Hierarchical species distribution modelling across high dimensional nested spatial scales

Craig Wilkie[1], Jafet Belmont[1], Claire Miller[1], Marian Scott[1], Tom August[2], Philip Taylor[3]

[1] University of Glasgow, Glasgow, UK
[2] UK Centre for Ecology and Hydrology, Wallingford, UK
[3] UK Centre for Ecology and Hydrology, Edinburgh, UK

E-mail for correspondence: `craig.wilkie@glasgow.ac.uk`

Abstract: We propose a two-stage modelling approach to evaluate how a large suite of environmental metrics available over nested spatial scales shape species distributions. We focus on dragonfly communities, where the data consist of partially observed presence records, making identifying the ecological processes driving the true species distribution/occupancy patterns difficult.

Keywords: Detectability; Dragonflies; Occupancy; High Dimensionality

1 Introduction

Understanding how species distributions are affected by environmental changes is of major interest in many ecological studies. However, describing such processes is no easy task due to the sources of uncertainty that occur at different spatial and temporal scales and that are induced by imperfect detectability.
We propose a two-stage statistical modelling framework for analysing how environmental metrics describing freshwater connectivity interact with land-use change to affect species distributions, while accounting for imperfect detectability of the species. Specifically, we look at UK dragonfly species richness, since their species presence records are only partially observed due to imperfect detection.

2 Data sets

Data were provided by Hydroscape (web: hydroscapeblog.wordpress.com), a project investigating how anthropogenic stressors and connectivity interact to influence biodiversity and ecosystem function in UK freshwaters. Dragonfly occupancy records (for over 4000 $\times$ 1km grid cells, matched to lakes) from 2000 to 2016 were taken for 41 non-invasive species from the National Biodiversity Network, Biological Records Centre and British Dragonfly Society repositories. Species-specific covariates that may affect species detection probability were taken from Powney et al. (2014). Anthropogenic stressors (% agricultural and urban land use) and connectivity metrics (e.g. perimeter, number of lakes, river length) were calculated on 7 spatial scales surrounding each lake to capture the impact of different types of connectivity on freshwater ecosystems (Taylor et al., in prep.).

3 Statistical Methods

Species richness (the total number of species occupying a grid cell) can be underestimated when the probability of detecting the different species is less than 1. The species occupancy is potentially affected by many environmental variables over nested spatial scales. We take a 2-stage approach, estimating detectability in stage 1 and identifying and modelling the effects of the covariates on the adjusted species richness in stage 2.

3.1 Stage 1: estimating occupancy, accounting for detectability

First, we analysed the observed occupancy of dragonfly species in each grid cell by fitting a species-specific multispecies occupancy model (eqn. 1) using the species-specific covariates from equation 2 (Kéry and Royle, 2008).

$$\begin{aligned} z_{ij} &\sim \text{Bernoulli}(\psi_i) && \text{State process} \\ \sum_{K_j} y_{ij\cdot} | z_{ij} &\sim \text{Binomial}(K_j, p_i z_{ij}) && \text{Aggregated observation process} \\ \text{logit}(\psi_i) \sim \text{N}(\mu_{\psi_i}, \sigma^2_{\psi_i}); \text{logit}(p_i) &\sim \text{N}(\mu_{p_i}, \sigma^2_{p_i}) && \text{Species heterogeneity model} \end{aligned} \tag{1}$$

$$\mu_{p_i} = \alpha_0 + \sum_{m=1}^{M} \alpha_m (m\text{th species-specific parameter})_i \tag{2}$$

where $y_{ij\cdot}$ is the number of times species i was detected in grid cell j across K visits, p_i is the detection probability for the ith species, z_{ij} is the latent variable for true species occupancy and ψ_i is the occupancy probability. Grid cell-level species richness is computed as a derived quantity of the predicted occupancy, as $S_j = \sum_i z_{ij}$. Noninformative priors were specified to run the Gibbs sampler in R through JAGS.

3.2 Stage 2: understanding the effects of the covariates on species richness

Second, we evaluated the effect of grid cell-level covariates on species richness, using two sub-steps:
(a) ***Random forests*** (Strobl et al., 2009; accounting for high correlations and differing scales) to identify the explanatory variables that are "important" to the response, with a reduced set selected using prediction MSE.
(b) The reduced set of potential explanatory variables are considered in a ***generalised additive model (GAM)***, allowing for smooth, nonlinear relationships, with interactions modelled using tensor products. Stage 1 uncertainties are included through inverse-variance weighting, via the gamm function in the mgcv package in R.

4 Results

The dragonfly occupancy and detection probabilities varied widely within the community, as shown in Figure 1. The estimated detection probability is below 50% for most species, showing the importance of accounting for uncertainty in observed species occupancy.

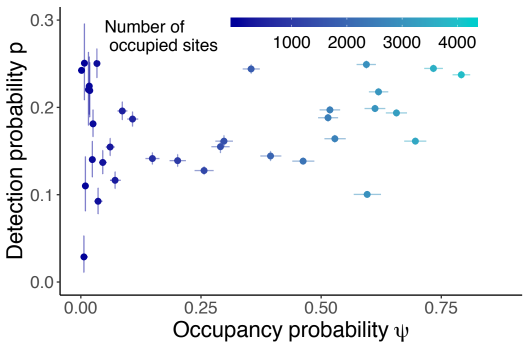

FIGURE 1. Estimated species occupancy and detection probabilites with number of sites each of the species is estimated to be present.

For stage 2, we provide an example for moderate alkalinity, deep lakes. Random forests identified the 6 most relevant potential explanatory variables to dragonfly richness from a dataset of 144 potential explanatory variables. Of these, two variables represented the same parameter (% agriculture) at two spatial scales. Only the most important of these two variables was retained. The 5 remaining potential explanatory variables were considered in a quasipoisson-response GAM, incorporating the inverse of the stage 1 prediction variance as weights. Figure 2 shows the smooths for the resulting model. Log(catchment mean rainfall) has a positive coefficient (1.18). Square root of 500m buffer Strahler 2 length per ha is a connectivity variable with a generally positive effect, but logit(% agriculture in catchment)

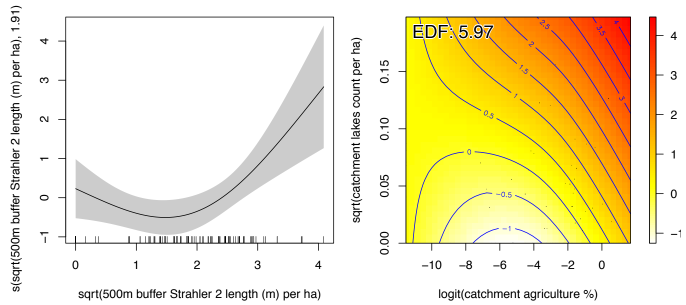

FIGURE 2. Fitted smooths for moderate alkalinity, deep lakes.

interacts with a connectivity variable (square root of catchment lake count per ha), suggesting that the effects of connectivity on dragonfly species richness vary with increasing stress caused by nearby agriculture. The model explains approximately 28% of variance in estimated species richness, appearing to be a moderately good fit to the data.

5 Discussion and conclusions

This two-stage approach presents a computationally efficient method for dimension reduction of the nested spatial covariate space to model species richness in the presence of imperfect detection.

Acknowledgments: JB's work is financially supported by the Mexican Council of Science and Technology (CONACyT) under scholarship number 494334. All others worked as part of NERC Hydroscape (NE/N005740/1). Steve Brooks (NHM) provided additional advice on dragonfly ecology.

References

Kéry, M. and Royle, J. (2008). Hierarchical Bayes estimation of species richness and occupancy in spatially replicated surveys. *Journal of Applied Ecology*, **45**, 589 – 598.

Powney, G., Brooks, S., Barwell, L., Bowles, P., Fitt, R., Pavitt, A., Spriggs, R., and Isaac, N. (2014). Morphological and geographical traits of the British Odonata. *Biodiversity Data Journal*, **2**, e1041.

Strobl, C., Hothorn, T., and Zeileis, A. (2009). Party on! A new, conditional variable importance measure for random forests available in the party package. *The R Journal*, **1**, 14 – 17.

Taylor, P., Carvalho, L., Law, A., Baker, A., Chapman, D., Willby, N., Wilkie, C. (in preparation). The connectivity of UK freshwaters.