Intro to Spatial Statistics

Mevin Hooten

Department of Statistics and Data Sciences
The University of Texas at Austin

Overview

- Motivation.
- Overview of Spatial Statistics.
- Continouous Spatial Processes:
  - Spatial correlation and covariance functions.
  - Simulating spatial random processes.
  - Assumptions.
  - Small-scale variability.
- Assessing spatial dependence.
  - Variograms and Covariograms.
- Geostatistical Modeling.
  - Bayesian model formulation.
  - Bayesian Kriging.
  - Generalized spatial models.

Hooten et al. (2003)

Landscape Ecology **18:** 487–502, 2003.

Research article

Predicting the spatial distribution of ground flora on large domains using a hierarchical Bayesian model

Mevin B. Hooten[1,*], David R. Larsen[2] and Christopher K. Wikle[1]
[1]Department of Statistics, University of Missouri, 222 Mathematical Sciences Building, Columbia, Missouri 65211, USA; [2]Department of Forestry, University of Missouri, Columbia, Missouri 65211, USA;
**Author for correspondence (e-mail: hooten@stat.missouri.edu)*

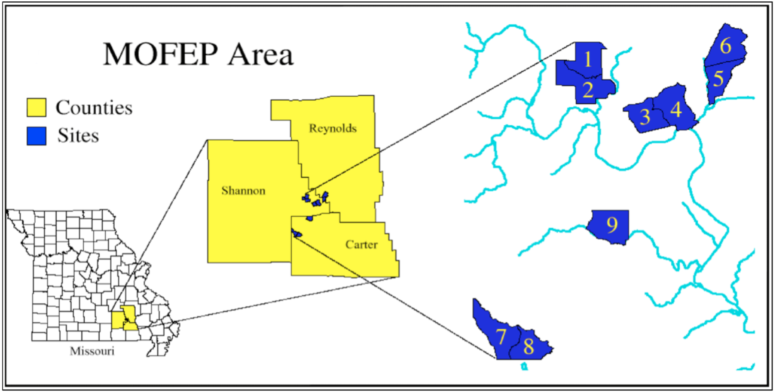
Received 5 July 2002; accepted in revised form 28 March 2003

Key words: Bayesian statistics, Hierarchical Bayesian models, Landscape vegetation prediction, Spatial modeling, Missouri, USA, Ozark Highlands

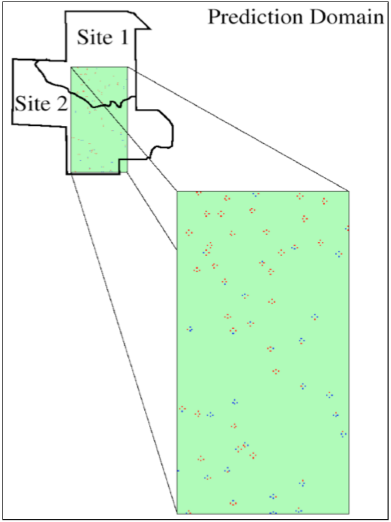
Abstract

Accomodation of important sources of uncertainty in ecological models is essential to realistically predicting ecological processes. The purpose of this project is to develop a robust methodology for modeling natural processes on a landscape while accounting for the variability in a process by utilizing environmental and spatial random effects. A hierarchical Bayesian framework has allowed the simultaneous integration of these effects. This framework naturally assumes variables to be random and the posterior distribution of the model provides probabilistic information about the process. Two species in the genus *Desmodium* were used as examples to illustrate the utility of the model in Southeast Missouri, USA. In addition, two validation techniques were applied to evaluate the qualitative and quantitative characteristics of the predictions.

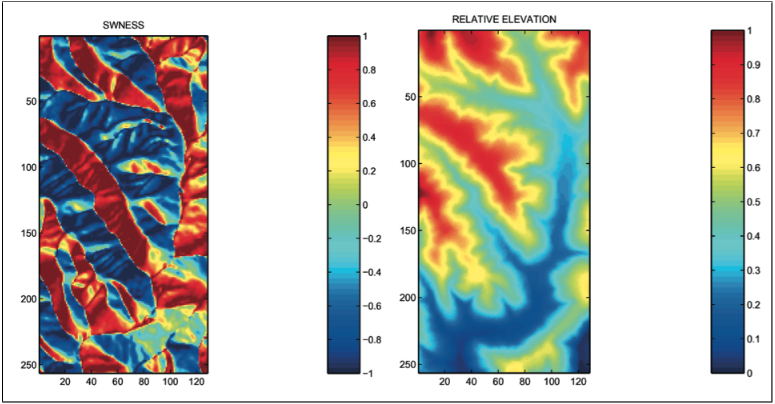
Hooten et al. (2003)



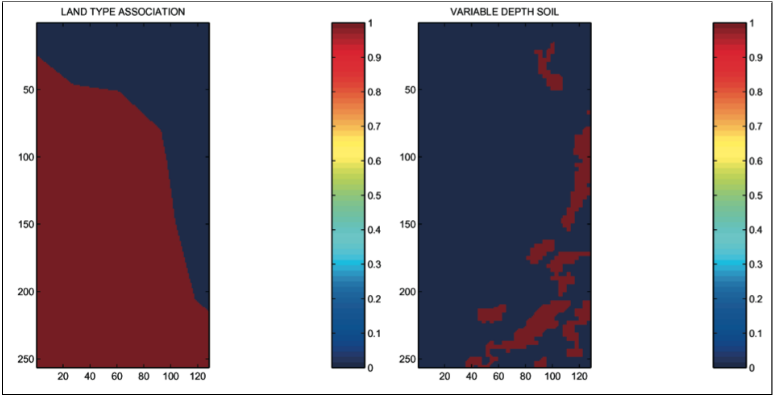
Hooten et al. (2003)



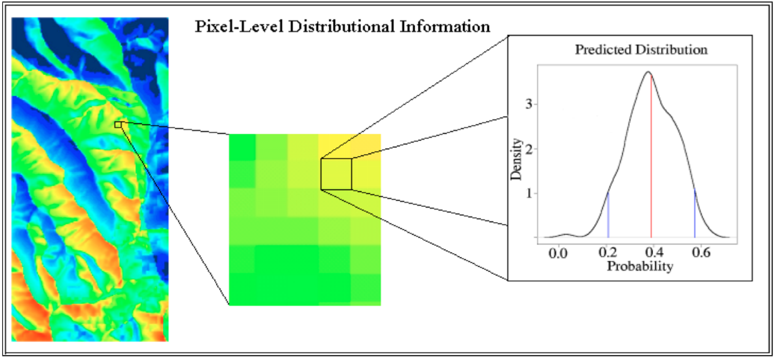
Hooten et al. (2003)



Hooten et al. (2003)



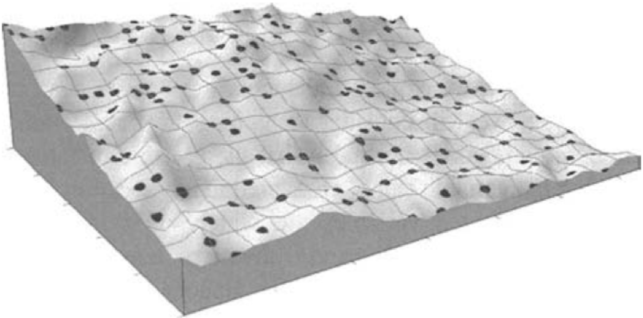
Hooten et al. (2003)



Spatial Processes

1. **Spatial Point Processes:** Random locations are of interest, sometimes associated point characteristics ("marks").
2. **Continuous Spatial Processes:** Random measurements at fixed locations are of interest.
3. **Areal Spatial Processes:** Random measurements in fixed regions are of interest.

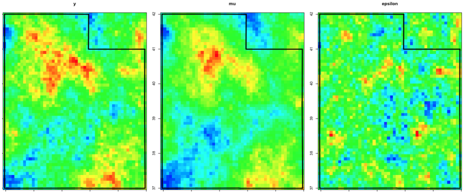
Continuous Spatial Processes



Imagine a smooth 2-D function

$y(\mathbf{s}) = \mu(\mathbf{s}) + \varepsilon(\mathbf{s})$, where $\mathbf{s} \in \Re$

1. μ: First order, the mean effect, a trend.
2. ε: Second order, often thought of as correlated error.



Gaussian Spatial Regression

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathsf{N}(\mathbf{0}, \boldsymbol{\Sigma})$$

Gaussian Spatial Regression

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$$

1. **First Order Structure:** $\mathbf{X}\boldsymbol{\beta}$, the trend.
2. **Second Order Structure:** $\boldsymbol{\varepsilon}$, where $\boldsymbol{\Sigma}$ can explain various forms of spatial autocorrelation.
3. **Prediction:** Kriging.

Note: This is referred to as “model-based geostatistics.”

Covariance function: Covariogram

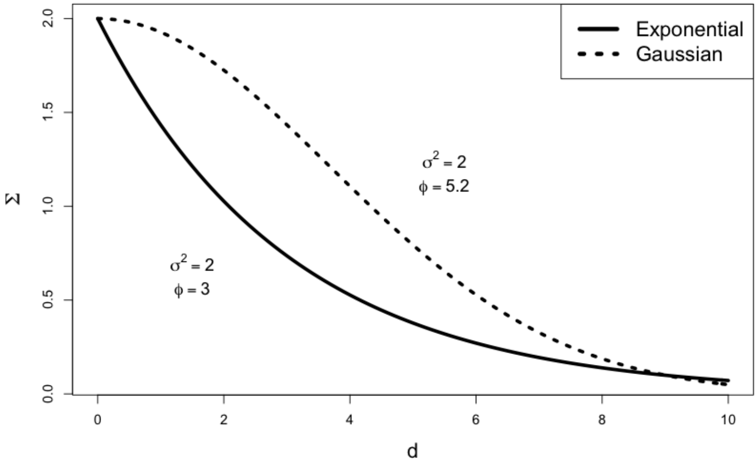
Parametric Covariance Functions:

- **Exponential**: $\Sigma_{i,j} = \sigma^2 \exp\left(-\frac{d_{i,j}}{\phi}\right)$

- **Gaussian**: $\Sigma_{i,j} = \sigma^2 \exp\left(-\frac{d_{i,j}^2}{\phi^2}\right)$

Note: $d_{i,j} =$ distance between locations i and j.

Parameteric Covariance Functions



Important Assumptions

- **Stationarity**: spatial structure does not vary with location.

- **Isotropy**: spatial structure does not vary with direction.

Two Sources of Error

Random Effects Approach:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\eta} + \boldsymbol{\varepsilon}$$

1. **Correlated Error:** $\boldsymbol{\eta} \sim \text{N}(\mathbf{0}, \boldsymbol{\Sigma})$
2. **Uncorrelated Error:** $\boldsymbol{\varepsilon} \sim \text{N}(\mathbf{0}, \sigma_\varepsilon^2\mathbf{I})$

Two Sources of Error

Hierarchical Approach:

$$\mathbf{y} \sim \mathsf{N}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\eta}, \sigma_{\varepsilon}^2 \mathbf{I})$$

$$\boldsymbol{\eta} \sim \mathsf{N}(\mathbf{0}, \boldsymbol{\Sigma})$$

Two Sources of Error

Hierarchical Approach:

$$\mathbf{y} \sim \mathsf{N}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\eta}, \sigma_{\varepsilon}^2 \mathbf{I})$$

$$\boldsymbol{\eta} \sim \mathsf{N}(\mathbf{0}, \boldsymbol{\Sigma})$$

These both imply:

$$\mathbf{y} \sim \mathsf{N}(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma} + \sigma_{\varepsilon}^2 \mathbf{I})$$

Simulate a correlated continuous spatial process

1. Choose locations s_i for $i = 1, \ldots, n$.
2. Choose the mean $\boldsymbol{\mu}$. This could be a scalar or it could vary spatially.
3. Choose range parameter ϕ and variance component σ^2.
4. Compute distance matrix $\mathbf{D}$ between all n locations of interest.
5. Calculate covariance matrix $\boldsymbol{\Sigma} = \sigma^2 \exp\left(-\frac{\mathbf{D}}{\phi}\right)$.
6. Sample the n-dimensional vector $\mathbf{y} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

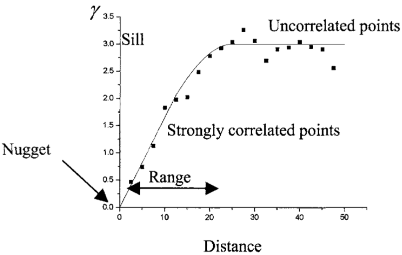
Assess the spatial correlation in a data set

1. Assume $\mathbf{y}$ is measured at n spatial locations.
2. Compute the residuals: $\mathbf{e} = \mathbf{y} - \boldsymbol{\mu}$.
3. Examine the residuals $\mathbf{e}$ for spatial correlation (i.e., autocorrelation).

Estimating spatial correlation

Empirical Semi-Variogram:

$$\hat{\gamma}(d) = \frac{\sum(e_i - e_j)^2}{2N(d)}$$



Fitted Variogram

Classic Estimation:

- After the empirical variogram is estimated at several bins for d, one can fit a parametric model to it.
- In this case, use $\hat{\gamma}(d)$ as the response variable and d as the covariate in weighted least squares or nonlinear regression to estimate σ^2, ϕ.

Semi-Variogram, Variogram, and Covariogram

- Semi-Variogram: $\gamma(d)$
- Variogram: $2\gamma(d)$
- Covariogram: $\text{cov}(d) = \text{cov}(0) - \gamma(d)$

Bayesian Geostatistical Model

- **Goal**: use Bayesian methods to estimate $\boldsymbol{\beta}$, σ^2, and ϕ.

$$\mathbf{y} \sim \mathsf{N}(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma})$$

- $\Sigma_{i,j} = \sigma^2 \exp\left(-\frac{d_{i,j}}{\phi}\right)$.
- $\boldsymbol{\beta} \sim \mathsf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
- $\sigma^2 \sim \mathsf{IG}(q, r)$.
- Many choices for $\phi \sim [\phi]$.

Posterior:

$$[\boldsymbol{\beta}, \sigma^2, \phi|\mathbf{y}] = c \times [\mathbf{y}|\boldsymbol{\beta}, \sigma^2, \phi][\boldsymbol{\beta}][\sigma^2][\phi]$$

Prior Selection

Choices for range parameter ϕ:

- $\phi \sim$ Gamma(γ_1, γ_2)
- $\log(\phi) \sim$ N$(\mu_\phi, \sigma^2_\phi)$
- $\phi \sim$ DiscUnif(Φ)
- $\phi \sim$ Half-Cauchy(γ)

Bayesian Kriging

- **Goal**: predict $y(\mathrm{s}_u)$ at unobserved location s_u, given the model and the data $y(\mathrm{s}_i)$ for $i = 1, \ldots, n$.

$$y(\mathbf{s}_i) = \mathbf{x}(\mathbf{s}_i)'\boldsymbol{\beta} + \varepsilon(\mathbf{s}_i)$$

- We need the posterior predictive distribution:

$$[y_u|\mathbf{y}] = \int\int\int [y_u|\mathbf{y}, \boldsymbol{\beta}, \sigma^2, \phi][\boldsymbol{\beta}, \sigma^2, \phi|\mathbf{y}]d\boldsymbol{\beta}d\sigma^2 d\phi$$

Predictive Full-Conditional

Notice that:

- $[y_u|\mathbf{y}, \boldsymbol{\beta}, \sigma^2, \phi] = N(\tilde{\mu}, \tilde{\sigma}^2)$

where,

- $\tilde{\mu} = \mathbf{x}_u' \boldsymbol{\beta} + \mathbf{c}' \boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$
- $\tilde{\sigma}^2 = \sigma^2 - \mathbf{c}' \boldsymbol{\Sigma}^{-1} \mathbf{c}$

and,

- $\mathbf{c} = (c_1, \ldots, c_n)'$
- $c_i = \text{cov}(\varepsilon_u, \varepsilon_i)$

- In MCMC, sample $y_u^{(k)} \sim \text{N}(\tilde{\mu}^{(k)}, \tilde{\sigma}^{2(k)})$.
- The Bayesian Kriging predictor is: $\text{E}(y_u|\mathbf{y}) \approx \sum_{k=1}^{K} y_u^{(k)}/K$.

Generalized Spatial Models

- Binary:

$$y_i \sim \text{Bern}(p_i)$$
$$\text{logit}(p_i) = \mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i$$

- Count:

$$y_i \sim \text{Pois}(\lambda_i)$$
$$\log(\lambda_i) = \mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i$$

Spatial Occupancy Model

$$y_i \sim \begin{cases} 0 & , z_i = 0 \\ \text{Binom}(J_i, p_i) & , z_i = 1 \end{cases}$$

$$z_i \sim \text{Bern}(\psi_i)$$

- $\text{logit}(p_i) = \mathbf{w}_i'\boldsymbol{\alpha} + \eta_i$
- $\text{logit}(\psi_i) = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$
- $\boldsymbol{\eta} \sim \text{N}(\mathbf{0}, \boldsymbol{\Sigma}_\eta)$
- $\boldsymbol{\varepsilon} \sim \text{N}(\mathbf{0}, \boldsymbol{\Sigma}_\varepsilon)$

References

- Hooten, M.B. and T.J. Hefley. (2019). Bringing Bayesian Models to Life. CRC Press.
- Hooten, M.B., D.S. Johnson, B.T. McClintock, and J.M. Morales. (2017). Animal Movement: Statistical Models for Telemetry Data. CRC Press.
- Johnson, D.S., P.B. Conn, M.B. Hooten, J. Ray, and B. Pond. (2013). Spatial occupancy models for large data sets. Ecology, 94: 801-808.
- Hooten, M.B., Larsen, D.R., and C.K. Wikle, (2003). Predicting the spatial distribution of ground flora on large domains using a hierarchical Bayesian model. Landscape Ecology, 18: 487-502.