Guest Editor's Introduction to the Special Issue on "Modern Dimension Reduction Methods for Big Data Problems in Ecology"

Christopher K. WIKLE, Scott H. HOLAN, and Mevin B. HOOTEN

With an ever-increasing amount of ecological data from remote sensing, long-term networks, long-term surveys, and computer models, there is a need to develop efficient statistical methods that can accommodate the unique dependence structures associated with ecological inference and prediction. Indeed, as the volume of such "big data" increases, scientists are interested in addressing increasingly complex questions—particularly those accounting for spatio-temporal dependence across multiple scales, as well as multivariate community-level responses. For this invited special issue, we have sought contributions from many of the leading researchers at the interface of statistics and ecology.

The methodological advancements presented here can all be characterized in some sense as "dimension reduction." The methods are concerned with multivariate and/or spatio-temporal processes and data, with the associated dimension reduction being either in terms of the parameters, state-process, or grouping. The challenges with high dimensionality are, in some cases, further increased by the non-Gaussian and/or nonlinear nature of the data and processes of interest.

Guhaniyogi et al. address several of these themes. In particular, they are concerned with the issue of inference in a high-dimensional multivariate spatial setting with spatial nonstationarity. They provide a computationally feasible approach for drawing inference in such environments in a manner that accommodates dimension reduction in both the parameter and data space.

Johnson et al. reduce the dimensionality through a clustering perspective designed to characterize metapopulation abundance trends using spatio-temporal data. In particular,

Christopher K. Wikle (✉) is Professor (E-mail: *wiklec@missouri.edu*), Scott H. Holan is Associate Professor (E-mail: *holans@missouri.edu*), Department of Statistics, University of Missouri, Columbia, MO, USA. Mevin B. Hooten is Associate Professor U.S. Geological Survey, Colorado Cooperative Fish and Wildlife Research Unit, Department of Fish, Wildlife, and Conservation Biology and Department of Statistics, Colorado State University, Fort Collins, CO, USA (E-mail: *hooten@rams.colostate.edu*).

Journal of Agricultural, Biological, and Environmental Statistics, Volume 18, Number 3, Pages 271–273
DOI: 10.1007/s13253-013-0151-0

rather than build computationally expensive spatial dependencies in the model, they group "common" sites together based on their temporal variability. This has the benefit of providing a nonparametric spatial association (i.e., no spatial model need be specified) and it reduces the dimension of the spatio-temporal trend process.

Leininger et al. are concerned with modeling spatially extensive classification data (e.g., land-use classifications). In addition to the high-dimensional multivariate outcome in space, they are also concerned with accommodating spatially extensive covariates, such as those associated with topography and census information. Finally, parallelization is utilized to improve the computational efficiency of their model.

Wu et al. present methodology to handle large spatio-temporal count datasets in which the observations can be either over- or under-dispersed. They utilize a hierarchical Bayesian Conway-Maxwell Poisson (CMP) model that allows for dispersion to vary dynamically. In addition, to handle nonlinear spatio-temporal dynamics, they incorporate a threshold autoregressive model for the CMP intensity parameter and, in their waterfowl settling pattern prediction, provide scientifically motivated regime switching based on climate conditions. Nonlinear spatial dimension reduction is facilitated through kernel principal component analysis.

Dunstan et al. are concerned with data arising from community-level or multi-species studies, where data are simultaneously collected on many species at the same set of sites. The goal of such analyses is to study how communities of species respond to the environment. They consider a finite mixture of regression models to classify species into one of a small number of archetypal forms of environmental response. They demonstrate that this method has improved predictive performance when compared to analyses of presence/absence data alone.

Clark et al. are concerned with a different type of dimensionality reduction that arises when one is attempting a sensitivity analysis of environmental models. Such analysis is challenging due to potentially large numbers of input and output variables and their associated feedbacks between them. This paper considers "dynamic inverse prediction," which facilitates interpretability and sensitivity analysis of variables associated with inputs and unobserved interval processes.

Hooten et al. are concerned with dimension reduction in the situation where one is interested in using a scientifically motivated mechanistic model to facilitate spatio-temporal process modeling. In particular, they are concerned with diffusion partial differential equations (PDEs) that are, by definition, specified in continuous time and space and must be solved numerically. In a traditional statistical estimation environment associated with high-dimensional spatial fields, such solutions can be computationally prohibitive given the need to evaluate the solution many times. Using a mathematical method of multiple scales called homogenization, they show that, for some PDE models, a natural statistical scaling procedure can be obtained that is consistent with the implied dynamics of the mechanistic model. In the case of ecological diffusion models specifically, they demonstrate a harmonic averaging technique that provides an optimal approximation to the PDE solver, which in turn dramatically improves computational efficiency.

Fu et al. are concerned with a binary response data over large spatial domains. They consider an autologistic regression framework that allows covariates. Traditionally, such

models can be computationally expensive when properly accounting for spatial dependence, particularly in the context of model selection. By utilizing asymptotic theory and regularization methods, they develop a penalized pseudolikelihood method for simultaneous variable selection and parameter estimation that is computationally feasible for large spatial datasets.

Yang et al. consider a different type of dimensionality issue. In particular, as high-frequency time signals become increasingly more common (e.g., from monitoring instruments or social media streams), one is increasingly interested in using these signals as covariates or predictors in ecological studies. In order to effectively utilize these signals in this context, this paper develops a class of nonlinear multivariate time-frequency functional models. The proposed methods utilize low-rank basis expansions and stochastic search variable selection to effectively reduce the dimensionality and to identify important time-frequency (and, hence, time-domain) features.

We believe that the contributions in this special issue provide an informative snapshot of research at the interface of statistics and ecology that is useful in addressing "big data" issues. This is not meant to be an exclusive or exhaustive collection of such papers, but we believe it does illustrate the depth and breadth of such research at the interface of the disciplines. It is our desire that these papers may entice others to see potential applications of their "big data" methodology to problems of interest to ecologists. We are certain that the contributions presented here, although a good start, have only scratched the surface in terms of the statistical issues associated with "big data" in ecology and we look forward to seeing many more such contributions in the future.

[Published Online July 2013.]