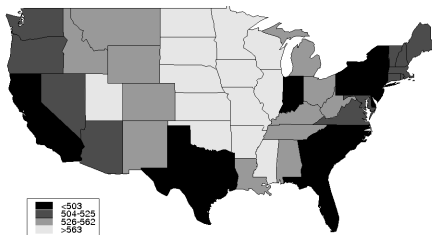


Figure: SAT scores across the 48 contiguous states in the US

# 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
  - The spatial information is represented in terms of a graph depicting the relative orientation of the regions

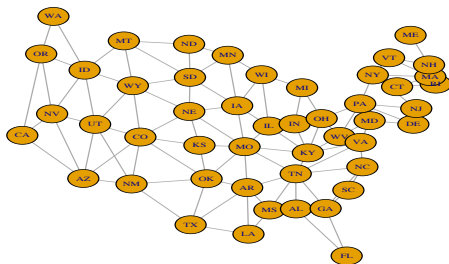


Figure: Graph for US state map

# 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.

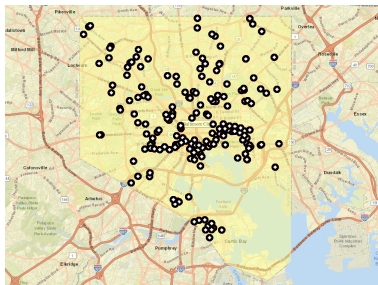
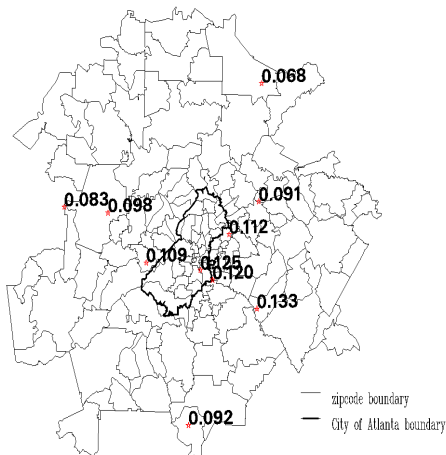


Figure: Locations of robberies in Baltimore in February 2017

# Types of spatial data



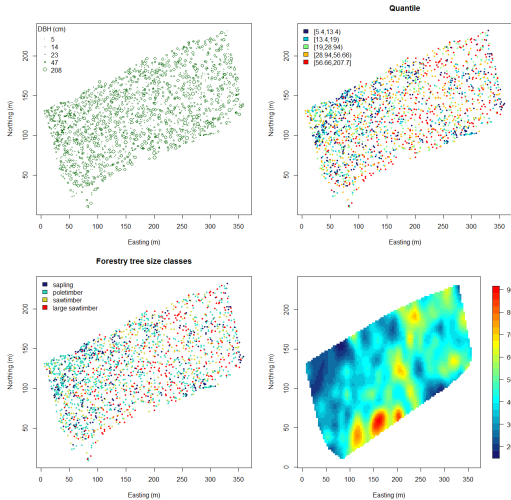
**Figure:** Areal and point-referenced data: Atlanta zip codes and 8-hour maximum ozone levels (ppm) at 10 sites, July 15, 1995

## Point-referenced data

- Point-level modeling refers to modeling of spatial data collected at locations referenced by **coordinates** (e.g., lat-long, Easting-Northing).
- Data from a spatial process  $\{Y(\mathbf{s}) : \mathbf{s} \in D\}$ ,  $D$  is a subset in Euclidean space.
- **Example:**  $Y(\mathbf{s})$  is a **pollutant level** at site  $\mathbf{s}$
- **Conceptually:** Pollutant level exists at all possible sites
- **Practically:** Data will be a partial realization of a spatial process – observed at  $\{\mathbf{s}_1, \dots, \mathbf{s}_n\}$
- **Statistical objectives:** **Inference** about the process  $Y(\mathbf{s})$ ; **predict** at new locations.
- **Remarkable:** Can learn about entire  $Y(\mathbf{s})$  surface. The **key:** Structured dependence



# Plotting point referenced data

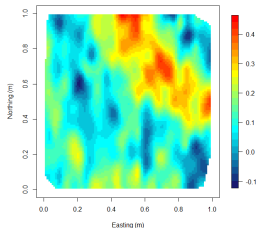


**Figure:** Western Experimental Forest (WEF) inventory data on diameter at breast height (DBH) of plants

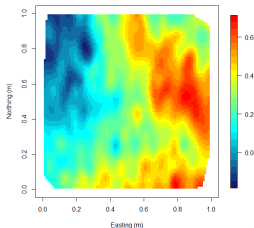
## What's so special about spatial?

- A typical setup: Data observed at  $n$  locations  $\{\mathbf{s}_1, \dots, \mathbf{s}_n\}$
- At each  $\mathbf{s}_i$  we observe the response  $y(\mathbf{s}_i)$  and a  $p \times 1$  vector of covariates  $\mathbf{x}(\mathbf{s}_i)'$
- Linear regression model:  $y(\mathbf{s}_i) = \mathbf{x}(\mathbf{s}_i)' \beta + \epsilon(\mathbf{s}_i)$
- $\epsilon(\mathbf{s}_i)$  are iid  $N(0, \tau^2)$  errors
- Although the data is spatial, this is an ordinary linear regression model
- $y = (y(\mathbf{s}_1)', y(\mathbf{s}_2)', \dots, y(\mathbf{s}_n)')'$ ;  $X = (\mathbf{x}(\mathbf{s}_1), \mathbf{x}(\mathbf{s}_2), \dots, \mathbf{x}(\mathbf{s}_n))'$
- Inference:  $\hat{\beta} = (X'X)^{-1}X'Y \sim N(\beta, \tau^2(X'X)^{-1})$
- Prediction at new location  $\mathbf{s}_0$ :  $\widehat{y(\mathbf{s}_0)} = \mathbf{x}(\mathbf{s}_0)' \hat{\beta}$
- Does this always suffice or we need any thing specialized method for such data?

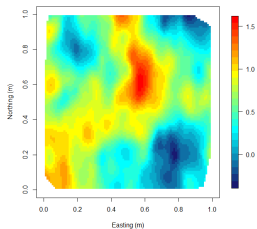
# Exploratory data analysis



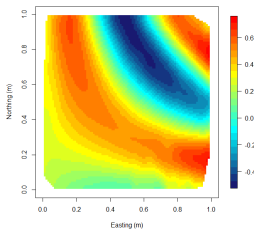
(a) Data 1



(b) Data 2



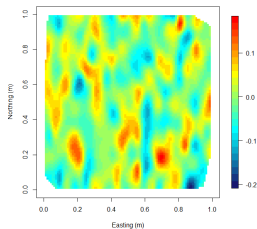
(c) Data 3



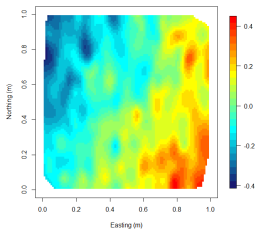
(d)  $x(s)$

# Residual plots

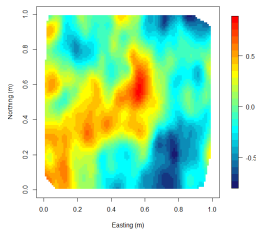
- Linear regression:  $y(\mathbf{s}_i) = \beta_0 + x(\mathbf{s}_i)\beta_1 + \epsilon(\mathbf{s}_i)$



(a) Data 1



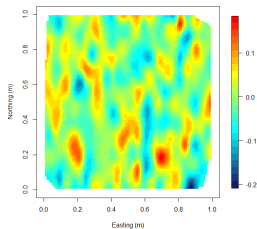
(b) Data 2



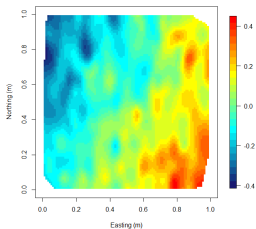
(c) Data 3

# Residual plots

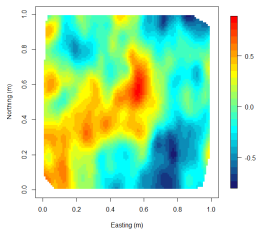
- Linear regression:  $y(\mathbf{s}_i) = \beta_0 + x(\mathbf{s}_i)\beta_1 + \epsilon(\mathbf{s}_i)$



(a) Data 1



(b) Data 2



(c) Data 3

- 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

- The residual surface seems continuous
- If a spatial surface  $Y(\mathbf{s})$  is continuous then  $(Y(\mathbf{s} + \mathbf{h}) - Y(\mathbf{s}))^2 \rightarrow 0$  as  $\|\mathbf{h}\| \rightarrow 0$
- In general  $(Y(\mathbf{s} + \mathbf{h}) - Y(\mathbf{s}))^2$  increasing with  $\|\mathbf{h}\|$  will imply a spatial correlation



# Empirical variogram

- Plot  $(Y(\mathbf{s}_i) - Y(\mathbf{s}_j))^2$  as function of  $\|\mathbf{s}_i - \mathbf{s}_j\|$  for all  $i, j$

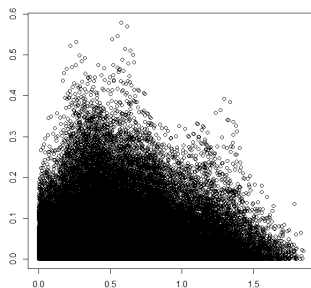


Figure: Points cloud for data 1

# Empirical variogram

- **Binning:** Grid up the  $t$  space into intervals  $I_1 = (0, t_1)$ ,  $I_2 = (t_1, t_2)$ , and so forth, up to  $I_K = (t_{K-1}, t_K)$ . Representing  $t$  values in each interval by its midpoint, we define:

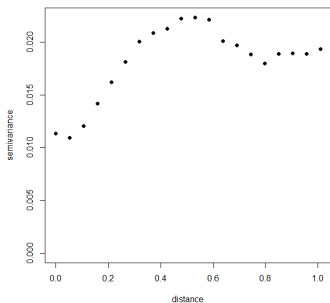
$$N(t_k) = \{(\mathbf{s}_i, \mathbf{s}_j) : \|\mathbf{s}_i - \mathbf{s}_j\| \in I_k\}, k = 1, \dots, K.$$

- Empirical Variogram:

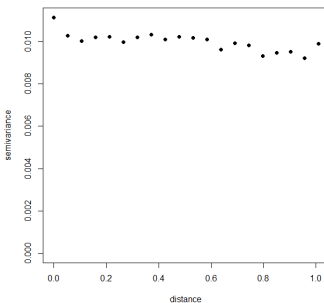
$$\gamma(t_k) = \frac{1}{|N(t_k)|} \sum_{\mathbf{s}_i, \mathbf{s}_j \in N(t_k)} (Y(\mathbf{s}_i) - Y(\mathbf{s}_j))^2$$

- Semivariogram =  $0.5 \times$  Variogram

# Empirical variogram: Data 1



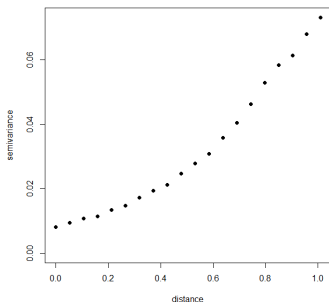
(a) Data 1: y



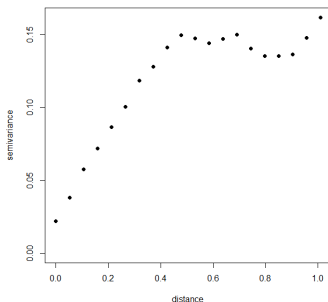
(b) Data 1: residuals

- Residuals display little spatial variation

## Empirical variograms: Data 2 and 3



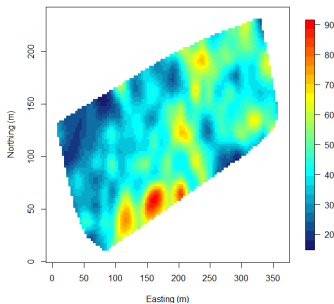
(c) Data 2: residuals



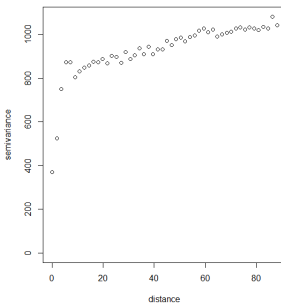
(d) Data 3: residuals

- Variogram of the residuals points to spatial variation

## EDA for WEF data



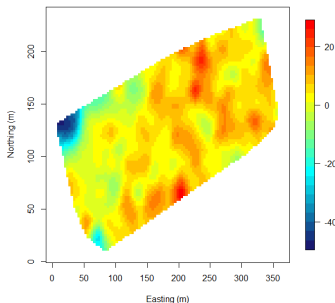
(a) DBH



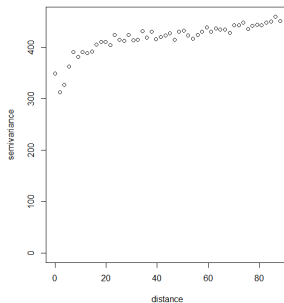
(b) Variogram of DBH

- Regression model:  $\text{DBH} \sim \text{Species}$

## EDA for WEF data



(c) DBH residuals

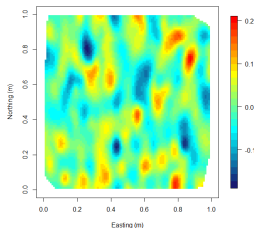


(d) Variogram of DBH residuals

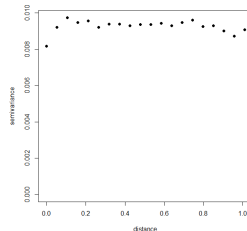
- Surface plot and variogram of residuals point to spatial variation

## Using the locations

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



(a) Residuals for data 2



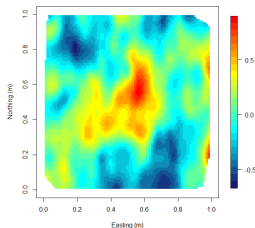
(b) Empirical variogram

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

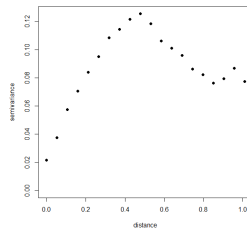
## Using the locations

- Linear regression with the co-ordinates added as regressors:

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



(a) Residuals for data 3



(b) Empirical variogram

- Dataset 3 still exhibits strong spatial correlation



## Using the locations

- 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 ?

## Using the locations

- 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 ? [Answer in next class](#)

# References

- BCG book chapters 1.1, 2.1.4, 2.5
- US forest biomass data: Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016). Hierarchical nearestneighbor gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, 111(514):800-812.
- Slovenia stomach cancer data: Zadnik V, and Reich B. Analysis of the relationship between socioeconomic factors and stomach cancer incidence in Slovenia. (2016) *Neoplasma*, 53(2):103
- EU PM<sub>10</sub> data: Datta, A., Banerjee, S., Finley, A. O., Hamm, N. A., and Schaap, M. (2016). Non-separable dynamic nearest-neighbor Gaussian Process models for spatio-temporal data with an application to particulate matter analysis. *Annals of Applied Statistics*, 10(3):1286-1316.
- Scallops data: BCG book figure 1.11(b)
- US SAT score data: BCG book Fig 4.1
- Baltimore robbery map: <http://maps.baltimorepolice.org/flexviewer/>
- Atlanta ozone data: BCG book figure 1.3
- WEF data: *spBayes* package in *R*