Categorical structured additive regression for the assessment of habitat suitability in the spatial distribution of mussel seed abundance

M. P. Pata[1,*], T. Kneib[2], C. Cadarso-Suárez[3], V. Lustres-Pérez[1] and E. Fernández-Pulpeiro[1]

[1] *Department of Zoology and Physical Anthropology, School of Biology, University of Santiago de Compostela (Spain); maria.pazos2@rai.usc.es, vicente.lustres@usc.es*
[2] *Department of Mathematics, Carl von Ossietzky Universität Oldenburg (Germany); thomas.kneib@uni-oldenburg.de*
[3] *Unit of Biostatistics, School of Medicine, University of Santiago de Compostela (Spain);carmen.cadarso@usc.es*
[*] *Corresponding author*

Abstract. *Categorical regression models are used to model relationships between a polytomus response and a set of regressor variables. Since these responses can be ordered or unordered, different models are proposed. In our example, the study of the distribution and abundance of mussel seed, the response variable is ordinal. Hence, a cumulative model would be the natural option, but we also opt for a multinomial model, since it is plausible that different effects can be present for each category. Therefore, the use of both models could be much more informative from a biological point of view.*

Keywords. *Categorical structured additive regression; Spatial effects; Mussel seed.*

1 Introduction

In ecology, Generalized Linear Models (GLM, [9]) have been the most widely used statistical models to assess relationships between species distribution and environment, but nowadays, the use of the Generalized Additive Models (GAM,[6] ; [10]), is extended to species distribution modelling, due to its ability to handle non-linear data [1], [5]. However, spatial autocorrelation often exists in this kind of data [8]. Since the spatial correlation is difficult to handle in the GAM framework, a more general regression model is then required. One of these kind of models are the Bayesian Structured Additive Regression models (STAR, [3]). These models have great advantages in the context of spatial data, as they allow to incorporate smooth effects of continuous covariates and spatial effects with flexible forms. The spatial effects can be split in a conditional (structured) part, and an unconditional (unstructured) part which allows to explain the overdispersion caused by unobserved heterogeneity or the presence of autocorrelation

[3]. In the present work, we model the distribution and abundance of mussel seed using the STAR models for categorical responses, based on a mixed model representation. The mussel is the most important marine resource in galician aquaculture sector, producing over 95% of mussel in Spain, the first producer of this resource in Europe (over 40% of production). This important sector depends on natural resources of mussel seed, so it is necessary to increase the knowledge of factors regulating the spatial distribution and abundance of this resource.

2 The mussel seed data

The study was carried out in 93 sites on the galician atlantic coast, during spring tides, from march to september, in 2005-2006 and 2008-2009. In each site, a transect perpendicular to the coastal line was placed in the intertidal zone. A sample quadrat of 20x20 cm was setting each 50 cm, and the percentage cover of mussel seed was measured. In the present work, two covariates have been selected: tidal height (in meters) and magnetic course (in degrees).
The response is given by a categorical variable representing percentages of observation plots covered by mussel seed, with four categories: (a) category 1 (reference category): low abundance $\leq$5%, (b) category 2: medium (5%-25%], (c) category 3: high (25%-50%], and (d) category 4: very high $>$50%. All computations are carried out with BayesX package [2].

3 Statistical methodology: Categorical structured additive regression

Categorical regression models are used to model relationships between a polytomus response and a set of regressor variables. Since these responses can be ordered or unordered, different models are proposed. In our example, the response variable is ordinal and a cumulative model would be the natural option. However, we also opt for a multinomial model, since it is plausible that different effects can be present for each category. Therefore, the use of both models could be much more informative from a biological point of view.
For nominal responses, the multinomial logit is the most common choice, where the probability of category r is defined as

$$P(Y=r)=\pi^{(r)}=h^{(r)}\left(\eta^{(1)},\ldots,\eta^{(q)}\right)=\frac{exp(\eta^{(r)})}{1+\sum_{s=1}^{q}exp(\eta^{(s)})}$$

with linear predictor $\eta^{(r)}=u'\alpha^{(r)}$ depending on covariates u and category-specific vector of regression coeficients $\alpha^{(r)}$. For ordered responses, the usual choice is the cumulative logit model, defined via the cumulative distribution function F of the standard extreme value distribution as $P(Y\leq r)=F\left(\eta^{(r)}\right)$ or

$$P(Y=r)=\pi^{(r)}=h^{(r)}\left(\eta^{(1)},\ldots,\eta^{(q)}\right)=F\left(\eta^{(r)}\right)-F\left(\eta^{(r-1)}\right),r>1$$

with linear predictor $\eta^{(r)}=\theta^{(r)}-\upsilon'\alpha^{(r)}$, where $\theta^{(1)}<\ldots<\theta^{(q)}$ are the ordered thresholds.
With the specification of categorical structured additive regression models, it is possible to include nonlinear effects of continuous covariates and spatial effects in an unified and flexible framework [3] [7]. The resulting predictors for both the multinomial logit and cumulative logit are be the following

$$\eta_i^{(r)}=\upsilon_i'\alpha^{(r)}+f_1^{(r)}(x_{1l})+\ldots+f_l^{(r)}(x_{il})+f_{spat}^{(r)}(s_i)$$

$$\eta_i^{(r)} = \theta_i^{(r)} - \upsilon_i'\alpha - f_1(x_{1l}) - \ldots - f_l(x_{il}) - f_{spat}(s_i)$$

where $f_1, \ldots, f_l$ are smooth functions of the covariates $x_1, \ldots, x_l$, and f_{spat} is the non-linear effect of spatial index $s_i \in \{1, \ldots, S\}$. For the multinomial model, the covariates are assumed to be independent of the category while effects are category-specific. For the ordinal model, all effects apart from the thresholds are constant across categories.

To estimate smooth effect functions and model parameters, an empirical Bayesian approach, based on mixed model representation is employed. Since we use a Bayesian approach, it is necessary to specify priors for the effects. For the fixed effects parameter γ, diffuse priors $p(\gamma) \propto const$ are assumed. For the unknown smooth functions f we assume Bayesian penalised splines with second order random walk priors. For the structured spatial effects, a Markov Random Field prior is selected, since the regions are clustered in connected geographical regions [3].

Inference is performed with empirical Bayes (EB) posterior analysis based on generalized linear mixed model (GLMM) methodology, once an appropiate reparameterization of the regression terms is given. Based on the GLMM approach, regression and variance parameters can be estimated using iteratively weighted least squares (IWLS) and (approximate) restricted maximum likelihood (REML) developed for GLMM's [4]; [7].

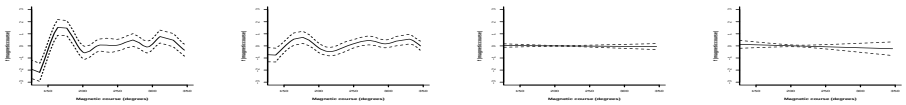
4 Results and conclusions

Multinomial and cumulative logit models were applied to assessing the possible influence of tidal height and magnetic course on the spatial distribution of mussel seed abundance. The estimated smooth effects of the two covariates are shown in Figure 1. Results from the cumulative model indicate that higher values of the predictors correspond to lower abundance of mussel seed. While the functional effect of tidal height is quite similar for the ordinal model and the multinomial model for categories 2 and 3, some deviations from this pattern were found for category 4. As regards of magnetic course, qualitatively similar results were obtained for both the ordinal model and the multinomial model for category 2, while no effect of this covariate was encountered for the two remaining categories, 3 and 4.

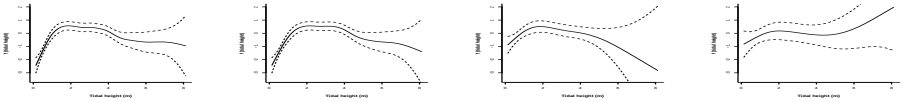
Almost no spatially structured effects were found when using ordinal model. However, a more detailed picture through the separated effects given by the multinomial model indicates that spatially structured effects are quite pronounced for categories 3 and 4. Also, larger variation of the unstructured effect among categories was observed, as compared to that encountered for the structured effect.

As a general conclusion, compared to the ordinal model, the multinomial model provides more insight into the factors that influence the spatial distribution of mussel seed. Since these models allow for different covariate effects along the various response categories, we can identify whether factors determining the presence of the resource are different from those determining its abundance.

Acknowledgments. The authors would like to express their gratitude for the support received in the form of the Spanish MEC Grant MTM2008-01603 and the Galician Regional Authority (Xunta de Galicia) projects INCITE08PXIB208113PR and 07MMA001200PR.



(a) Estimated effects of magnetic course.



(b) Estimated effects of tidal height.

Figure 1: Estimated smooth effects of covariates with 95% pointwise credible intervals, for cumulative (first column), and multinomial models: category 2 (second column), category 3 (third column) and category 4 (fourth column). Category 1 (<5%) is taken as reference category.

(a) Posterior mode estimates of the structured spatial effects.

(b) Posterior mode estimates of the unstructured spatial effects

Figure 2: Estimated spatial effects, for cumulative (first column), and multinomial models: category 2 (second column), category 3 (third column) and category 4 (fourth column).

References

[1] Austin, M. P.(2002). Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological Modelling* **157**, 101–118.

[2] Belitz, C., Brezger, A., Kneib, T., Lang, S. (2009): *BayesX - Software for Bayesian inference in structured additive regression models*. Version 2.0.

[3] Fahrmeir, L. and Lang, S. (2001). Bayesian inference for generalized additive models based on Markov random field priors. *Applied Statistics*. **50 (2)**, 201–220.

[4] Fahrmeir, L., Kneib, T. and Lang, S. (2004). Penalized structured additive regression for space-time data: a Bayesian perspective. *Statistica Sinica*. **14**, 731–761.

[5] Guisan, A., Edwards, T.C. and Hastie, T. (2002). Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological Modelling*. **157**, 89–100.

[6] Hastie, T. J and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman and Hall, London.

[7] Kneib, T. and Fahrmeir, L. (2006). Structured additive regression for categorical space-time data: a mixed model approach. *Biometrics*. **62**, 109–118.

[8] Kneib, T., Müller, J. and Hothorn, T. (2008). Spatial smoothing techniques for the assesment of habitat suitability. *Environmental and Ecological Statistics*. **15**, 343–364.

[9] McCullagh, P., Nelder, J.A. (1997). *Generalized Linear Models*, second ed. Chapman and Hall, London.

[10] Wood, S. N. (2006). *Generalized additive models: an introduction with R*. CRC Press, Boca Raton, FL.