nonlinear inversion: how many parameters can we estimate and which measurements are most useful? Global Change Biology 7:495–510.

Williams, M., P. A. Schwarz, B. E. Law, J. Irvine, and M. R. Kurpius. 2005. An improved analysis of forest carbon dynamics using data assimilation. Global Change Biology 11:89–105.

Xu, T., L. White, D. Hui, and Y. Luo. 2006. Probabilistic inversion of a terrestrial ecosystem model: analysis of uncertainty in parameter estimation and model prediction. Global Biogeochemical Cycles 20. [doi: 10.1029/2005GB002468]

Ecological Applications, 19(3), 2009, pp. 574–577
© 2009 by the Ecological Society of America

The importance of accounting for spatial and temporal correlation in analyses of ecological data

Jennifer A. Hoeting[1]

Department of Statistics, Colorado State University, Fort Collins, Colorado 80523-1877 USA

I congratulate Cressie, Calder, Clark, Ver Hoef, and Wikle for their insightful overview on hierarchical modeling in ecology. These authors are all experts in this field and have contributed to many of the advances that have been made in the field of ecological modeling. Spatial and temporal correlation is a major theme in Cressie et al. (2009). Below I expand on these ideas, discussing some of the advantages and disadvantages of accounting for spatial and/or temporal correlation in analyses of ecological data. While much of the focus in this discussion is on spatial statistical models, similar problems occur when temporal or spatiotemporal correlation is ignored.

An Example of a Statistical Model that Accounts for Spatial Correlation

Unless data are observed within a very specific experimental design, ecological data are often correlated. As an example, consider the problem of estimating stream sulfate concentrations in the eastern United States. We consider data collected as part of the EPA's Environmental Monitoring and Assessment Program (EMAP). The sample sites were mainly located in Pennsylvania, West Virginia, Maryland, and Virginia. For more details about this example and the issues described below, see Irvine et al. (2007).

The response $\mathbf{Y} = [Y_1, \ldots, Y_n]'$ is stream sulfate concentration at each of the n stream sites. In this simplified example we consider four predictors $[\mathbf{X}_1, \ldots, \mathbf{X}_4]$ which are geographic information system (GIS) derived covariates of the percentage of landscape covered by forest, agriculture, urban, and mining within the watershed above each stream site. To investigate the relationship between these covariates and the response, we might first consider a standard multiple regression model for $i = 1, \ldots, n$ given by

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \varepsilon_i \qquad (1)$$

where X_{ij} is the ith observation from the jth predictor and the error terms ε_i are independent normally distributed random variables with mean 0 and variance σ^2. If we follow through with this analysis, we might find that various predictors have statistically significant effects, and we might provide some estimate of the variance, σ^2. We might also undertake some form of model selection to determine which covariates best explain observed patterns of stream sulfate concentration. However, what if, after accounting for all available covariates, the errors are not independent? What if there is remaining spatial correlation, such that observations that are in close proximity in space are related, and that these predictors haven't fully accounted for such correlation? In this case, we can learn a lot from modeling the nonindependent errors.

As an alternative to a multiple regression model with independent errors, we might consider a spatial regression model where the errors are assumed to be spatially correlated. In this case, we could use the model in Eq. 1, but now would assume that the errors, ε_i, for stream sites close to one another are more similar than the errors for stream sites that are far apart. In mathematical terms, the independent error model assumes $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ where $\mathbf{I}$ is the $n \times n$ identity matrix and $\mathbf{0}$ is a vector of length n, and the spatial regression model assumes $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is an $n \times n$ correlation matrix. This model goes by many names in the literature and is a type of general (or generalized) linear model. For our stream sulfate problem, we might adopt a model for the covariance matrix $\boldsymbol{\Sigma}$ (see e.g., Schabenberger and

Manuscript received 2 May 2008; revised 28 May; accepted 30 June 2008. Corresponding Editor: N. T. Hobbs. For reprints of this Forum, see footnote 1, p. 551.

[1] E-mail: jah@lamar.colostate.edu

Gotway 2005) and conclude that observations that are less than 200 km apart are related. This might lead us to think about the biological and physical processes that could lead to this correlation; such research avenues might suggest additional covariates that should be included in our model.

Disadvantages of Ignoring Spatial Correlation

What could go wrong if we use the independent error model (Eq. 1) instead of a spatial regression model when the latter is the appropriate model? Plenty. There is a long history of research that demonstrates the many disadvantages of ignoring spatial correlation. Some of the highlights and relations to ecology are described here.

One key issue is sample size. If an independent-error model is adopted (Eq. 1) but the model errors are not independent, then the "effective sample size" will be smaller than the number of observations collected (Schabenberger and Gotway 2005:32). Effective sample size decreases as the correlation between observations increases. If an independent-error model is adopted but the data are correlated, standard errors can be underestimated. For example, when the independent-error model is used and maximum likelihood estimates of the regression coefficients $\beta_1, \ldots, \beta_4$ in Eq. 1 are obtained, the parameter estimates will be unbiased but the standard errors of these estimates can be too small (Schabenberg and Gotway 2005:324). In ecology this underestimate of uncertainty can be critical: a covariate may be deemed to be important only because an inappropriate model is selected.

In the area of model selection, Hoeting et al. (2006) showed that ignoring spatial correlation when selecting covariates for inclusion in regression models can lead to the exclusion of relevant covariates in the model. Ignoring spatial correlation in model selection can also lead to higher prediction errors for estimation of the response.

The drawbacks described above were based on research for non-Bayesian spatial modeling. In addition to those drawbacks, non-Bayesian spatial models can lead to underestimation of uncertainty. For example, traditional estimation methods for the spatial regression model assume that the covariance matrix $\mathbf{\Sigma}$ is fixed even when the parameters in the model for $\mathbf{\Sigma}$ are estimated. This leads to estimates of standard errors that do not account for the uncertainty in all parameters.

Spatial correlation also plays a factor in sampling design. Cressie et al. (2009) made an important point that hierarchical models allow for direct incorporation of the sampling design in the modeling. The advantages of a sound sampling design cannot be overemphasized. Too many ecological studies involve sites selected for convenience. When the goal of an analysis is to provide a map or some other inference across a sampling area, then additional considerations should be made when designing the study. It has been shown in a number of contexts that a cluster sampling design is appropriate for spatially correlated data (e.g., Zimmerman 2006, Irvine et al. 2007, Ritter and Leecaster 2007). A cluster design includes some observations observed at close distances as well as sampling coverage over the entire sampling area. Xia et al. (2006) proposed methodology that produces an optimal design for spatially correlated data where the optimization and subsequent design depends on the goals of the study. For example, a design which emphasizes accurate estimation of the regression coefficients in Eq. 1 will be different than a design which emphasizes accurate estimation of the degree of spatial correlation. While such informed design is not always possible, even cursory consideration of these ideas should lead to improved sampling designs and thus more accurate models.

Advantages of Bayesian Spatial and Spatiotemporal Hierarchical Models

Numerous examples demonstrate the advantages of accounting for spatial and/or temporal correlation in Bayesian hierarchical models for ecological problems. In the area of species distribution, a series of papers by Gelfand and coauthors (Gelfand et al. 2005, 2006, Latimer et al. 2006) developed a complex hierarchical modeling framework that led to new insights into the spatial distributions of 23 species of a flowering plant family in South Africa. These authors showed that accounting for spatial correlation facilitated the assessment of the factors that impact species distributions, produced accurate maps of species occurrence, and allowed for honest assessment of uncertainty. This work is a particularly good example of the advantages of a Bayesian analysis. The Bayesian paradigm allows for in depth exploration of a virtually unlimited set of results through careful thought and collaboration between ecologists and statisticians. One warning, however, is that the South African species distribution analyses were complex and required many person-hours to produce results. While this work is a terrific example of the possibilities of a Bayesian analysis, it is also an example of the complexities involved in doing such a careful analysis.

In the area of disease ecology, Waller et al. (2007) considered the county-specific incidence of Lyme disease in humans for the northeastern United States. They examined a suite of models ranging from a standard least-squares independent-errors regression model to a hierarchical Bayesian model that accounted for spatial correlation among counties. The inclusion of spatial and temporal components in the model led to new insights into the spread of Lyme disease over space and time and produced maps showing disease trends over space and time. The Bayesian model also allowed for natural incorporation of missing data; the model provided estimates for sites where the predictors were known but the response (Lyme disease counts) was unknown. This work led to new insights into the factors that might contribute to the spread of Lyme disease.

In the area of wildlife disease, Farnsworth et al. (2006) used spatial models to link spatial patterns to the scales over which generating processes operate. They developed a Bayesian hierarchical model to relate scales of deer movement to observed patterns of chronic wasting disease (CWD) in mule deer in Colorado. The Bayesian hierarchical model allowed for investigation of the effects of covariates observed at different scales; covariates for individual deer (e.g., sex and age) and covariates observed across the landscape (e.g., percentage of low-elevation grassland habitat) were included in the model for the probability than an individual deer was infected by CWD. The modeling framework also facilitated a comparison of models for CWD across different scales of deer movement via a model for the unexplained variability in the probability that an individual deer was infected by CWD. The model with the strongest support suggested that unexplained variability has a small-scale component. This led the authors to suggest that future investigations into the spread of chronic wasting disease should focus on processes that operate at smaller, local-contact scales. For example, deer congregate in smaller areas during the winter and disperse across the landscape during the summer; thus CWD may be spread more easily during the winter months.

Hooten and Wikle (2008) considered the spread of an invasive species over time and space. They demonstrated that many insights can be gained via a spatiotemporal model that incorporates a reaction–diffusion component to model the spread of the invasive Eurasian Collared-Dove in the United States. This paper demonstrates another strength of the Bayesian approach as it allows for a natural incorporation of partial differential equation models, long used in mathematics but typically not parameterized to allow for process and data error. The Hooten and Wikle model allows for uncertainty and nonlinearity in the diffusion model (process error) as well as an error term that allows for both observer error and small-scale spatiotemporal variability (data error). The analyses provided a series of maps estimating the extent of the Eurasian Collared-Dove invasion over time for the southeastern United States. The authors concluded that there is remaining variability associated with the rate of species invasion and not attributable to human population. This remaining variability has an estimated spatial range of 1/10 the size of the United States. Such conclusions allow biologists to do a targeted search for other factors that might contribute to the spread of this invasive species.

Challenges and Education

All of the spatial and spatiotemporal modeling cited in the previous section involved close collaboration between ecologists and statisticians. While such collaborations advance the fields of statistics and ecology, statisticians need to develop more approachable interfaces to allow scientists to apply complex Bayesian hierarchical models. As the field of Bayesian hierarchical modeling has matured, software packages such as WinBUGS (*available online*)[2] have made it possible for non-experts to implement the Markov chain Monte Carlo methods required for estimation and inference in many Bayesian hierarchical models. However, much more work needs to be done in this area, particularly for the more complex models that account for spatial and/or temporal correlation.

In addition, a push for more modern statistical education in ecology and other sciences is needed. In the meantime, where can an ecologist learn more about statistical models to account for spatial and/or temporal correlation? An introduction to these issues with an ecological focus is given in the book by Clark (2007). Waller and Gotway (2004) focus on spatial statistics in public health. Books that require a higher-level understanding of statistics but that are still quite accessible include Banerjee et al. (2004), which focuses on Bayesian hierarchical models for spatial data, and Schabenberger and Gotway (2005), which provides a broad overview to statistical methods for spatial data. Both these books include sections on spatiotemporal modeling. Other overviews of spatiotemporal models with an ecological focus include Wikle (1993) and Wikle et al. (1998).

Acknowledgments

This work was supported by National Science Foundation grant DEB-0717367.

Literature Cited

Banerjee, S., B. P. Carlin, and A. E. Gelfand. 2004. Hierarchical modeling and analysis for spatial data. Chapman and Hall/CRC Press, Boca Raton, Florida, USA.

Clark, J. S. 2007. Models for ecological data: an introduction. Princeton University Press, Princeton, New Jersey, USA.

Cressie, N. A. C., C. A. Calder, J. S. Clark, J. M. Ver Hoef, and C. K. Wikle. 2009. Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. Ecological Applications 19:553–570.

Farnsworth, M. L., J. A. Hoeting, N. T. Hobbs, and M. W. Miller. 2006. Linking chronic wasting disease to mule deer movement scales: a hierarchical Bayesian approach. Ecological Applications 16:1026–1036.

Gelfand, A. E., A. M. Schmidt, S. Wu, J. A. Silander, Jr., A. Latimer, and A. G. Rebelo. 2005. Modelling species diversity through species level hierarchical modeling. Journal of the Royal Statistical Society, Series C (Applied Statistics) 54:1–20.

Gelfand, A. E., J. A. Silander, Jr., S. Wu, A. Latimer, P. O. Lewis, A. G. Rebelok, and M. Holder. 2006. Explaining species distribution patterns through hierarchical modeling. Bayesian Analysis 1:41–92.

Hoeting, J. A., R. A. Davis, A. A. Merton, and S. E. Thompson. 2006. Model selection for geostatistical models. Ecological Applications 16:87–98.

Hooten, M., and C. Wikle. 2008. A hierarchical Bayesian nonlinear spatio-temporal model for the spread of invasive species with application to the Eurasian Collared-Dove. Environmental and Ecological Statistics 15:59–70.

Irvine, K, A. I. Gitelman, and J. A. Hoeting. 2007. Spatial designs and properties of spatial correlation: effects on

[2] ⟨http://www.mrc-bsu.cam.ac.uk/bugs⟩

covariance estimation. Journal of Agricultural, Biological and Environmental Statistics 12(4)450–469.

Latimer, A. M., S. Wu, A. E. Gelfand, and J. A. Silander, Jr. 2006. Building statistical models to analyze species distributions. Ecological Applications 16:33–50.

Ritter, K., and M. Leecaster. 2007. Multi-lag cluster designs for estimating the semivariogram for sediments affected by effluent discharges offshore in San Diego. Environmental and Ecological Statistics 14:41–53.

Schabenberger, O., and C. A. Gotway. 2005. Statistical methods for spatial data analysis. Chapman and Hall/CRC, Boca Raton, Florida, USA.

Waller, L. A., B. J. Goodwin, M. L. Wilson, R. S. Ostfeld, S. Marshall, and E. B. Hayes. 2007. Spatio-temporal patterns in county-level incidence and reporting of Lyme disease in the northeastern United States, 1990–2000. Environmental and Ecological Statistics 14:83–100.

Waller, L., and C. Gotway. 2004. Applied spatial statistics for public health data. Wiley, New York, New York, USA.

Wikle, C. K. 2003. Hierarchical Bayesian models for predicting the spread of ecological processes. Ecology 84:1382–1394.

Wikle, C. K., L. M. Berliner, and N. Cressie. 1998. Hierarchical Bayesian space–time models. Environmental and Ecological Statistics 5:117–154.

Xia, G., M. L. Miranda, and A. E. Gelfand. 2006. Approximately optimal spatial design approaches for environmental health data. Environmetrics 17:363–385.

Zimmerman, D. L. 2006. Optimal network design for spatial prediction, covariance parameter estimation and empirical prediction. Environmetrics 17:635–652.

Ecological Applications, 19(3), 2009, pp. 577–581

Hierarchical Bayesian statistics: merging experimental and modeling approaches in ecology

Kiona Ogle[1]

Departments of Botany and Statistics, University of Wyoming, Laramie, Wyoming 82071 USA

Introduction

This is an exciting time in ecological research because modern data analytical methods are allowing us to address new and difficult problems. As noted by Cressie et al. (2009), hierarchical statistical modeling provides a statistically rigorous framework for synthesizing ecological information. For example, hierarchical Bayesian methods offer quantitative tools for explicitly integrating experimental and modeling approaches to address important ecological problems that have eluded ecologists due to limitations imposed by classical approaches. Fruitful interactions between ecologists and statisticians have spawned dialogue and specific examples demonstrating the utility of such modeling approaches in ecology (e.g., Wikle 2003, Clark and Gelfand 2006*a*, *b*, Ogle and Barber 2008), and I applaud Cressie et al. for introducing ecologists to some of the fundamental statistical and probability concepts underlying hierarchical statistical modeling. Readers are also referred to Ogle and Barber (2008) for a more in-depth treatment of the hierarchical modeling framework, fundamental probability results, and examples that illustrate the advantages of this approach in plant physiological and ecosystem ecology.

Hierarchical statistical modeling approaches are promising for addressing complex ecological problems, and I imagine that in the next 10–20 years, these approaches will be commonly employed in ecological data analysis. However, application of these approaches requires appropriate training, but training opportunities in hierarchical modeling methods that integrate experimental and/or observational data with models are lacking (Hobbs and Hilborn 2006, Hobbs et al. 2006, Little 2006). Thus, overview papers such as those by Cressie et al. (2009) and Ogle and Barber (2008) are expected to stimulate interests and motivate new curriculums that deliver training in modern, model-based approaches to data analysis. Cressie et al. discuss several strengths and limitations of hierarchical statistical modeling, but no single topic is treated in great detail. Thus, I expand upon the importance of the "process model" because I see this as a key element of hierarchical statistical modeling that facilitates explicit integration of experiments (data) and ecological theory (models).

Experimental vs. Modeling Approaches

Ecologists generally take one of two approaches to tackling scientific problems: experimental vs. modeling (Herendeen 1988, Grimm 1994). Perhaps the most common approach is to couch a study within an experimental or sampling design framework that "controls" for sources of variability, followed up by standard, frequentist-based hypothesis testing (Cottingham et

Manuscript received 26 March 2008; accepted 16 June 2008. Corresponding Editor: N. T. Hobbs. For reprints of this Forum, see footnote 1, p. 551.

[1] E-mail: kogle@uwyo.edu