Hierarchical Bayesian Modeling and Analysis for Spatial BIG Data

Sudipto Banerjee, University of California Los Angeles

Abstract: In this course we will describe hierarchical modeling and related Markov chain Monte Carlo (MCMC) methods for spatial statistics, with special emphasis on methods for analyzing very large or "BIG" spatial datasets. We will describe both exploratory data analysis tools and traditional modeling approaches for different types of spatial data. Our approach will be fully model-based through the use of Gaussian processes. Therefore, we will develop the basics of spatial Gaussian process models. Approaches from traditional geostatistics (variogram fitting, kriging, etc.) will be briefly covered here. We then turn to areal data models, again starting with exploratory displays and progressing towards more formal model specifications, e.g., Markov random fields that underlie the conditional, intrinsic, and simultaneous autoregressive (CAR, IAR, and SAR) models widely used in areal data settings. The remainder of our presentation will cover hierarchical modeling for both univariate and multivariate spatial response data, including Bayesian kriging and lattice modeling, as well as more advanced issues pertaining to BIG data sets. We also discuss modern computational approaches for very large spatial and spatiotemporal data sets. Short course participants should have an M.S. understanding of mathematical statistics at, say, the Hogg/Craig/Tanis or Casella/Berger levels, as well as basic familiarity with Bayesian modeling and computing at the Carlin/Louis or Gelman et al. levels. We will not assume any significant previous exposure to spatial or spatiotemporal methods.

The course will comprise roughly three lectures.

Lecture 1: Modeling and analysis for point-referenced data (Geostatistics).
This lecture will focus upon the modeling and analysis for point-referenced geographic data referenced by the coordinates (e.g., latitude-longitude, Easting-Northing) where the variables (outcomes, predictors and so on) have been observed or measured. The lecture will discuss the primary inferential questions for point-referenced data and introduce the concept of a spatial process. The spatial process will be shown to be a model for point-referenced data and the lecture will illustrate its use in answering several inferential questions (e.g., interpolation or "kriging", and inference on spatial range, variances and measurement errors) in classical geostatistics. The fully model-based approach will be presented within a Bayesian hierarchical framework and computational methods for fitting such models will be discussed and demonstrated using readily available software packages for spatial data analysis in the R statistical computing environment.

Lecture 2: Modeling and analysis for high-dimensional spatial and spatiotemporal data.
With rapid developments in GIS, statisticians today routinely encounter spatial and temporal data containing observations from a large number of spatial locations and time points. However, fitting hierarchical spatial-temporal models often involves expensive matrix computations with complexity increasing in cubic order for the number of spatial locations and temporal points. This renders such models unfeasible for large data sets. This lecture will present two approaches for constructing well-defined spatial-temporal stochastic processes that accrue substantial computational savings. Both these processes can be used as "priors" for spatial-temporal random fields. The first approach constructs a low-rank process operating on a lower-dimensional subspace. The second approach constructs a Nearest-Neighbor Gaussian Process (NNGP) that can be exploited as a dimension-reducing prior embedded within a rich and flexible hierarchical framework to deliver exact Bayesian inference.

Lecture 3: Modeling and analysis for areally-referenced data.
This lecture will focus upon the modeling and analysis for regionally aggregated geographic data referenced by counties, states, zip codes, census tracts or other such administrative clusters over which the data have been collected. Such data are inherently clustered by sampling design and one needs to account for spatial dependence across these clusters. This lecture will introduce spatial Markov random fields and show how these can be used to capture associations arising from different spatial sampling designs for regionally aggregated data. Bayesian algorithms and software for their implementation using R packages as well as the BUGS language will be discussed.