

Proposal: Bayesian Models for Environmental Spatial Data analysis with R

Andrew O. Finley and Sudipto Banerjee

April 22, 2009

1 General overview

This is a proposal for the first edition of the book “Bayesian Models for Environmental Spatial Data analysis with R,” to be authored by Dr. Andrew O. Finley at Michigan State University and Dr. Sudipto Banerjee at University of Minnesota.

Recent advances in Geographical Information Systems (GIS) and Global Positioning Systems (GPS) enable accurate geocoding of locations where scientific data are collected. This has encouraged collection of spatiotemporal datasets in many fields and has generated considerable interest in statistical modeling for time and location-referenced data. This accumulation of data and need for analysis is especially common in the broad fields of forestry and ecology. In these fields, spatially and temporally indexed data, typically consisting of one or more response variables and associated covariates, are used to estimate natural resource inventory, presence/absence, counts, and change. In these settings, the focus of inference is often on specific model parameters and/or subsequent prediction at a new location or time. In these modeling exercises, rarely is it safe, or even desirable, to assume that model residuals are *independent* and *identically* distributed. The propensity to violate these assumptions is especially great in environmental datasets because the data often exhibit temporal, spatial, or hierarchical structure, or all three¹.

This book details recent advancements in statistical (and in particular Bayesian) computing pertaining to hierarchical random effects models using

¹“The first law of ecology is that everything is related to everything else.” *The Closing Circle: Nature, Man, and Technology*. New York : Knopf, 1971.

Markov chain Monte Carlo (MCMC) methods. In the spirit of the **useR** series, the book will emphasize programming, computing, and data analysis within the **R** statistical environment. The modeling focus will be on generalized linear model frameworks that accommodate spatial and temporal associations. Diverse settings for spatial and spatiotemporal models are considered, mostly motivated by a range of studies that employ forestry and ecological datasets.

2 How the book differs from its competitors

Spatial statistics continue to be an area of very active research and, therefore, it is no surprise that a number of new texts on the subject continue to enter the market. We believe our proposed book will be the only text on spatial statistics offering in-depth spatial statistical modeling within **R** from a fully Bayesian standpoint. Further, the focus on environmental data further differentiates our book from its competitors. Perhaps the closest competitor is offered by Bivand, Pebesma, and Gómez-Rubio's recent book (*Applied Spatial Data Analysis with R*, 2008). However, other than a brief overview of spatial data input/output, our texts will have very little overlap. In fact we believe our text will nicely complement Bivand et al., offering an extended hierarchical modeling perspective on their chapter 8, *Interpolation and Geostatistics*. Similar to Bivand et al., we will illustrate several supporting **R** packages (e.g., `geoR`, `geoRglm`, `sp`, `fields`, `rgdal`, `gstat`, and others that appear in the **R** CRAN Analysis of Spatial Data <http://cran.r-project.org/web/views/Spatial.html>); however, we will emphasize the use of our `spBayes` package for hierarchical modeling. Like the proposed text, `spBayes` is unique because it offers a suite of functions to fit fully Bayesian models. Several less computing oriented texts exist. For instance, the book by Schabenberger and Gotway (*Statistical methods for spatial data analysis*, Chapman and Hall/CRC, 2004), has a predominantly non-Bayesian focus. Furthermore, its coverage is not as broad as the one being proposed here, which adds multivariate spatial models, spatial predictive process models, and spatial survival models. Waller and Gotway (*Applied Spatial Statistics for Public Health Data*, 2004) is another excellent text, but focuses almost exclusively on public health applications. LeSage and Pace (*Introduction to Spatial Econometrics*, 2009, CRC Press) is another new text that provides excellent coverage of spatial models in econometrics. They dis-

cuss likelihood and Bayesian methods but they focus predominantly upon the spatial autoregression models and their extensions. The emphasis, as is evident from the title itself, is on econometric models and their applications. The text by Diggle and Ribeiro (*Model-based Geostatistics*, 2008, Springer) is a clear competitor to ours, providing in-depth discussions of geostatistics or point-referenced data models, but offers relatively few computing exercises (compared to a **useR** book). The texts by Wackernagel (*Multivariate Geostatistics*, 2004, Springer) and Möller and Waagepetersen (*Statistical Inference and Simulation for Spatial Point Processes*, 2003, Chapman and Hall/CRC) limit themselves almost exclusively to multivariate and point-process models, respectively, and offer little or no computing. Finally, the text by Banerjee, Carlin, and Gelfand (*Hierarchical Modeling and Analysis for Spatial Data*, 2004) is another clear competitor to our proposed book; however, our emphasis on R programming, computing, and modeling of environmental data differentiates the texts.

3 Proposed Table of Contents

The following provides an outline of our plan for the chapters. The approximate page numbers are also indicated. We do not expect the total number of pages to exceed 200, inclusive of appendices and indices.

1. Introduction

- 1.1 Motivating examples from environmental sciences
- 1.2 Introduction to spatial data and models
- 1.3 R programming and statistical pre-requisites

2. Bayesian Hierarchical models

- 2.1 Introduction to hierarchical modeling and Bayes' Theorem
- 2.2 Bayesian linear regression
 - 2.2.1 Analysis with conjugate NIG priors
 - 2.2.2 Analysis with flat priors – classical analysis
 - 2.2.3 Prediction
 - 2.2.4 Bayesian model comparisons

2.3 Bayesian computation

2.3.1 The Gibbs sampler

2.3.2 The Metropolis-Hastings algorithm

2.3.3 Adaptive Markov Chain Monte Carlo

2.3.4 Convergence diagnosis

3. **Spatial data compilation in R**

3.1 Import and Export

3.2 Geographic distance and projection

3.2.1 Distance computations

3.2.2 Spatial data projection

3.3 Exploratory data analysis

3.3.1 Non-spatial exploratory analysis

3.3.2 Spatial data exploratory analysis

3.3.3 Representing continuous spatial data

4. **Model-based geostatistics**

4.1 Elements of point-referenced modeling

4.1.1 Stationarity

4.1.2 Variograms

4.1.3 Isotropy

4.1.4 Variogram model fitting

5. **Hierarchical spatial process models**

5.1 Ingredients for spatial models

5.2 Formal modeling theory for spatial processes

5.2.1 Covariance functions

5.2.3 Simulating spatial data

5.3 Gaussian spatial models

5.3.1 Exact Bayesian inference

5.3.2 Bayesian inference using MCMC

- 5.4 Non-Gaussian spatial models
- 5.5 Gaussian predictive process models
 - 5.5.1 Low-rank kriging and the predictive process
 - 5.5.2 Bias-adjusted predictive process models
 - 5.5.3 Knot selection

6. **Multivariate spatial models**

- 6.1 Multivariate spatial processes
- 6.2 Multivariate Gaussian spatial models
 - 6.2.1 Missing observations
- 6.3 Multivariate non-Gaussian spatial models
- 6.4 Multivariate predictive process

7. **Advanced topics**

- 7.1 Spatially varying coefficient models
 - 7.1.1 Univariate models
 - 7.1.2 Multivariate models
- 7.2 Spatiotemporal models
 - 7.2.1 Covariance functions
 - 7.2.2 Nonseparable spatiotemporal model
- 7.3 Spatial survival and frailty models
- 7.4 Spatial point-process models

Appendices

- A **Building a high performance R computing environment**
- B **Essentials of matrix theory**

References

Author Index

Subject Index