SORT 34 (1) January-June 2010, 67-78

Modelling spatial patterns of distribution and abundance of mussel seed using Structured Additive Regression models

María P. Pata[1], María Xosé Rodríguez-Álvarez[2,3], Vicente Lustres-Pérez[1]
Eugenio Fernández-Pulpeiro[1], Carmen Cadarso-Suárez[2,3]

Abstract

As mussel farming depends on sources of natural mussel seed, knowledge of factors is required to regulate both the spatial distribution and abundance of this resource. These spatial patterns were modelled using Bayesian STructured Additive Regression (STAR) models for categorical data, based on a mixed-model representation. We used Bayesian penalized splines for modelling the continuous covariate effects and a Markov random field prior for estimating the spatial effects.

MSC: 62F15, 62G08, 62P12, 92D40.

Keywords: Mussel seed, Bayesian structured additive regression (STAR) models, spatial effects, Bayesian P-splines.

1. Introduction

Knowledge of spatial patterns of distribution and abundance of species is essential in order to understand the ecological processes that have generated such processes (Underwood, Chapman and Connell (2000)). In the case of marine resources, knowledge of these patterns is of crucial interest.

Mussel farming is widely developed along most of Galicia's Atlantic coastline, and indeed this region is the largest producer in Europe (200,000 MT/year). As mussel

[1] Departamento de Zoología y Antropología Física, Universidade de Santiago de Compostela (USC), Spain.
[2] Unidad de Bioestadística, Departamento de Estadística e Investigación Operativa, USC, Spain.
[3] Instituto de Investigación Sanitaria de Santiago (IDIS), Santiago de Compostela, Spain.
Received: November 2009
Accepted: May 2010

farming depends on natural mussel seed resources, knowledge of its distribution and abundance is fundamental to prevent depletion of natural populations.

In ecology, Generalized Linear Models (GLM, McCullagh and Nelder (1997)) are the most widely used statistical models to assess relationships between species distribution and environment. In recent years, however, biomedical researchers have shown a great interest in the use of Generalized Additive Models (GAM, Hastie and Tibshirani (1990); Wood (2006)), due the latter's ability to cover the complex non-linear effects had by continuous covariates on the outcome of interest. Recent applications of GAMs in ecology (see, for instance, Austin (2002), Austin (2007), Guisan, Edwards and Hastie (2002)) show that GAM regression models are useful tools for analysing relationships between species' distributions and their environment. Yet, spatial autocorrelation often exists in the data because the sample points are close to one another and subject to the same environmental factors (see Kneib, Müller and Hothorn (2008)). Since spatial correlation is difficult to handle within a GAM framework, a more general regression model is thus called for.

Accordingly, this study modelled the spatial distribution of mussel seed within a Bayesian STructured Additive Regression Model (STAR, Fahrmeir and Lang (2001)) framework. Inference was based on a mixed-model representation (Kneib and Fahrmeir (2006)). The use of STAR models affords several advantages when analysing spatial data, including, among others, the possibility of incorporating: (a) flexible forms of the effects of continuous covariates, by using Bayesian P-splines (Eilers and Marx (1996), Lang and Brezger (2004)); (b) flexible spatial effects; and, (c) random effects to explain the overdispersion caused by unobserved heterogeneity or the presence of autocorrelation in spatial data (Fahrmeir and Lang (2001)). Models that enable smooth effects of continuous covariates and spatial effects with flexible forms to be incorporated are known as geoadditive models (Kammann and Wand (2003)). In this paper, we used a geoadditive multicategorical regression model (Kneib and Fahrmeir (2006)), in which the response variable was assumed to follow a multinomial distribution.

The paper is structured as follows: the mussel seed data are introduced in Section 2; the statistical methodology is described in Section 3; the results from fitting the proposed STAR models to mussel seed data are shown in Section 4; and the paper concludes with a Discussion Section.

2. The mussel seed data

This study was undertaken during spring tides at 62 sites along Galicia's Atlantic seaboard, between 43°21′ N, 8°21′ W and 42°44′ N, 9°04′ W, from March to September in 2005 and 2006.

At each site, a transect perpendicular to the coastline was placed in the intertidal zone. A sample quadrant (20×20 cm) was set at 50-centimetre intervals and the percentage cover of mussel seed then measured. Information from a set of covariates

was taken in order to explain the distribution pattern of the mussel seed. These covariates were tidal height (in metres), percentage of pools, and positioning related to cardinal points divided into the following five categories: NN; NE; SE; SW; and NW.

For study purposes, the outcome of interest was percentage cover (from 5% upwards, in multiples of 5%). This variable was treated as categorical, and the following four categories were established:

Category 1: low abundance, [0% - 5%]. This was used as the reference category;
Category 2: medium, (5% - 25%];
Category 3: high, (25% - 50%];
Category 4: very high, $> 50\%$.

Within the STAR framework, several approaches can be used to analyse categorical responses, such as the multinomial model for nominal categories or the cumulative logit probit models, among others, for ordered categories. Despite the fact that a better option might have been the cumulative model, we nevertheless chose to use a multinomial model in view of the biological interest that this option could afford.

All computations were performed using the BayesX package (Belitz *et al.* (2009)).

3. Statistical methodology: geoadditive multicategorical regression model

In multicategorical data the response variable Y is observed in categories $r \in (1,\ldots,k)$. Analysis of this type of data calls for an appropriate model to take into account the additional information supplied by these categories (Boeck and Wilson (2004)). In this paper, a multinomial logit model was considered, with the probability of the category r expressed as follows:

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

with k as reference category, and the linear predictor $\eta^{(r)} = u'\alpha^{(r)}$, depending on covariates u and category-specific vector of regression coeficients $\alpha^{(r)}$. It is possible to obtain the general multinomial model

$$\pi = h(\eta), \qquad \eta = V\gamma,$$

by defining the design matrix

$$V = \begin{pmatrix} \upsilon_1' \\ \vdots \\ \upsilon_q' \end{pmatrix} = \begin{pmatrix} u' & & 0 \\ & \ddots & \\ 0 & & u' \end{pmatrix}$$

and the overall vector of regression parameters (Kneib (2006); Kneib and Fahrmeir (2006))

$$\gamma = \left(\alpha^{(1)}, \ldots, \alpha^{(q)}\right).$$

To take into account the spatial information for each unit (administrative areas in our example), the following geoadditive multicategorical model (defined by the geoadditive predictor) is then considered

$$\eta_i^{(r)} = u_i' \alpha^{(r)} + f_1^{(r)}(x_{i1}) + \ldots + f_l^{(r)}(x_{il}) + f_{spat}^{(r)}(s_i),$$

where $f_1^{(r)}, \ldots, f_l^{(r)}$ are unknown smooth functions of the covariates $x_1, \ldots, x_l$, and $f_{spat}^{(r)}$ is the non-linear effect of spatial index $s_i \in \{1, \ldots, S\}$ (administrative area in our example).

This specification of the model allows for flexible incorporation of non-linear effects of continuous covariates and spatial effects. Furthermore, the different types of covariates are considered in a unified framework (Fahrmeir and Lang (2001); Kneib and Fahrmeir (2006)).

Since spatial correlation and/or heterogeneity due to unobserved spatially varying covariates are usually present in spatial data, it seems appropiate for the spatial effect to be broken down into a spatially correlated part (structured part: f_{str}) and a spatially uncorrelated part (unstructured part: f_{unstr}):

$$f_{spat}^{(r)}(s) = f_{str}^{(r)}(s) + f_{unstr}^{(r)}(s).$$

This representation of the spatial effects makes it possible to distinguish between the two kinds of unobserved covariates, namely, those that display a strong spatial structure and those that are present locally (Besag, York and Mollié (1991); Fahrmeir *et al.* (2003)).

To estimate smooth effect functions and model parameters, an empirical Bayesian approach based on mixed model representation is used. Assigning appropriate priors for parameters and functions is crucial. For the fixed effects parameter γ, diffuse priors $p(\gamma) \propto const$ are asssumed.

For specifying smoothness priors for continous covariates, a Bayesian version of the P-splines approach of Eilers and Marx (1996) is used (Lang and Brezger (2004)). This approach assumes that the effect f of a covariate x can be approximated by a polinomial spline of degree l defined on a set of equally spaced knots $x_{min} = \xi_0 < \xi_1 < \ldots < \xi_{r-1} < \xi_r = x_{max}$. This can be written in terms of a linear combination of $M_j = r_j + l_j$ B-spline basis functions

$$f_j(x) = \sum_{m=1}^{M_j} \beta_{jm} B_m(x),$$

where β_j is the vector of the unknown regression coefficients.

The main problem when dealing with these splines lies in the selection of the number of knots and their placement. The idea of P-splines is to select a generous number of knots and define a roughness penalty on adjacent regression coefficients to regularise the problem and avoid overfitting (Eilers and Marx (1996)). In the frequentist approach, first- or second-order differences are usually used. From a Bayesian perspective, these are replaced by their stochastic analogues, namely, first-or second-order random walks. For the purposes of this study, we used second-order random walks for the regression coefficients, defined as

$$\beta_{jm} = 2\beta_{j,m-1} - \beta_{j,m-2} + u_{jm},$$

with Gaussian errors $u_{jm} \sim N\left(0, \tau_j^2\right)$. (Lang and Brezger (2004)). The variance parameter τ_j^2 controls the amount of smoothnes.

Since the spatial locations are clustered in connected geographical regions, a Markov Random Field prior (Besag *et al.* (1991)) is selected for the structured spatial effects. This spatial smoothness prior is defined by

$$\left\{f_{str}(s)|f_{str}(s'); \, s \neq s', \tau^2\right\} \sim N\left(\sum_{s\in\delta_s} \frac{f_{str}(s')}{N_s}, \, \frac{\tau^2}{N_s}\right),$$

where N_s is the number of adjacent sites, $s \in \delta_s$ ndicates that site $s^{'}$ is neighbour of site s, that is, they share a common boundary (Fahrmeir and Lang (2001), Kneib (2006)).

The unstructured spatial effects are assumed to be i.i.d. random effects $f_{unstr}(s) \sim N\left(0, \tau^2\right)$ (Fahrmeir *et al.* (2004); Kneib and Fahrmeir 2006).

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. For empirical Bayes inference, the variances τ_j^2 are considered as unknown constants to be estimated from their marginal likelihood. 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. For detailed description of the estimation procedure see Fahrmeir *et al.* (2004) and Kneib and Fahrmeir (2006).

4. Results

To analyse the spatial distribution of mussel seed with respect to the relevant explanatory variables, a geoadditive multinomial logit model was applied. The parametric effects of sites' positioning in terms of cardinal points as well as the smooth effects of tidal height and percentage of pools were included in the model.

A summary of the estimated effects of site positioning for each category is shown in Table 1. The category with the highest frequency, SW, was chosen as the reference category. As can be seen from Table 1, the results were only significant for Category 2 (mussel seed abundance of 5%-25%), and as NN-, NE- and NW-positioning of sites reduced the presence of this category, SW-positioning was therefore the best for the presence of Category 2.

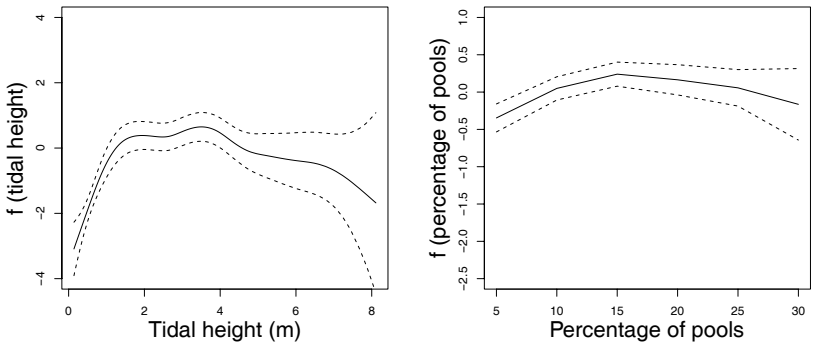
Table 1: *Estimates, standard deviations (S.D) and 95% credible confidence interval for the fixed effects. Category 1 (< 5%) is taken as reference category.*

	Estimated effects	S.D.	95%CI	
Category 2 (5%-25%]				
NN	−1.73	0.662	−3.02	−0.43
NE	−0.86	0.179	−1.21	−0.51
SE	−0.17	0.280	−0.72	0.37
NW	−0.36	0.161	−0.68	−0.04
Category 3 (25%-50%]				
NN	−0.04	0.435	−0.86	0.54
NE	−0.03	0.365	−0.74	0.68
SE	0.41	0.441	−0.44	0.44
NW	0.06	0.251	−0.42	0.25
Category 4 $>$ 50				
NN	−0.49	0.687	−1.04	0.06
NE	−0.09	0.769	−1.60	1.41
SE	0.36	0.784	−1.16	1.90
NW	0.44	0.553	−0.64	1.53

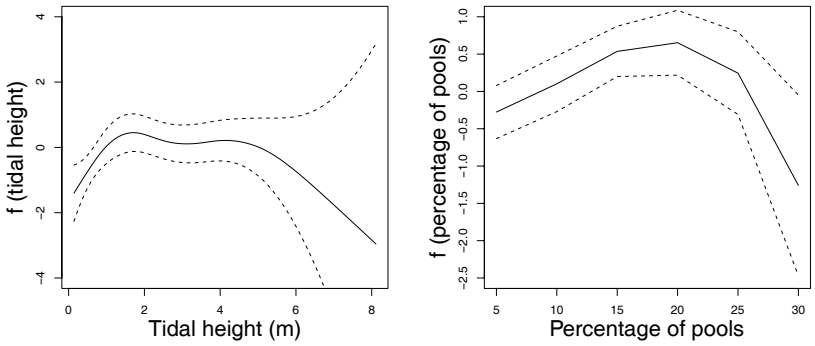
The estimated smooth effects of tidal height and percentage of pools on the presence of mussel seed are shown in Figure 1. As can be seen, these are complex nonlinear effects. Hence, the use of purely linear models to describe such data could lead to estimations and, by extension, conclusions that were erroneous. STAR models enable flexible forms of the effects of continuous covariates to be incorporated in the response and better knowledge of the biological process so obtained.

The effect of tidal height appeared to be similar for the above three categories in the initial metres, with a much more pronounced shape for Category 2. The function increased until a tidal height of about two metres, and then decreased from 4 metres. It seems that the most suitable tidal heights for the presence of mussel seed range from 2 to 4 metres, particularly for Category 2. For Category 3, tidal height appeared to have no effect from a height of about 2 metres.

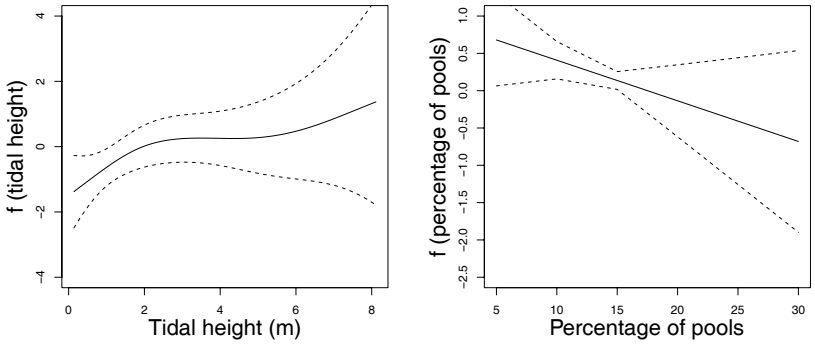
The effects of percentage of pools are depicted in the right panels of Figure 1. For the presence of Categories 2 and 3, which plotted similar patterns, sites with 15% to 20% of pools would seem to be more suitable. The presence of these categories decreased



(a) Category 2 (5%-25%]



(b) Category 3 (25%-50%]



(c) Category 4 (>50)

Figure 1: *Estimated smooth effects of tidal height (left panel) and percentage of pools (right panel),with 95% pointwise credible intervals. Category 1 (<5%) is taken as reference category.*

thereafter but posterior probabilities were non-significant above 20%. For Category 4 (very high abundance), in contrast, the presence of mussel seed decreased linearly with percentage of pools.

Figure 2 displays the spatial effects on a grey scale, after controlling for the covariates: white colour refers to a positive spatial effect signifying higher abundance of the category, while dark colour refers to a negative effect signifying lower abundance. The significance of the structured spatial effects are shown in the third column of Figure 2, with black areas indicating strictly negative credible intervals, white ones indicating strictly positive, and grey areas indicating no effect. In cases where the structured spatial effects proved non-significant, the map of posterior probabilities is not shown.

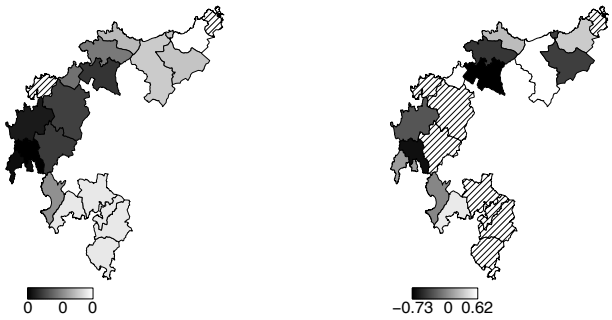
As can be seen from the maps (Figure 2, first row), the structured spatial effects for Categories 3 and 4 displayed a clear regional pattern but were not significant for Category 2. There is a descending south-north gradient, with southern regions appearing to be more appropriate and northern areas unsuitable for the presence of mussel seed. These results seem to be plausible because the northern areas are extremely exposed and steep, and even the nature of the substrate is less suitable for settlement and, by the same token, a higher abundance of mussel seed.

In the maps of unstructured effects (Figure 2, second row), the dashed areas denote regions in which unstructured effects are not estimated. No clear pattern in unstructured effects is displayed in these maps and, compared to the structured effects, the local effects were smaller. Moreover, the posterior probabilities (maps not shown) indicate that no region has a significant effect on response.

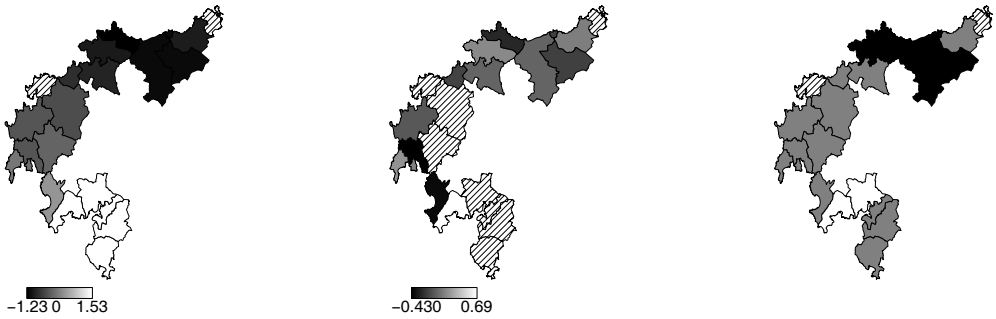
5. Conclusions

This study proposes a novel application of STAR models to the field of marine resources. The use of geoadditive multicategorical models for mussel seed data demonstrates that these models can be very useful tools for fitting this type of biological data. STAR models enable flexible non-linear effects of covariates as well spatial effects to be incorporated. Moreover, since spatial effects can be split into spatially correlated and uncorrelated parts, it becomes possible to distinguish between unobserved covariates that display a strong spatial structure and those that are only present locally. Our data revealed marked, downward, south-north spatial pattern. However, since the unstructured effects were not significant, the distribution of the mussel seed would not seem to be affected by locally present covariates.

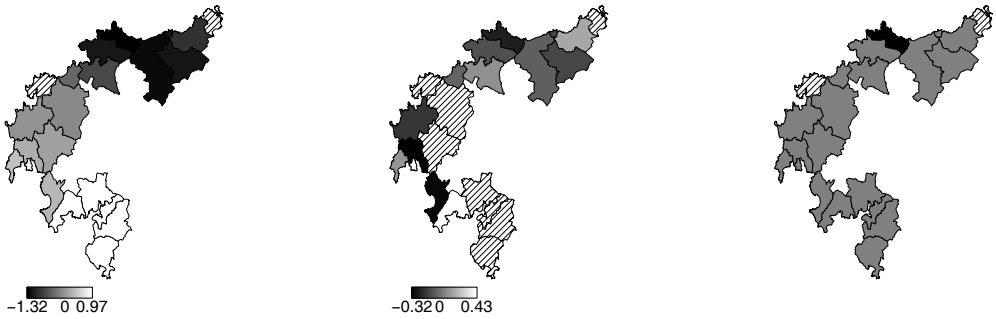
As pointed out in Section 2 above, though the response variable was treated as nominal (with multinomial distribution) in this study, this outcome could also be considered ordinal, in which case other STAR models, such as cumulative probit/logit models, could be used in our application. Future extensions of our work include a statistical comparison of categorical versus cumulative models for fitting mussel seed distribution.



(a) Category 2 (5%-25%]



(b) Category 3 (25%-50%]



(c) Category 4 (>50%)

Figure 2: *From left to right, averages estimates of the structured spatial effects (first column), unstructured spatial effects (second column) and posterior probabilities (third column). Category 1 (<5%) is taken as reference category. For category 2 the map of posterior probabilities is not shown since the structured spatial effects proved non-significant.*

Finally, an additional advantage of using STAR models for fitting ecological data lies in the flexibility of incorporating temporal effects in a simple manner, something that makes it possible to offer flexible spatio-temporal models, which are of great interest in many biomedical fields.

Acknowledgements

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. We are also grateful to the referee for her/his valuable comments and suggestions, which served to make a substancial improvement to this paper.

References

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

Austin, M. P. (2007). Species distribution models and ecological theory: a critical assessment and some possible new approaches. *Ecological Modelling*, Review, 200, 1-19.

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

Besag, J., York, J. and Mollié, A. (1991). Bayesian image restoration with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics*, 43, 1-59.

Boeck, P. and Wilson, M. eds. (2006). *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*. Springer, New York.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing using B-splines and penalties (with comments and rejoinder). *Statistical Science*, 11, 89-121.

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

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

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.

Kammann, E. E. and Wand, M. P. (2003). Geoadditive models. *Journal of the Royal Statistical Society C*, 52, 1-18.

Kneib, T. (2006). Mixed model based inference in structured additive regression. PhD thesis, Dr.Hut-Verlag.

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

Kneib, T., Müller, J. and Hothorn, T. (2008). Spatial smoothing techniques for the assessment of habitat suitability. *Environmental and Ecological Statistics*, 15, 343-364.

Lang, S. and Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics*, 13, 183-212.

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

Underwood, A. J., Chapman, M. G. and Connell, S. D. (2000). Observation s in ecology: you can't make progress on processes without understanding the patterns. *Journal of Experimental Marine Biology and Ecology*, 250, 97-115.

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