Bayesian Analysis (2006) **1**, Number 1, pp. 93–98

Some Perspectives on Modeling Species Distributions (Comment on article by Gelfand et al.)

Jennifer A. Hoeting*

I'd like to congratulate the authors for their important contributions to the study of species distributions. This paper and the authors' other publications that have resulted from this research clearly demonstrate that interdisciplinary research can advance several disciplines simultaneously.

While this paper deals with the scientific problem of species richness and diversity, the authors should also be complimented for the richness and diversity of their statistical results. This paper should be assigned reading for graduate students in statistics, as an example of the range of results that can be examined via a Bayesian analysis. Similarly, students of ecology should read this paper for both the ecological insights and as motivation to take more statistics courses.

1 Modeling individual species level presence–absence

The authors make a number of contributions in the area of modeling individual species level presence–absence. I examine several of these issues below.

One of the important contributions of this work is that the authors model species level presence–absence instead of classifying the sites by some measure of species diversity. As noted by the authors in the introduction, many ecological studies model an index which is a summary over many species. In stream studies, for example, scientists use an index of biotic integrity, which quantifies a stream's ability to support and maintain a natural biological community. Scientists often relate these indices to environmental covariates. However, as noted by Gelfand et al., effects of environmental covariates may be different depending on the species. For example, the effect of minimum July temperature varies across species (see Table 2 and Figure J). Understanding the effect of environmental covariates on individual species is potentially useful and can lead to new insights into species patterns. By examining species individually, Gelfand et al. answer pertinent questions for ecologists.

While the authors can examine individual species with their models, the beauty of their approach is that the results allow for examination of effects over all the species. The authors present a number of insightful measures for this purpose (Section 6). One overall measure considered by the authors is a summary of the effect of covariate l over all species under consideration, or $\sum_{k=1}^{40} \beta_l^{(k)}$ where β_l is the posterior for coefficient

*Department of Statistics, Colorado State University, Fort Collins, CO, http://www.stat.colostate.edu/~jah/

l (Table 4). This quantity seems to be ad-hoc and some weighting of the individual models is probably appropriate, e.g., weighting by the uncertainty in β_l which is clearly shown in Figures H–M, or weighting by the posterior model probability computed for the model for each species.

I would like to emphasize the authors' careful definition of a binary outcome for an areal process (see discussion surrounding equations (1) and (4)). While block averages have a long history in the literature (e.g., Cressie (1993)), this subtlety has been missed by a number of other authors in applications ranging from species distributions to disease mapping. The careful consideration of this issue by Gelfand et al. should encourage others to properly examine this issue.

2 Issues of model assumptions, assessment, and selection

The authors' model for species distribution is the most comprehensive of its kind to date. Their advances lead to suggestions for future studies and areas where additional focus may lead to new insights.

The authors make several simplifying assumptions which are necessary to expedite inference. Two assumptions may benefit from further examination. A basic building block of their model is the assumption that the probability of habitat suitability for species k in grid cell i, $p_i^{(k)}$, is independent of the probability of land transformation, $(1 - U_i)$ in cell i, so

$$P\left(V_i^{(k)} = 1\right) = (1 - U_i)\, p_i^{(k)},$$

where $V_i^{(k)} = 1$ is the event that a randomly selected location in cell i is suitable for species k when cell i has been impacted or changed by human use. I question the validity of this assumption. In the United States, for example, housing developers and foresters must pay close attention to the Endangered Species Act before developing an area. Future authors may wish to explore this issue further.

Another simplifying assumption has to do with relationships between species. The authors claim that the data are collected on such a small scale (1 min by 1 min grid) that "interactions between species are not likely to be of substantial concern." Simple exploratory analyses may be in order to justify such a statement. The authors further examine this issue in their consideration of vicariance in Section 8, and continued work in these areas may be fruitful.

Additional examination of model adequacy may lead to insights about the quality of fit. One could argue that there are so many parameters that the model has to fit the data, but then how can be we sure that there are sufficient data to inform the posterior for all the parameters of interest? Additional studies on the extent of Bayesian learning from the prior to the posterior distribution may be worthwhile.

Another model assessment issue is the spatial scale for the analysis. The authors chose 1 min by 1 min as the scale of the analysis, but inferences may depend on the choice of scale. Exploration of the impact of scale may lead to useful ecological insights.

As the authors point out, model selection is a complex issue for this problem. Here the authors choose models where the model selection statistics are computed over all species. As future work, it might be a worthwhile exercise to investigate model selection at the species level. Just as the parameter estimates associated with environmental covariates vary over species, so might the models themselves.

3 Bayesian Computation

The authors include little discussion of Bayesian computation. Such discussion was probably omitted due to space constraints and due to the fact that many of these issues are examined elsewhere. MCMC computations for the models adopted here require a high degree of expertise and finesse. I briefly touch on several relevant issues below. Several of these topics are discussed in the context of spatial models in the book by Banerjee, Carlin, and Gelfand (2004).

In the previous section I questioned the assumptions of independence between potential species presence and land conversion and also independence in distribution patterns between species. However, incorporating dependence between these quantities may make the already difficult calculations extremely computer intensive. Such complex models will become more tractable as statisticians continue to develop innovative solutions to the problem of MCMC computations. The structured Markov chain Monte Carlo method offers one such innovation by facilitating faster convergence for problems with highly correlated parameters (Cowles 2003; Sargent, Hodges, and Carlin 2000).

For the models considered here, one issue of concern is sensitivity to the prior parameters, particularly for the random effects. This issue has received a great deal of interest in the literature (e.g., Bernardinelli, Clayton, and Monomoli (1995); Carlin and Perez (2000); Haneuse and Wakefield (2004)). Was sensitivity to the prior distributions for the random effects investigated for this problem?

Another issue of continuing concern is how to diagnose convergence of the MCMC runs when there are thousands of parameters to monitor. For the model considered here, there are 2444 independent parameters in equations (5) and (6). The authors do not discuss convergence diagnosis, but clearly it is a challenging problem for this model. Brooks and Roberts (1998) provide one recent overview of work on issues related to MCMC convergence, with some discussion of diagnostic methods for multiple chains.

4 Opportunities and challenges in ecology

Statisticians have much to offer to the field of ecology. There are a wealth of problems that have not been adequately solved and these present interesting challenges for ecologists and statisticians.

For species distribution modeling, incorporating knowledge from the fields of population genetics, evolutionary biology, and biogeography may lead to useful inferences. These types of models should help ecologists gain a more fundamental understanding

of patterns of species distribution. However, many challenges lie ahead, not the least of which are computational and mathematical problems. Gelfand et al. are taking some major steps in this direction in their integration of results from population genetics into models for interspecies dependence.

Mapping species distribution patterns over very large geographical regions (e.g., Africa or the Northern Hemisphere) and estimating trends in species distribution patterns over time are also problems of keen interest for scientists and some policymakers. The sparsity and quality of relevant data in space and time make such analyses challenging. Species distribution survey data are rarely collected in an optimal manner. Some very old natural history museum data arise from studies where only sites with presences were recorded and no records were kept of sites that were searched without finding the species. In addition, it can be difficult to quantify other impacts such as the effects of population growth, land conversion, and climate change. Models to address these issues will continue to present new and interesting challenges for statisticians.

Another modeling challenge involves rare and/or hard-to-find species. A goal may be to produce a map that shows scientists where to find the species. If the species is very rare, creating the models can be problematic as there is so little data to use. These issues are of key concern in the U.S. with the continuing importance of the Endangered Species Act and the use of the courts to ensure its enforcement.

One problem we will probably always face as statisticians is how to communicate results such as the ones in this paper in an understandable manner for other scientists. In my experience many good ecologists lack an understanding of even simple mathematical notation, such as the definition of the transpose symbol. These same people are now being asked to use Bayesian techniques and complicated models. (I don't mean to pick on ecologists; this is a universal problem). The need for mathematical and statistical expertise is not new, but a study like the one presented here demonstrates the need for continued focus on educating mathematically and statistically-savvy scientists. At Colorado State University we are trying to meet these challenges by educating statisticians and mathematicians in the area of ecology and ecologists in the areas of statistics and mathematicians. Our PRogram for Interdisciplinary Mathematics, Ecology, and Statistics (PRIMES), supported by a Integrative Graduate Education and Research Training (IGERT) grant from the National Science Foundation, provides funding for students to gain degrees in their home department while gaining training in other fields and experience in interdisciplinary research (www.primes.colostate.edu). Other efforts, such the short course on Bayesian statistics in ecology that was put on by Gelfand and colleagues at Duke University in summer 2004, are new and creative ways to educate experienced ecologists in these "new" methods. I hope that support for these programs will continue as the need for continued statistical education is clearly great.

Bibliography

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Boca Raton, FL: Chapman & Hall/CRC. 95

Bernardinelli, L., Clayton, D., and Monomoli, C. (1995). "Bayesian estimates of disease maps: How important are priors?" *Statistics in Medicine*, 14:2411–2431. 95

Brooks, S. P. and Roberts, G. O. (1998). "Assessing Convergence of Markov Chain Monte Carlo Algorithms." *Statistics and Computing*, 8:319–335. 95

Carlin, B. P. and Perez, M.-E. (2000). "Robust Bayesian analysis in medical and epidemiological settings." In *Robust Bayesian Analysis*, volume 152 of *D. Rios-Insua and F. Ruggeri, eds., Lecture Notes in Statistics*, 251–372. New York: Springer-Verlag. 95

Cowles, M. K. (2003). "Efficient model-fitting and model-comparison for high-dimensional Bayesian geostatistical models." *Journal of Statistical Planning and Inference*, 111:221–239. 95

Cressie, N. A. C. (1993). *Statistics for Spatial Data*. New York: Wiley. 94

Haneuse, S. and Wakefield, J. (2004). *Ecological Inference: New Methodological Strategies*, chapter Ecological Inference Incoprorating Spatial Dependence, 266–301. Gary King and Ori Rosen and Martin Tanner, eds., Cambridge University Press. 95

Sargent, D. J., Hodges, J. S., and Carlin, B. P. (2000). "Structured Markov chain Monte Carlo." *Journal of Computational and Graphical Statistics*, 9(2):217–234. 95

Acknowledgments

Research supported in part by the U.S. Environmental Protection Agency (EPA) as part of the STAR Research Assistance Agreement CR-829095 awarded to Colorado State University. The views expressed here are solely those of author. EPA does not endorse any products or commercial services mentioned here.

Bayesian Analysis (2006) **1**, Number 1, pp. 99–102

Comment on article by Gelfand et al.

Jay M. Ver Hoef*

I have enjoyed reading the paper by Gelfand et al. (2005). My congratulations go to the authors, as they have given us an important advance in the science of modeling species diversity. First, I would like to emphasize the importance of this topic. I fully agree with the authors that species diversity has been a central concept in ecology for many years, yet the mechanisms that determine species diversity are still enigmatic. How then has this paper helped us?

One of the first problems in assessing species diversity is to know where a species occurs. While this may seem simple, it is actually very difficult. The authors have a very fine data set that was systematically sampled in a very interesting, diverse part of the world, where high species diversity is compacted into a relatively small space. One of the questions that I want to ask is, "Can the methods of Gelfand et al. (2005) be used more generally?" That is, can I use them in Alaska? Alaska is a rather large state, but if we consider plants, being far to the north, it is not really very diverse. We know of only about 1600 different plant species in Alaska. Rhode Island has more plant species (2600). The methods of Gelfand et al. (2005) are fairly complex, but in principle it seems that they could be adapted for hundreds (perhaps thousands) of different plant species as computational power increases. However, for a more general application, there are problems with species presence data that do not occur for Gelfand et al. (2005). Sampling has not occurred uniformly over my state, or any large geographic area that I know of. For example, I'm pretty sure that if we added a covariate such as distance to the nearest university, there would be a highly significant, negative regression coefficient when modeling species presence or diversity. The reason is clear. For years, botany professors have been sending out legions of graduate students and classes to collect plants, and they stay relatively close to home. Thus, not all zeros are created equal. This is known as ascertainment bias in the epidemiology literature. Gelfand et al. (2005) have done an outstanding job in distinguishing other factors that do create zeros, such as transformed landscapes. This is an important step, but it is information that is relatively easy to gather as compared to effort. Eventually, it will be important to solve the effect of effort (ascertainment bias).

Now, what about prior information? Gelfand et al. (2005) use a hierarchical model with vague priors. This makes sense, given the complexity of the model. Eliciting priors from most plant collectors that I know would be very difficult. It would be hard for them to make sense of priors on parameters in a model with the complications of the potential and transformed surfaces, hidden random effects, etc. Still, these same plant collectors have a wealth of prior knowledge; they have spent years crawling through the bushes. Early in my career I collected plants as my job, and I lived by the maps drawn in Hulten's (1968) *Flora of Alaska*. It was a big deal to extend any of the species ranges drawn in his book. Plant collectors, such as Hulten, simply used their experience and

*National Marine Mammal Laboratory, Seattle, WA, jay.verhoef@noaa.gov

knowledge of terrain, climate and the known collection locations for a species to draw a line on a map that formed the species range. How can we tap such information? One interesting approach has been taken by Lele and Das (2000), who did not adopt a Bayesian formulation. Their thesis is that we should elicit predictions, not priors, on parameters. I think this is the right idea, and it would be interesting to incorporate a Bayesian approach that uses elicited predictions into the models developed by Gelfand et al. (2005), and indeed many others.

The model that Gelfand et al. (2005) propose is very interesting; it is a major improvement on many other approaches. As noted earlier, it is fairly complex compared to almost all other approaches so far. Nevertheless, there is one part of the model that is perhaps too simple. In eq. (5), Gelfand et al. (2005) give us

$$\log\left(\frac{p_i^{(k)}}{1-p_i^{(k)}}\right) = \boldsymbol{w_i'\beta_k} + \psi_k + \rho_i.$$

This model allows each species to have its own intercept ψ_k and covariate response vector β_k, but all species have a common spatial pattern ρ_i in the "residuals" – or that part of the model not explained by the fixed effects. A model such as

$$\log\left(\frac{p_i^{(k)}}{1-p_i^{(k)}}\right) = \boldsymbol{w_i'\beta_k} + \psi_k + \rho_i^{(k)}$$

has too many parameters, allowing a separate spatial pattern (in the residuals) for each species. Undoubtedly, some species are responding to similar spatial effects. As the authors point out, this residual spatial random effect accounts for (at least in part) unmeasured spatially-patterned covariates. Some species will respond in a similar manner to a particular unmeasured covariate, while other species will respond in a similar way to another unmeasured covariate. A more flexible approach that does not have too many more parameters would be to allow for just a few spatial patterns, and then assume that each species' residuals are a linear combination of those spatial random effects:

$$\log\left(\frac{p_i^{(k)}}{1-p_i^{(k)}}\right) = \boldsymbol{w_i'\beta_k} + \psi_k + \sum_{m=1}^{M} \eta_k^{(m)} \rho_i^{(m)},$$

where M is, say, 1 to 5. Bayes factors, DIC, or reversible jump MCMC methods could be used to choose M.

None of this detracts from the fundamental contributions that Gelfand et al. (2005) have given us. I hope that both statisticians and ecologists take notice, and that they use and build upon the models and ideas that these authors have developed. The synergy of collaboration among statisticians and ecologists is apparent from this article.

Bibliography

Gelfand, A., Silander, J., Shanshan, W., Latimer, A., Lewis, P., Rebelo, A., and Holder, M. (2006). "Explaining species distribution patterns through hierarchical modeling."

Bayesian Analysis, (in press).

Hulten, E. (1968). *Flora of Alaska and Neighboring Territories*, 1008. Stanford University Press.

Lele, S. and Das, A. (2000). "Elicited data and incorporation of expert opinion for statistical inference in spatial studies." *Mathematical Geology*, 465–487.

Bayesian Analysis (2006) **1**, Number 1, pp. 103–104

Rejoinder

Alan E. Gelfand, John A. Silander Jr., Shanshan Wu, Andrew Latimer,
Paul O. Lewis, Anthony G. Rebelo and Mark Holder

We very much appreciate the positive comments of both Jay Ver Hoef and Jennifer Hoeting. We were particularly delighted by their appreciation of our "synergy between statisticians and ecologists" and our demonstration "that interdisciplinary work can advance several disciplines simultaneously." We briefly address the key points they have brought up.

Regarding Jennifer's criticism of the "independence" assumption in expression (2) of the paper, we agree that this is surely not true though it may be roughly true. However, the "correlation" calculation in (2) is a bit more complicated than it initially appears in that the calculation is with respect to a uniform distribution over say, the locations in unit i. In fact, the randomness arises from the fact that the objects being integrated are random functions, rather than from randomness in the choice of locations. So, in fact, the integral in (2) is a stochastic integral and the assumption demands that the resulting random variable factor almost surely into the product of the random variables $1 - U_i$ and $p_i^{(k)}$. In the absence of this assumption, we would have little choice but to model $P(V_i^{(k)} = 1)$ analogously to (5) in the paper and, as a consequence, we would sacrifice the ability to consider both potential and transformed species distributions.

Jay raises the important issue of sampling bias in the data collection. We recognize that this occurs in most presence-absence species sampling and, indeed, it does for our data as well. The expert botanist on our team (Rebelo) was *sure* that for large areas within the CFR, no protea would be found and thus that there was no need to sample in these areas. We did not take this information as "data," e.g., in the form of null sites; rather, we counted upon the second stage spatial modelling to provide smoothing for the random effects associated with unsampled grid cells. In this regard, Jennifer also comments upon sampling concerns, particularly with, say if one were working with museum data where there are no nulls. We note that there seems to be a component of the ecology community that is comfortable with developing species ranges in this setting (see, for instance, Engler et al., 2004). We are troubled by such inference for handling presence-only data (as most statisticians would be) and, in a manuscript in preparation, will attempt to illuminate more clearly the flaws in this work along with possible remedies.

Jay has noted the limitations in expression (5) of the additive form in species random effects and spatial random effects. We completely agree and have looked at for example, an additional multiplicative term of the form $\alpha\psi_k\rho_i$ as well as other possibilities. However, in forthcoming work (Latimer, et al., 2005) we have focused on species level spatial random effects, $\rho_i^{(k)}$. Assuming these to be independent across species enables us to fit our model one species at a time. This allows simple parallelization of the computation and is permitting us to make our way through range prediction for the more than 300 protea species in the CFR. This approach also fits in nicely with Jennifer's suggestion

that we investigate model selection at the species level. Indeed, in the Latimer et al paper we carry out a modest version of this. We note that in other new work we consider the issue of spatial scale, working with 1 min $\times$ 1 min grids as well as 4 min $\times$ 4 min and 16 min $\times$ 16 min grids. The differences in predicted species range are expected and noteworthy.

Finally, two issues were raised which we have not investigated. First, Jennifer suggests that an uncertainty weighted sum of the β's would be more appropriate to work with. The challenge in this case would be to specify these weights so that the resulting sum is still a parametric function in order that we can examine its posterior. In our current formulation the $\beta_l^{(k)}$ are i.i.d. for each l making it unclear where the weights would come from? Second, Jay suggests that informative prior specification could be elicited in the form of predictions. This is an attractive idea and has some history in the Bayesian community. See, for instance the work of Ibrahim and colleagues, e.g., Ibrahim (1997).

In summary, we thank Jay and Jennifer for their thoughtful remarks and expect that they, along with us, will continue to collaboratively address challenging model problems that arise when studying complex ecological processes. There is much opportunity for statisticians to contribute in such enterprise.

Bibliography

Engler, R., Guisan, A., and Rechsteiner, L. (2004). "An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data." *Journal of Applied Ecology*, 41:263–274.

Ibrahim, J. (1997). "On properties of predictive priors in linear models." *The American Statistician*, 51:333–337.

Latimer, A., Wu, S., Gelfand, A., and Silander, J. J. (2005). "Building statistical models to analyze species distributions." *Ecological Modelling*. (forthcoming).