Overview of Spatial Statistics

Brian Reich and Safraj Shahul Hameed

North Carolina State University and the Public Health Foundation of India

May 31, 2016

SAMSI Workshop on
Statistical Methods and Analysis of Environmental Health Data

Overview

- Spatial data are everywhere in environmental applications
- With modern technology such as satellites and remote sensing, datasets are becoming larger and more precise
- The field of spatial statistics is fairly mature (methods, software, books, etc.)
- However, there is active research, especially in developing new ways to analyze massive datasets

Three types of spatial data

- **Point-referenced (geostatistical) data**: a response (e.g., PM) is measured at a finite number of spatial locations (e.g., monitor stations)
- **Areal data**: The spatial domain is partitioned into a finite number of regions (e.g., states) and a single summary of each region is recorded (e.g., percent unemployed)
- **Point-pattern data**: The spatial location of an event (e.g., earthquakes) is the response of interest
- There are different (connected) tools for each data type
- We will focus on point-referenced data

Common objectives

- Test for spatial correlation
- Estimate the range of spatial correlation
- Estimate the effects of covariates while accounting for residual spatial dependence
- Predict and map (with uncertainty) the response at unmonitored locations

Plotting spatial data

- `R` has many nice spatial packages including:
  - `maps`: standard mapping and projection tools
  - `fields`: useful tools for plotting and manipulating spatial data
  - `ggplot2`: general plotting tools with nice spatial functions
- The two main types of maps are the values at the monitoring locations and a map of predicted values
- Example: `http://www4.stat.ncsu.edu/~reich/workshop/Ozone_Example.html`

Fitting a spatial model

$$Y(s) = X(s)^T \beta + \varepsilon(s)$$

- $Y(s)$ is the response at spatial location s
- $X(s)$ are covariates at s (e.g., temperature or elevation)
- β are the regression coefficients, interpreted the same as in non-spatial linear regression
- $\varepsilon(s)$ is the Gaussian residual
- This is standard linear regression if the residuals are independent

Fitting a spatial model

- In a spatial model the residuals $\varepsilon(s)$ are not assumed to be independent
- We model the correlation between at two sites as a decreasing function of the distance between sites
- The residuals are split into two components

$$\varepsilon(s) = \theta(s) + \epsilon(s)$$

- **Nugget**: The pure (uncorrelated) measurement error is

$$\epsilon(s) \overset{iid}{\sim} \text{Normal}(0, \tau^2)$$

- The spatial errors $\theta(s)$ are correlated

Fitting a spatial model

- **Partial sill**: The variance of the spatial errors is

$$\text{Var}[\theta(s)] = \sigma^2$$

- **Sill**: The total variance is

$$\text{Var}[\varepsilon(s)] = \sigma^2 + \tau^2$$

- Most analyses assume the correlation between points is:
  - **Stationary**: the same throughout the spatial domain
  - **Isotropic**: the same for all angles
- In this case the correlation between the residuals at sites s and t is a function of only the distance between sites, *d*

Fitting a spatial model

- There are many correlation functions (Matern, powered-exponential, spherical, etc.)
- We will use the exponential correlation

$$\text{Cor}[\theta(s), \theta(t)] = \exp\left(-\frac{d}{\phi}\right)$$

- Correlation decays exponentially with distance, d
- **Range**: the parameter ϕ controls the range of spatial correlation

Fitting a spatial model

- The parameters β, σ^2, τ^2 and ϕ can be estimated using maximum likelihood estimation
- The `R` package `GeoR` can be used
- Estimation can be slow for large datasets because the likelihood involves large matrices
- Example: `http://www4.stat.ncsu.edu/~reich/workshop/Ozone_Example.html`

Spatial prediction

- We use the observed data at the monitors to estimate the model parameters
- Once we have parameter estimates, we can make predictions at other locations
- There are many ways to do this: nearest neighbor, average of observations in a window, etc
- **Kriging** is the optimal method in the sense that it is the Best Linear Unbiased Predictor (BLUP)

Spatial prediction

- The Kriging prediction at location s_0 given the data at $s_1, \ldots, s_n$ is

$$\hat{Y}(s_0) = X(s_0)^T \beta + \sum_{i=1}^{n} \lambda_i [Y(s_i) - X(s_i)^T \beta]$$

- The prediction is a linear combination of the residuals
- The weights λ_i are determined by the spatial correlation
- Intuitively, points close to s_0 are weighted highest

Spatial prediction

- Prediction standard deviations have a similar form
- The `R` package `GeoR` performs Kriging
- To make a map, you apply Kriging to a fine grid of points covering the area of interest
- Example: `http://www4.stat.ncsu.edu/~reich/workshop/Ozone_Example.html`

Spatiotemporal data

- A natural extension is to processes that evolve over space and time
- For example, $Y(\mathrm{s}, t)$ is the PM at location s and day t
- The methods are very similar to those discussed above
- The main difference is that we need to estimate both the correlation across space and the correlation across time
- Kriging weights observations in space and time based on the relative strength of the two types of correlation

Other extensions

- Multivariate spatial analysis of multiple outcomes, e.g., PM and ozone
- Non-Gaussian data, e.g., counts or binary outcomes
- Spatially-varying coefficients, e.g., $\beta(s)$
- More sophisticated models such as nonstationary covariance functions
- Spatial analysis of extreme values
- Methods to handle large n
- Many more!

Resources

- Books on theory: Cressie (1993), Stein (1999)
- Book on applied methods for health data: Waller and Gotway (2004)
- Book on recent methods: Handbook of Spatial Statistics (2010)
- Book on spatiotemporal data: Wikle and Cressie (2011)
- More computing: `geoRglm`; `OpenBUGS`; `Proc Mixed`
- My info: `http://www4.stat.ncsu.edu/~reich/`