Boosted negative binomial hurdle models for spatiotemporal abundance of sea birds

Benjamin Hofner[1], Adam Smith[2]

[1] Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
[2] University of Rhode Island, USA

E-mail for correspondence: benjamin.hofner@fau.de

Abstract: Modelling the abundance of sea birds is hampered by various difficulties: Sightings might be scarce (excess of zeros), overdispersion might be present, and the effects of biophysical covariates might be non-linear. Additionally, one should consider the spatiotemporal data structure in the model and apply variable selection to achieve a sparse model representation. We propose a spatiotemporal negative binomial hurdle model that considers all raised issues. Variable selection and model choice are achieved by boosting methods with stability selection.

Keywords: GAMLSS; variable selection; spatiotemporal modelling; hurdle model.

1 Background

Nantucket Sound, Massachusetts, USA, is an important wintering area for seaducks in southern New England. Wind energy development has been fully permitted on 62 km^2 of Horseshoe Shoal in the northwest portion of Nantucket Sound. We conducted 30 aerial strip-transect (180 m wide) surveys throughout 1,100 km^2 of Nantucket Sound (Figure 1) during the winters of 2003–2005 to evaluate sea duck distribution and relative abundance. Here we consider counts only for Common Eider (*Somateria mollissima*). We aggregated eider counts within 2.25km^2 segments in the study area (Figure 1).

2 Modelling approach

We related spatiotemporal variation in sea duck occupancy (presence/ absence) and abundance to potentially relevant biophysical and spatiotemporal covariates. The data shows a high prevalence of zero counts (e.g., 75 % of segments contained no eider observations). Thus, we applied a

This is a slightly modified version of a paper which was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

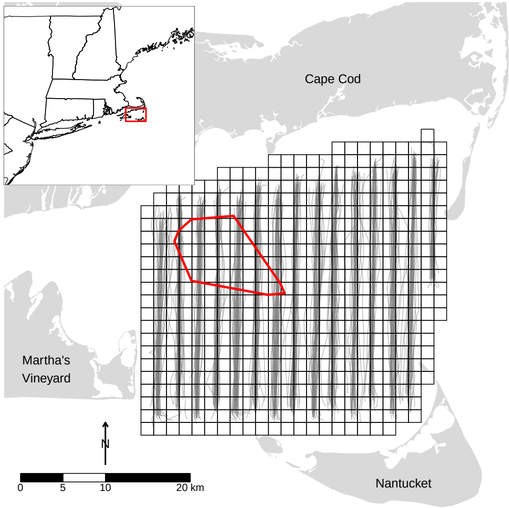

FIGURE 1. Nantucket Sound, Massachusetts, US study area. The grid indicates the extent of the study area and its division into 504 2.25km^2 segments. Gray lines indicate all aerial transects flown over the course of the study. The red polygon indicates the location of permitted wind energy development on Horseshoe Shoal.

negative binomial hurdle model that separately modeled the probability of occurrence of at least one individual (*"occupancy model"*; logistic regression model) in a given segment and the abundance of eider in that segment conditional on their presence (*"count model"*; truncated negative binomial model). The negative binomial density was specified such that the mean counts are given by $E(Y|x) = \mu$ and the variance as $Var(Y|x) = \mu(1 + \mu\sigma)$ with dispersion parameter σ.

We evaluated biophysical covariates expected to influence the distribution and availability of benthic prey or the distribution, abundance, and movements of eider. Relevant biophysical covariates included, for example, bathymetry ($depth$), distance to nearest land ($dist$), sea floor roughness (SAR), sea bottom sediment grain size (SGS), tendency for summer stratification of the water column ($strat$), sea surface temperature relative to other segments (SST_r), average monthly sea surface temperature (SST_m), water temperature near the sea bottom (SBT), dissolved organic material concentrations (DOM), average epibenthic tidal velocity (ETV_{avg}) and standard deviation of epibenthic tidal velocity (ETV_{sd}). All continuous covariates were standardized, i.e., mean centered and scaled, before entering the model. We specified *a priori* expected time-varying effects for relative sea surface temperature and water depth to allow the effects to vary over time within a given winter. Additionally, we include spatial, temporal and spatiotemporal effects to account for unmeasured heterogeneity.

We used GAMLSS (Rigby & Stasinopoulos, 2005) to model the zero-truncated negative binomial *count model*, i.e., we regressed both the mean μ and the dispersion parameter σ on covariates. The *occupancy model* was based on a generalized (logistic) additive model, where the expected occupancy $E(Y = 1|x) = \pi$ was modeled.

2.1 Boosting

Model-based boosting (Hothorn *et al.*, 2010) was applied to fit all models and to allow for the selection of variables: Boosting aims at minimizing the negative log-likelihood of the model at hand. In each step, all effects are fitted separately to the gradient of the log-likelihood and only the best-fitting effect is selected and updated. For the GAMLSS *count model* we cycled through both parameters, location and dispersion, and for each parameter through all specified effects (Mayr *et al.*, 2012). In both the count and the occupancy model we used smooth P-spline base-learners and applied a model decomposition (Hofner, Hothorn, Kneib & Schmid, 2011) to allow the boosting algorithm to select the appropriate model complexity.

2.2 Cross-validation

We used 25-fold subsampling to determine the optimal stopping iteration for each model. This is done to prevent overfitting and to minimize the (out-of-bag) prediction error (measured by the negative log-likelihood of the model). Specifically, we randomly drew (without replacement) data sets of size $n/2$ from the original data set. We used these data sets to estimate the model and used the other half of the data to determined the out-of-bag prediction accuracy.

2.3 Stability selection

To obtain even sparser models we used stability selection (Meinshausen & Bühlmann, 2010) to extract variables and effects that were stably selected while controlling the per-family error rate. Stability selection and an improved version (complementary-pairs stability selection; Shah & Samworth, 2013) was recently investigated in the context of boosting models and showed promising results in conjunction with boosted GAMs (Hofner, Boccuto & Göker, 2015). Here, we applied complementary-pairs stability selection with unimodality assumption, used an upper bound for the per-family error rate of six and allowed each model to select $q = 35$ variables.

3 Results

3.1 Count model

We obtained sparse models that included only a small fraction of the initially specified effects. The model for the mean parameter of the *count*

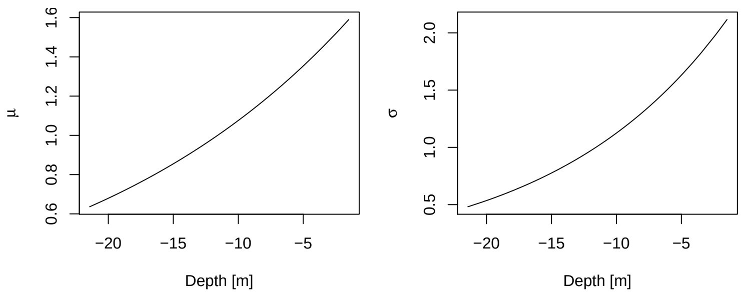

FIGURE 2. Partial effect of depth on the eider count μ and on the dispersion σ. Eider prefer shallower areas but at the same time counts show a greater variability in these areas.

model included categorical and linear effects (i.e., *ferry* and *depth*), and smooth effects f for various biophysical parameters. Temporal, spatial and spatiotemporal effects were not selected.

$$\begin{aligned}\log(\mu) = \text{ferry} + \text{depth} + \\ + f(\text{SAR}) + f(\text{SGS}) + f(\text{SST}_r) + f(\text{SBT}) + f(\text{DOM}).\end{aligned}$$

The dispersion of the *count model* included linear and smooth effects for various covariates. Additionally a temporal effect (*y2004*) and a spatial effect was selected. Spatiotemporal effects or time-varying effects were not selected.

$$\begin{aligned}\log(\sigma) &= \text{depth} + \text{SST}_m + \text{y2004} \\ &\quad + f(\text{DOM}) + f(\text{ETV}_{\text{avg}}) + f(\text{ETV}_{\text{sd}}) + f_{\text{spatial}},\end{aligned}$$

where μ is the (conditional) mean count of eiders, σ is the (conditional) dispersion in the negative binomial model. As an example, the effects of depth on μ and σ are depicted in Figure 2.

3.2 Occupancy model

The binomial logit model for *occupancy* was far more complex and included linear and smooth effects of various biophysical predictors, time-varying effects, spatial effects as well as a spatiotemporal effect:

$$\begin{aligned}\text{logit}(\pi) &= \text{ferry} + \text{SAR} + \text{DOM} + \text{y2005} \\ &\quad + f(\text{dist}) + f(\text{SGS}) + f(\text{strat}) + f(\text{SST}_m) + f(\text{SBT}) \\ &\quad + f(\text{time}) + f(\text{depth}, \text{time}) + f_{\text{spatial}} + f_{\text{spatial}} \cdot \text{time},\end{aligned}$$

where π represents the (conditional) occupancy probability. Additionally, the occupancy probability was dependent on the size of the observation window (i.e., the area of transect surveyed in a given segment). The spatiotemporal effect is displayed in Figure 3. A slight shift of occupancy during the winter, which is not covered by other measured covariates, is apparent at the end of the winter season.

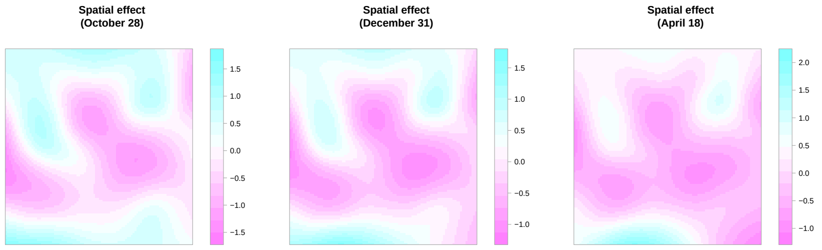

FIGURE 3. Partial spatiotemporal effect of the eider occupancy probability during winter season.

4 Implementation

All models were fitted using the statistical environment R. The *occupancy model* was fitted using the package **mboost** (Hothorn *et al.*, 2010, 2015; Hofner, Mayr, Robinzonov & Schmid, 2014), the *count model* was fitted using the package **gamboostLSS** (Hofner, Mayr, Fenske & Schmid, 2015; Mayr *et al.*, 2012) and stability selection was implemented in the package **stabs** (Hofner & Hothorn, 2015, Hofner, Boccuto & Göker, 2015).

5 Summary

Our boosting approach results in a very flexible model class for count data: It allows to incorporate excess zeros and to model the dispersion. Smooth, spatial and spatiotemporal effects can be easily included. Model choice and variable selection are available within the boosting framework. Additional sparsity with error control can be obtained by applying stability selection.

Acknowledgments: Special thanks to Massachusetts Audubon (particularly T. Allison and S. Perkins) for providing the aerial survey data. J. Osenkowski (Rhode Island Department of Environmental Management) and P. Paton and S. McWilliams (University of Rhode Island) funded A. Smith's contribution.

References

Hofner, B., Boccuto, L., and Göker, M. (2015). Controlling false discoveries in high-dimensional situations: Boosting with stability selection. *Accepted for publication in BMC Bioinformatics.*

Hofner, B., and Hothorn, T. (2015). stabs: Stability Selection with Error Control, *R package version 0.5-1*, http://cran.r-project.org/package=stabs.

Hofner, B., Hothorn, T., Kneib, T., and Schmid, M. (2011). A Framework for Unbiased Model Selection Based on Boosting. *Journal of Computational and Graphical Statistics*, **20**, 956 – 971.

Hofner, B., Mayr, A., Fenske, N., and Schmid, M. (2015). gamboostLSS: Boosting Methods for GAMLSS Models, *R package version 1.1-3*, http://cran.r-project.org/package=gamboostLSS.

Hofner, B., Mayr, A., Robinzonov, N., and Schmid, M. (2014) Model-based boosting in R – A hands-on tutorial using the R package mboost. *Computational Statistics*, **29**:3 – 35.

Hothorn, T., Bühlmann, P., Kneib, T., Schmid, M., and Hofner, B. (2010). Model-Based Boosting 2.0. *Journal of Machine Learning Research.* **11**, 2109 – 2113.

Hothorn, T., Bühlmann, P., Kneib, T., Schmid, M., and Hofner, B. (2015). mboost: Model-Based Boosting. *R package version R package version 2.4-2.* http://cran.r-project.org/package=mboost.

Mayr, A., Fenske, N., Hofner, B., Kneib, T., and Schmid, M. (2012). Generalized additive models for location, scale and shape for high-dimensional data - a flexible approach based on boosting. *Journal of the Royal Statistical Society: Series C (Applied Statistics).* **61**, 403 – 427.

Meinshausen, N. and Bühlmann, P. (2010). Stability Selection (with Discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **72**, 417 – 473.

Rigby, R.A. and Stasinopoulos, D.M. (2005). Generalized Additive Models for Location, Scale and Shape (with Discussion). *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **54**, 507 – 554.

Shah, R.D. and Samworth, R.J. (2013). Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **75**, 55 – 80.