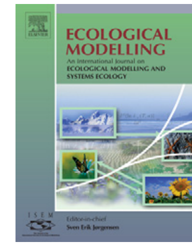
ECOLOGICAL MODELLING 199 (2006) 129–131

available at www.sciencedirect.com
ScienceDirect
journal homepage: www.elsevier.com/locate/ecolmodel



Editorial

Further advances in predicting species distributions

In 2001, a workshop focused on the use of generalized linear models (GLM: McCullagh and Nelder, 1989) and generalized additive models (GAM: Hastie and Tibshirani, 1986, 1990) for predicting species distributions was held in Riederalp, Switzerland. This topic led to the publication of special issues in Ecological Modelling (Guisan et al., 2002) and Biodiversity and Conservation (Lehmann et al., 2002). A second workshop, held in 2004, also in Riederalp, brought together a larger group of modeling specialists from the fields of ecology and statistics to discuss advances in predictive species distribution models since the first gathering. Three collections of papers resulted from this more recent workshop, one in the Journal of Applied Ecology (Guissan et al., in press), one in the Journal of Biogeography, (Araújo and Guisan, in press), and this special issue. The papers contained here offer a diverse look into current modelling issues for spatial prediction of both plant and animal species distributions at a variety of scales. Topics of these papers span all stages of the species distribution modelling process, including: choosing between purposive and probabilistic sampling schemes; handling problematic characteristics of ecological data such as disproportionate numbers of zero values, small sample sizes, and presence only data; comparing predictive modeling tools; and making best use of ecological theory in modelling practices. All of these topics are central to the proper modeling of species distributions, and this series of papers provides new insights to each topic, presenting a broad-based context in which to evaluate modeling tools.

We begin our series of papers with an exploration of some sampling issues and their possible effects on species distribution models. Much distribution modelling is conducted using samples collected in a purposive manner, where species of interest are targeted, rather than relying on a probabilistic design. The potential consequences of these purposive surveys on our ability to correctly model species distributions are rarely examined. The paper by Thomas Edwards and colleagues explores the effects of probabilistic versus purposive sampling on several measures of accuracy of classification models used for predicting the presence of four lichen species in the Pacific Northwest, USA. These probabilistic and non-probabilistic based models were compared using an independent evaluation data set. Although probabilistic and purposive surveys had somewhat similar accuracies when tested on training data, those based on probabilistic surveys had much greater accuracy when applied to unsampled regions.

Regardless of the underlying sample design, the majority of classification tools used in species distribution modelling require both presence and absence data. However, many extant data sets, especially records of museums and herbariums, contain only presence data. Modelling approaches exist for dealing with these presence-only data, but it is unclear if commonly used measures of accuracy, such as kappa (Cohen, 1960), MaxKappa (Guisan et al., 1998), and area under receiver operating characteristic curves (AUC; DeLong et al., 1988), are applicable. Using a data set of 114 plants species, Alexandre Hirzel and his co-authors compared accuracies obtained from these estimators to three metrics used for assessing the performance of presence-only models. Overall, the presence-only assessment metrics are highly correlated with measures commonly applied to presence–absence data.

While much work on spatial prediction to date has focused on presence–absence or presence-only data, Joanne Potts and Jane Elith point out that conservation management often requires abundance measures. Using data on a rare Australian plant *Leionema ralstonii* (F. Muell.), they investigate how alternative models perform when predicting abundance rather than presence/absence. Typical of much biological abundance data, their data were zero-inflated (having an excessive number of zero values) and over-dispersed (with variance exceeding the mean). The authors compare the familiar Poisson and negative binomial models with variants which cope better with large numbers of zeros. The best performer was the hurdle model (Cragg, 1971), which couples a presence–absence model with a count model. The negative binomial performed the worst. They recommend that researchers consider the hurdle model under the conditions of zero-inflation and over-dispersion because it performed well while being relatively easy to interpret and implement.

Austin (2002) has previously stressed the underlying importance of ecological theory when specifying predictive models for distributions. What most of us try to do in applying statistics to distributional data is to capture the response of a species to an environmental gradient. The totality of these responses over all gradients characterizes the ecological niche. A key question that is addressed by Christophe Coudun and Jean-Claude Gegout is how the sampling regime affects our ability to recover species response curves. In a simulation study using both artificial and real data sets, they show that response curves are poorly characterized when a species is too rare or when its theoretical optimum lies near the end of the sampled gradient. As a rule of thumb, they suggest that at least 50 occurrences are needed to characterize a species' response curve when using logistic regression.

The paper by Gretchen Moisen and her colleagues examines the performance of classification and regression trees relative to GAMs and stochastic gradient boosting (SGB; Friedman, 2001, 2002) when modelling broad-scale forest attributes. While variants of classification and regression trees are frequently used in these kinds of modelling exercises, especially when remotely-sensed information serves as the basis of the predictors, the results are frequently uninterpretable. Also, the most commonly used software packages operate using proprietary computational algorthms which are not transparent to the user. Modelling presence and prevalence of 13 tree species in the Intermountain West of the United States, they found that both the GAM and SGB models performed better than tree-based methods, as measured with AUC. SGB was the most stable of the techniques compared for predicting continuous species prevalence.

In another technique comparison paper, John Leathwick and co-authors explore the utility of multivariate adaptive regression splines (MARS; Friedman, 1991) as an alternative technique to GAMs, comparing GAM-based predictions against those of MARS for predicting the distributions of 15 fish species in New Zealand rivers and streams. Both GLMs and GAMs have been widely applied to species distribution modelling, even with their known limitations. GLMs, for example, are limited in their ability to fit the complex, non-linear relationships often occurring between species and environmental predictors (e.g., Austin et al., 1990), while the complexity of GAMs can make their use difficult, especially if the goal is to link model predictions to a GIS. The authors, results found few differences in model performance for individual species, as assessed using measures of deviance and AUC, but did note clear advantages to using MARS for analysis and spatial prediction. In addition, the authors discuss the predictive and computational advantages to taking a community approach to modelling species presence, where multiple species are modeled simultaeously in MARS, particularly in instances where rare species are of interest.

Last, the capability of statistical tools to accurately model species distributions is difficult to evaluate, in part because field data may not truly represent "truth". One possible solution is to use artificial data sets for evaluation purposes. As Mike Austin and colleagues argue, however, artificial data sets themselves must be based on sound ecological principles, not just derived from known statistical distributions, as if often the case. Using two commonly accepted theoretical bases in vegetation science, the Swan–Ter Braak (Swan, 1970; Ter Braak, 1985; Jongman et al., 1987), and the Ellenberg–Minchin (Mueller-Dombois and Ellenberg, 1974; Austin, 1976; Minchin, 1987) models, they created an artificial data set to then evaluate the performance of a GLM and a GAM. The data set contained both direct and indirect predictors of several hypothetical plant species. Both the GLM and the GAM performed well with Gaussian responses and the direct predictors, while the GAM outperformed the GLM with indirect predictors. The authors emphasize the importance of ecological knowledge and statistical skills of the analysts in modelling species presence.

The set of papers within this special issue provide in-depth explorations of several aspects of species distribution models, including effects of sampling regimes, comparisons of techniques, and the utility of artificial data sets for evaluating model performance. Together, they shed new light on the use of statistical tools for modeling and understanding species distributions in ecology.

Acknowledgements

The papers presented here had their origins in the second workshop on generalized regression analyses and spatial predictions which was attended by 30 scientists from 11 countries. The guest editors are grateful to the hosts of the workshop, Antoine Guisan, Anthony Lehmann, Jake Overton, Simon Ferrier, and Richard Aspinall, for their work in fostering a wonderful environment for scientific discussion on species modeling, and for encouraging the publication of these papers. We also acknowledge the following for their support of the workshop: the Swiss Academy of Science, as well as the Faculty of Geosciences and Environment, the Foundation of the 450th, the Herbette Foundation, and the Department of Ecology and Evolution, all at the University of Lausanne. Finally, thanks go to the authors for their contributions, referees for their thoughtful reviews, and to S.E. Jørgensen for his support of this special issue.

REFERENCES

Araújo, M.B., Guisan, A., in press. Five (or so) challenges for species distribution modelling. J. Biogeogr.

Austin, M.P., 1976. Non-linear species response models in ordination. Vegetation 33, 33–41.

Austin, M.P., 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. Ecol. Model. 157, 101–118.

Austin, M.P., Nicholls, A.O., Margules, C.R., 1990. Measurement of the realized qualitative niche: environmental niches of five eucalypt species. Ecol. Monogr. 60, 161–177.

Cohen, J., 1960. A coefficient of agreement of nominal scales. Educ. Psychol. Measur. 20, 37–46.

Cragg, J.G., 1971. Some statistical models for limited dependent variables with application to the demand for durable goods. Econometrica 39, 829–844.

DeLong, E.R., Delong, D.M., Clarke-Pearson, D.L., 1988. Comparing areas under two or more correlated receiver operating characteristic curves: a non-parametric approach. Biometrics 44, 837–845.

Friedman, J.H., 1991. Multivariate adaptive regression splines (with discussion). Ann. Stat. 19, 1–141.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 29, 1189–1232.

Friedman, J.H., 2002. Stochastic gradient boosting. Comput. Stat. Data Ann. 38, 367–378.
Guisan, A., Edwards Jr., T.C., Hastie, T.J., 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecol. Model. 157, 89–100.
Guissan, A., Lehmann, A., Ferrier, S., Austin, M., Overton, J., Aspinall, R., Hastie, T.J., in press. Making better biogeographic predictions of species distributions. J. Appl. Ecol.
Guisan, A., Theurillat, J.P., Kienast, F., 1998. Predicting the potential distribution of plant species in an Alpine environment. J. Veg. Sci. 9, 65–74.
Hastie, T.J., Tibshirani, R.J., 1990. Generalized Additive Models. Chapman & Hall, New York.
Hastie, T.J., Tibshirani, R.J., 1986. Generalized additive models. Stat. Sci. 1, 297–318.
Jongman, R.G.H., ter Braak, C.J.F., van Tongeren, O.F., 1987. Data analysis in community and landscape ecology. Pudoc. Wageningen.
Lehmann, A., Overton, J.M., Austin, M.P., 2002. Regression models for spatial prediction: their role for biodiversity and conservation. Biodiv. Conserv. 11, 2085–2092.
McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models. Chapman & Hall, New York.
Minchin, P.R., 1987. Simulation of multidimensional community patterns: towards a comprehensive model. Vegetation 71, 145–156.
Mueller-Dombois, D., Ellenberg, H., 1974. Aims and Methods of Vegetation Ecology. Wiley, New York.
Swan, J.M.A., 1970. An examination of some ordination problems by use of simulated vegetation data. Ecology 51, 89–102.
Ter Braak, C.J.F., 1985. Correspondence analysis of incidence and abundance data: properties in terms of a unimodal response model. Biometrics 41, 859–873.

Gretchen G. Moisen*
USDA Forest Service, Rocky Mountain Research Station, 507 25th Street, Ogden, UT 84401, USA

Thomas C. Edwards Jr.
USGS Utah Cooperative Fish and Wildlife Research Unit, College of Natural Resources, Utah State University, Logan, UT 84322-5290, USA

Patrick E. Osborne
Centre for Environmental Sciences, School of Civil Engineering and the Environment, University of Southampton, Highfield, Southampton SO17 1BJ, United Kingdom

* *Corresponding author*. Tel.: +1 801 625 5384; fax: +1 801 625 5723.
E-mail address: gmoisen@fs.fed.us (G.G. Moisen)

0304-3800/$ – see front matter
© 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2006.05.015

Published on line 21 July 2006