# Uncovering the Secrets of Rare Species: Can Community Level Modelling Help?

Brenton S. Chatfield[1, 2]

[1] School of Earth and Geographical Sciences
[2] School of Plant Biology
University of Western Australia
35 Stirling Highway, Crawley, Western Australia, Australia 6009
Telephone: (+61) 8 6488 4235
Fax: (+61) 8 6488 1037
Email: chatfb01@student.uwa.edu.au

## 1. Introduction

Spatial predictive modelling is now regularly used to explore patterns of species distribution and to investigate species-environment relationships in both terrestrial and aquatic environments (see reviews by (Franklin 1995; Guisan and Zimmermann 2000; Scott *et al*. 2002; Elith *et al*. 2006).

While most applications have developed models for single species, a recent review of community level modelling (Ferrier and Guisan 2006) has suggested that this approach deserves further consideration, as it has advantages over single species models in certain situations. One such community level modelling approach is multivariate adaptive regression splines (MARS) (Friedman 1991) which use piecewise linear basis functions to describe complex non-linear responses between a species or groups of species (multiresponse models) and environmental predictors. The ability to identify a common set of environmental predictors for community composition (Leathwick *et al*. 2006; Elith and Leathwick 2007) with a similar predictive ability to generalised additive models (GAMs) (Leathwick *et al*. 2006) suggests MARS can offer an approach for constructing distribution models for rare species or species with low occurrence data which could not be modelled individually due to statistical constraints.

Given the expense and logistical constraints involved in collecting data from the marine environment, the amount of data available to develop predictive models is often very sparse. Consequently, models have only been able to be used to investigate and predict the distribution of the more prevalent species when it is often the rare or low occurrence species which are of most interest, especially when determining conservation strategies.

Thus, the ability of the MARS approach to describe and predict the distribution of demersal fish species in the Recherche Archipelago, Western Australia was investigated, and specifically, how well the community level approach was for investigating and predicting distributions of species with very low occurrence data.

## 2. Methods

### 2.1 Study Site and Data Collection

The study focused around the Remark group of islands in the Recherche Archipelago, which is located on the south coast of Western Australia near the coastal town of Esperance. The sampling plan was designed to ensure that samples were taken across the main environmental gradients (substrate type and water depth), and it consisted of 3 components. A systematic component provided samples placed at 800 m intervals across the entire area. A stratified random component was used to determine the sampling locations around the islands where the substrate class (seagrass, reef, rhodolith and sand) and exposure levels were known. Finally, a small cluster of samples were taken to investigate rapid changes in depth seen towards the west of the study area.

Presence and ‘pseudo-absences’ (Elith and Leathwick 2007) were collected using baited remote underwater video systems (BRUVS) which recorded a 1 hour ‘snapshot’ of the fish species present at each of the 231 sites. For each site, all species presence and absence data was recorded from the footage. The predictor variables chosen for modelling had all been previously identified as influencing fish distribution and included substrate type, water depth, maximum wave height, maximum shear stress, the type and density of macroalgae, and the presence or absence of filter feeding organisms. The substrate type, macroalgae data, and the presence or absence of filter feeders was determined from the video footage. Water depth was recorded at the time of sampling and the maximum wave height and maximum shear stress for each site was derived from oceanographic models that were available for the study area (Kendrick *et al*. 2006)

Of the 81 species identified from the footage, only those species whose known depth range was within the depth range sampled were considered for modelling.

### 2.2 Model Fitting

MARS models for both single species and multiple species (multiresponse models) were constructed using the free statistical software, R, version 2.4.1 (R Development Core Team 2006) with the mda library and by modifying the custom code described and available in Elith and Leathwick (2007).

Model performance was evaluated by calculating the area under the receiver operating characteristic curve, AUC (Fielding and Bell 1997) and deviance (Elith and Leathwick 2007) from cross validated folds of data. Where AUC quantifies the ability of the models to discriminate between presences and absences, deviance expresses the magnitude of the deviations of the observed and fitted values (Elith and Leathwick 2007). The relative contribution of each term to the overall deviance was calculated and a comparison of the terms retained in the single species models was made with the terms retained in the multiresponse model.

In order to determine the ecological relevance of the models and to assess the influence that multiple gradients have on the probability of occurrence, the response curves and the fitted probabilities of occurrences were plotted for each species.

## 3. Results

Preliminary results indicate that both the single species and multiresponse MARS models were able to produce accurate predictions of the distributions, with mean AUC values for both approaches being greater than 0.85. There was no obvious decline in predictive performance of the multiresponse model for predicting the low occurrence species compared to more abundant species.

While the single species and multiresponse models had similar prediction accuracy, they did vary in which terms were used to make those predictions. On average, the multiresponse models retained more terms than the single species models and the relative importance of common terms varied between the modelling approaches.

## 4. Discussion

The results indicate that multiresponse models developed using MARS are certainly able to accurately reflect and predict the distribution of both high and low occurring species. There was minimal difference in the predictive ability between multiresponse and single species MARS models based on AUC and deviance values.

What was different were the variables being used by each modelling approach to make the predictions for the same species. It is suggested that while multiresponse models are able to determine the dominant environmental drivers of distribution from groups of species, it should not necessarily be done instead of, but as a supplement to single species modelling. Single species modelling still allows a more thorough investigation of the environmental factors influencing individual species and is valuable to improve our understanding of a species' ecology. Obviously, for species with low occurrence, multiresponse MARS models are the only option and their accuracy to predict distribution of these species will improve the quality of information available for developing conservation and resource management strategies.

This research has shown that community level modelling is definitely advantageous if predicting distributions of rare or low occurrence species is required. What will be interesting to determine is how well these models are able to predict distributions of all species using independent data compared to using cross validation predictions.

## 6. Acknowledgements

This work was conducted with funding from the CRC for Coastal Zone and Estuary and Waterway Management while the author was receiving an Australian Postgraduate Award Scholarship. The author would also like to thank John Leathwick (National Institute for Water and Atmospheric Research, Hamilton, New Zealand) for his assistance with developing the models in R.

## 7. References

Elith J, Graham CH, Anderson RP, Dudik M, Ferrier S, Guisan A, Hijmans RJ, Huettmann F, Leathwick JR, Lehmann A, Li J, Lohmann LG, Loiselle BA, Manion G, Moritz C, Nakamura M, Nakazawa Y, Overton JM, Peterson AT, Phillips SJ, Richardson KS, Scachetti-Pereira R, Schapire RE, Soberón J, Williams S, Wisz MS, Zimmermann NE (2006) Novel methods improve prediction of species' distributions from occurrence data. Ecography 29: 129-151

Elith J, Leathwick J (2007) Predicting species distributions from museum and herbarium records using multiresponse models fitted with multivariate adaptive regression splines. Diversity and Distributions 13: 265-275

Ferrier S, Guisan A (2006) Spatial modelling of biodiversity at the community level. Journal of Applied Ecology 43: 393-404

Fielding AH, Bell JF (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. Environmental Conservation 24: 38-49

Franklin J (1995) Predictive vegetation mapping: geographic modelling of biospatial patterns in relation to environmental gradients. Progress in Physical Geography 19: 474-499

Friedman JH (1991) Multivariate adaptive regression splines. The Annals of Statistics 19: 1-67

Guisan A, Zimmermann NE (2000) Predictive habitat distribution models in ecology. Ecological Modelling 135: 147-186

Kendrick G, Harvey E, McDonald J, Pattiaratchi C, Cappo M, Fromont J, Shortis M, Grove S, Bickers AN, Baxter KJ, Goldberg N, Kletczkowski M, Butler J (2006) Characterising the fish habitats of the Recherche Archipelago. Fisheries Research and Development Corporation Report. Project No. 2001/060

Leathwick JR, Elith J, Hastie T (2006) Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. Ecological Modelling 199: 188-196

Scott MJ, Heglund PJ, Morrison ML, Haufler JB, Raphael MG, Wall WA, Samson FB (2002) Predicting Species Occurrence: issues of accuracy and scale. Island Press, Washington, pp 868