Influence of offshore wind farms on distribution and abundance of Gaviidae: Methodological overview

Moritz Mercker - BIONUM

November 12, 2018

Abstract

We present a statistical approach to investigate the influence of offshore wind farms on the density and distribution patterns of Gaviidae. For this purpose, we combine and apply two state-of-the-art regression techniques frequently used in environmental modelling and impact studies, namely a "before-after control-impact" (BACI) approach as well as "generalised additive models" (GAM's). The combination of these approaches allows for a proper discrimination of natural/stochastic spatial-temporal distribution patterns from the pure effect of wind farms. Especially, the presented method provides (1) spatial distribution maps as well as total population numbers before vs. after wind farm construction; (2) the evaluation of a minimal significant avoidance distance; and (3) significance and effect size of abundance reduction in close vicinity to wind turbines.

1. Introduction

A frequent challenge in environmental impact studies is the discrimination of natural/stochastic spatio-temporal population fluctuations from the true influence of the impact. Simple before-after designs, for example, always suffer from the fact that population changes in the "impact-period" vs. the "before-period" may also represent natural population fluctuation between subsequent years. Analogously, a comparison of an "impact-area" with a "control area" within the same year always runs the risk to be biased due to spatial population fluctuations depending on other (often unknown) covariates than the impact.

Thus, the modern "before-after control-impact" (BACI) approach is increasingly suggested [25, 20, 26], since it allows to evaluate the possible effect of anthropogenic impacts on animal populations while eliminating the above mentioned bias due to natural spatial or temporal abundance fluctuations. Especially, an "impact area" and a "control area" are monitored before and after the activation of the impact, and relative comparisons of spatial and temporal differences allow to extract the unbiased impact [25, 26].

Although originally, such data have been evaluated with an ANOVA [26], it appears that BACI-analyses can also be formulated in the framework of modern regression techniques [25], since ANOVA and linear regression are highly related to each other [10]. This is an important feature in order to allow for an appropriate analysis, since ecological data often require regression methods specifically tailored to an actual data structure and research question [33, 17, 5, 14].

Especially, the use of modern regression techniques e.g. enables that count data can be modelled with appropriate probability distributions in the framework of generalised linear modelling [8, 16, 15, 18, 30, 19, 38] or that highly nonlinear dependencies (as frequently observed in ecological systems) can be appropriately described based on additive modelling [12, 11, 32, 34].

Indeed, in the presented study, the combination of the above mentioned modelling approaches is required in order to obtain unbiased results, since we deal with overdispersed count data as well as highly nonlinear relationships (such as the dependency of the abundance on the distance to a wind turbine). This finally leads to a BACI analysis formulated within the framework of generalised additive models (GAM's), the latter being a well established analysis method in the context of ecological studies and is frequently described in the literature of statistical ecology (e.g., [33, 37, 34, 35, 32, 11]).

2. Material and Methods

In this section we present in detail the underlying data and applied methods.

Bird count raw data

Bird count data are given based on observer-based aerial or ship surveys as well as on digital-based aerial surveys from 2000-2017 and restricted to the spring (01.03.-30.04.). The data comprise 25.077 detected Gaviidae and a total monitored area of 47.985 km^2. Furthermore, the data have been assigned to wind farm clusters based on the minimal distance to the next turbine (after construction); taken together the clusters ”Nördlich Borkum”, ”Helgoland”, ”Butendiek”, ”DanTysk”, as well as ”Bard/Austerngrund” have been investigated.

Correcting for incomplete bird detection

It is well known that at least for observer-based surveys, bird detection is incomplete. On the one hand, the probability of detection may decrease with the distance to the observer (where the shape and strength of distance-dependent decrease may again depend on other covariates such as survey method, bird flock size, or sea state [27, 28, 7]). On the other hand, also the detection probability on the transect line (i.e., distance-independent detection) may vary between different methods and may depend on additional covariates (such as sea state).

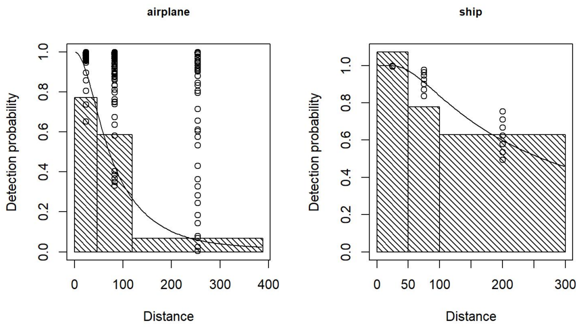

Figure 1: *Average detection functions in the context of distance sampling for observer-based counts from airplanes (left-hand side) as well as ships (right-hand side).*

To account for the first, we applied distance sampling methods (e.g., as presented in Ref. [6, 7]) to the observer-based raw data. Especially, we tested different detection functions (half-normal vs. Hazart-rate) as well as various different predictor combinations (main-effects as well as interaction terms based on the predictors *sea state, bird flock size, log(bird flock size)*) via AIC-analysis [33, 10] separately for aerial and ship-based data. The best detection function has been subsequently used to correct the raw-data case sensitive (i.e. depending on the distance class, the method, and all other covariates appearing as predictors in the best detection model). Raw data based on digital-based aerial surveys have not been distance-corrected, assuming that detection probability is distance-independent here.

To account for differences in the detection on the transect line, in final regression analysis (i.e., applied to distance-corrected data), the detection-related variables *sea_state* and *method* have been used as predictors, such that relative differences between methods are considered and estimated. Indeed, all three methods have various overlapping time frames, so that an estimation of such differences is possible.

Data pooling

For the sake of moderate computation times as well as a manageable amount of autocorrelation, it was necessary to pool the (distance-corrected) raw data. Hence, separately for each method (observer-based aerial survey vs. observer based ship survey vs. digital-based aerial surveys), and period (before construction vs. after construction), data have been spatially pooled in a grid with cells of $1.0 \times 1.0\,km$. This optimal grid size (i.e., large enough to be not too strongly autocorrelated, but small enough to provide a sufficient spatial resolution) has been determined based on semi-variogram- and bubble-plot-analyses of final regression model residuals (e.g., described in [17, 33]).

During the spatial pooling, for each grid cell separately, bird numbers and monitored area has been summed up whereas geographical coordinates and environmental covariates (such as water depth or distance to the coast) have been averaged.

Covariates

Beside the above mentioned detection-related categorical covariates *sea_state* and *method*, in final regression analyses (c.f., following subsections) we also considered smooth terms of the location-specific variables *dist_land* (= nearest distance to the mainland) and *depth* (=mean water depth). The aim was to further reduce the amount of unexplained variance and thus increase the power and quality of predictions in final regression models. These variables were given in an extra data sheet on a regular grid with a high spatial resolution of $0.2 \times 0.2\,km$, and have been transferred to the bird count data using cubic interpolation.
Furthermore, we introduced a 2D-spatial smooth predictor (especially a thin plate regression spline), depending on *Longitude* and *Latitude*. The aim was to account for additional spatial abundance heterogeneities not explained by the other used covariates. Since 2D thin plate splines are optimised for variables on the same scale [32], we rescaled geographical coordinates before analysis such that they are given in Kilometers.
For all smooth terms, the optimal amount of smoothing has been determined based on generalised cross-validation methods [32].

Importantly, this spatial smooth as well as the above mentioned environmental covariates were not allowed to vary between the periods (before construction vs. after construction) when direct effect of wind turbines has been evaluated. Thus, they account for long-term spatially varying bird densities, but do not interfere with the in the following introduced variables investigating the change in patterns and densities due to wind farm construction.

In order to investigate the influence of wind farm construction on bird distribution patterns and abundance, firstly, for each wind farm cluster separately, a ”before construction” and ”after construction” period has been defined and introduced as the additional binary variable *period*. Furthermore, the distance to the nearest wind turbine (after construction of all turbines) was given by the variable *dist_owp*, most importantly giving the same distance for both periods (before vs. after). Thus, in the before-period, this variable measures the distance to a wind turbine which does not yet exist, which is important to evaluate changes in response to this variable in the before vs. the after period.

The exact integration of *dist_owp* into regression analyses followed the general principles of a BACI-analysis [25, 20, 26]: If the change in abundance within a distinct impact-area (delimited from a control-area) has been investigated, the binary variable *B_dist_owp* has been defined (based on *dist_owp*), e.g. defining a certain proximity to a wind turbine as ”inside”, and ”outside” else. In regression analyses, B_dist_owp has been used as a main effect (just as the variable *period*) but also in interaction with the variable *period*. This interaction term $B\_dist\_owp : period$ finally represents the relative change in abundance in the impact vs. control area due to the wind turbines. Especially, it measures the relative difference in abundance in the impact vs. control area in the after-period, and corrects this value for the ”natural difference” between these areas as given in the before-period. This value is called the ”BACI-value” or ”BACI-reduction effect”.

In order to investigate the minimal avoid-distance (Meide-Distanz), B_dist_owp has been defined such that all locations with $dist\_owp > x\,km$ and $dist\_owp < y\,km$ have been defined as ”inside”, all locations with $dist\_owp > y\,km$ have been defined as ”outside”, and all location with $dist\_owp < x\,km$ has been excluded from the analysis. Thus, the impact area is given by a ”ring” or ”belt” around the wind turbines with inner radius x and outer radius y, where the control area is given by the area outside this belt, and the area inside is excluded from the regression. Stepwise increasing the diameter of the ring (using an annulus width of 3 km) and repeating the BACI-analysis allows to investigate at which distance to the wind turbines the BACI-effect within the ring is still significant, eventually leading to an estimate of the avoid-distance.

In order to investigate the overall effect of wind turbines on Gaviidae abundance, B_dist_owp has been defined such that $> 10\,km$ distance to a wind turbine has been defined as ”outside”, and ”inside” else. This order of magnitude has been derived from different studies (including this work), showing that Gaviidae abundance is strongly reduced within a distance of 10 km around wind turbines.

If the ratio ”inside vs. outside” has been calculated separately for the before vs. the after period, two separate regression models have been fitted to the data, each applied to the data restricted to one of the periods. Especially, the variable *period* thus has been neglected, and B_dist_owp appeared only as a main effect (defining all data with $dist\_owp < 10\,km$ as inside). Importantly, here, all dependencies on $depth$, $dist\_coast$ and $Longitude/Latitude$ have been neglected, since they would interfere with (and thus bias) the measured effect.

In order to investigate distribution patterns and census estimates in the before- vs. after-period, $dist\_owp$-related variables have been excluded from regression models. Instead, the dependency on $depth$, $dist\_coast$ as well as on the spatial smooth were allowed to vary with the period, and *period* was introduced only as a main effect. Thus, distribution patterns and bird numbers were estimated separately for both periods completely independent of any information of wind turbine location, leading to the most objective estimation of abundance and distribution patterns.

Regression model structure

For final analyses, we formulate the BACI-approach as well as our population models (the latter for distribution pattern- and census-estimates) in the framework of modern regression methods, as suggested frequently in recent biostatistical literature [33, 37, 38, 17, 25]. Especially (and as motivated within the introduction-section), we make use of generalised additive models (GAM’s), which allow to adapt the analysis to various characteristics of our data, such as count-data and overdispersion (requiring generalised modelling [16, 18, 33, 37, 4]), offset-modelling (since counts have to be related to varying sizes of monitored area [17, 37]), and strongly nonlinear dependencies (requiring additive modelling [12, 9, 11, 32, 34]).

The ”most complex” BACI-GAM (which has not yet been thinned regarding its predictors as described in the following subsection) is given by

$$\begin{aligned} log(y_j) &= \beta + method_j + sea\_state_j \quad (1)\\ &+ s(depth_j) + s(dist\_coast_j)\\ &+ s(latitude_j, longitude_j)\\ &+ B\_dist\_owp + period_j\\ &+ B\_dist\_owp \times period_j\\ &+ \text{offset}(log(area_j)) + \epsilon_j, \end{aligned}$$

with $\epsilon_j \sim N(0, \sigma^2)$ i.i.d. Here, y_j is the vector of bird numbers, where the index j refers to the observation number. β is the intercept, $s(.)$ depicts a cubic regression spline, where the optimal number on knots has been estimated via generalised cross-validation. For each wind farm cluster separately, an appropriate probability distribution as well as an appropriate subset of predictors has been selected based on AIC analysis [2] (c.f., following subsection).

As motivated within the previous subsections, for distribution-pattern- and census analyses, a slightly modified version of the model has been used, namely:

$$\begin{aligned} log(y_j) &= \beta + method_j + sea\_state_j \quad (2)\\ &+ period_j + s(depth_j, by = period)\\ &+ s(dist\_coast_j, by = period)\\ &+ s(latitude_j, longitude_j, by = period)\\ &+ \text{offset}(log(area_j)) + \epsilon_j, \end{aligned}$$

where the term $by = period$ indicates that smooth terms are estimated independently for each period.

Finally, the ratio ”inside vs. outside” has been calculated using the following model

$$\begin{aligned} log(y_j) &= \beta + method_j + sea\_state_j \quad (3)\\ &+ B\_dist\_owp\\ &+ \text{offset}(log(area_j)) + \epsilon_j, \end{aligned}$$

separately applied to the data restricted to each period.

Model validation strategy

In order to obtain and validate the optimal GAM-model, we modified the selection and validation strategies as described e.g. by Ref. [36, 37, 34, 38, 17, 10]. Especially, for each wind farm cluster separately, this included the following steps:

1. Based on this ”maximal complex model” as given in the previous subsection, choosing an appropriate probability distribution / stochastic part of the model based on the Akaike Information Criterion (AIC) [2]. Namely we compared a Poisson-, negative binomial-, Tweedie- , and a zero-inflated Poisson-distribution among each other. All four probability distributions have been shown to describe the stochastic part in regression models of (p.r.n. overdispersed) count data reasonable well [8, 16, 15, 18, 30, 19, 38];

2. Using the favoured probability distribution, selecting an optimal subset of predictors (again based on the AIC). Especially, we permuted over all possible combinations and formulations of dispensable predictors, leading to the comparison of 16 different models;

3. based on the model with the favoured probability distribution and subset of predictors, performing model validation (mainly relied on graphical analysis via residual plots [36]) in order to test all required model assumptions. P.r.n., adding corresponding auto-correlation structures to the model.

Final measures of relative bird reduction / avoidance

In summary, the following measures of relative bird reduction / avoidance have been extracted from the regression models as introduced above:

- From the model based on Eq. (1), the relative change of bird abundance within a

10 km radius around wind turbines compared to outside the 10 km zone has been evaluated. Importantly, this ratio has been corrected for the ”natural ratio” (inside vs. outside) as given before construction of the turbines (”BACI-effect”);

- Based on the same principles, the relative change of bird abundance can be evaluated with rings/belts around the wind turbines (compared to the area outside these rings). The annulus width of these rings has been set to 3 km. Increasing stepwise the average distance of the rings from the wind turbines allows to investigate up to which distance the bird abundance is significantly reduced;
- From the model based on Eq. (2), the relative change of abundance inside 10 km vs. outside 10 km distance from wind turbines can be evaluated independently for each period (before vs. after construction). A direct comparison of these values shows on the one hand the reduction effect (respectively the measures underlying the BACI-effect). On the other hand, it gives an impression if wind turbines are placed within areas with relative high or low Gaviidae abundance values (compared to the surroundings).

Calculation of distribution maps and total bird numbers

In order to calculate distribution maps and total numbers of Gaviidae in the before vs. after-period, we used the model based on Eq. (3). Especially, we used the fitted model to predict bird densities on a prediction map of the investigated area with a resolution of 1 km^2 and including values for all environmental covariates. Regarding the detection-related variables *method* and *sea_state*, we investigated which *method-sea_state*-combination led to the highest predictions, and used them subsequently within the predict-routine. However, this implies the assumption that at least for one *method-sea_state*-combination, detected bird numbers (after distance-correction) are close to the real bird numbers, i.e. detection is close to 100 %.

Software

All statistical analysis, validation procedures and visualisations have been performed using the statistical software R [24]. Especially, we used the following packages: *sp* [23] and *gstat* [22] for the analysis of spatial auto-correlation (e.g. via variograms and bubble-plots); *ggplot2* [31] for all other visualisations and plots; the *Rmisc* [13] and *matrixStats* [3] for different functions regarding data analysis and utility operations, *MASS* [29], *pscl* [1], and *mgcv* [32] for regression analyses, *Distance* [27, 28, 7, 21] for distance sampling-related procedures, and *parallel* [24] for the use of parallel computing.

References

[1] Simon Jackman Achim Zeileis, Christian Kleiber. Regression models for count data in r. url http://www.jstatsoft.org/v27/i08/. *Journal of Statistical Software*, 27(8), 2008.

[2] H. Akaike. Information theory and an extension of the maximum likelihood principle. *International Sympossium on Information Theroy*, Second Edition:267–281, 1973.

[3] Henrik Bengtsson. matrixstats: Functions that apply to rows and columns of matrices (and to vectors). *R package version 0.51.0.*, (2016). 2016.

[4] Benjamin M. Bolker, Mollie E. Brooks, Connie J. Clark, Shane W. Geange, John R. Poulsen, M Henry H. Stevens, and Jada-Simone S. White. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends Ecol Evol*, 24(3):127–135, Mar 2009.

[5] D. Borcard, F. Gillet, and P. Legendre. *Numerical Ecology with R*. Springer, 2011.

[6] S.T. Buckland, D.R. Anderson, K.P. Burnham, J.L. Laake, D.L. Borchers, and L. Thomas. *Introduction to Distance Sampling: Estimating Abundance of Biological Populations*. Oxford University Press, New York, 2001.

[7] S.T. Buckland, E.A. Rexstad, T.A. Marques, and C.S. Oedekoven. *Distance Sampling: Methods and Applications*. Springer, 2015.

[8] S. Candy. Modelling catch snd effort data using generalized linear models , the tweedie distribution, random vessel effects and random stratum-by-year effects. *CCAMLR Science*, 11:59–80, 2004.

[9] R.M. Fewster, S.T. Buckland, G.M. Siriwardena, S.R. Baillie, and J.D. Wilson. Analysis of population trends for farmland birds using generalized additive models. *Ecology*, 81(7):1970–1984, 2000.

[10] A. Field, J. Miles, and Z. Field. *Discovering statistics using R*. SAGE Publications Ltd, 2012.

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

[12] T. Hastie and R.J. Tibshirani. *Generalized Additive Models*. London, UK: Chapman and Hall, 1990.

[13] Ryan M. Hope. Rmisc: Ryan miscellaneous. *R package version 1.5.*, 2013.

[14] M. Kery and J. A. Royle. *Applied Hierarchical Modeling in Ecology*. Elsevier, 2016.

[15] C. C. Kokonendji, S. Dossou-Gbete, and C. G.B. Demetrio. Some discrete exponential dispersion models: Poisson-tweedie and hinde-demetrio classes. *SORT*, 2:201–214, 2004.

[16] C.C. Kokonendji, C.G.B. Demetrio, and S. Dossou-Gbete. Overdispersion and poisson-tweedie exponential dispersion models. *Monographie del Seminaro Matematico Garcia de Galdeano*, 31:365–374, 2004.

[17] F. Korner-Nievergelt, T. Roth, S. von Felten, J. Guelat, B. Almasi, and P. Korner-Nievergelt. *Bayesian Data Analysis in Ecology Using Linear Models with R, BUGS, and Stan*. Elsevier, 2015.

[18] A. Linden and S. Maentyniemi. Using the negative binomial distribution to model overdispersion in ecological count data. *Ecology*, 92(7):1414–1421, Jul 2011.

[19] Tara G. Martin, Brendan A. Wintle, Jonathan R. Rhodes, Petra M. Kuhnert, Scott A. Field, Samantha J. Low-Choy, Andrew J. Tyre, and Hugh P. Possingham. Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. *Ecol Lett*, 8(11):1235–1246, Nov 2005.

[20] T.L. McDonald, W.P. Erickson, and L.L. McDonald. Analysis of count data from before-after control-impact studies. *JABES*, 5(3):262–279, 2000.

[21] D.L. Miller, M.L. Burt, E.A. Rexstad, and L. Thomas. Spatial models for distance sampling data: recent developments and future directions. *Methods in Ecology and Evolution*, 4:1001–1010, 2013.

[22] E.J. Pebesma. Multivariable geostatistics in s: the gstat package. *Computers & Geosciences*, 30:683–691, 2004.

[23] R.S. Bivand Pebesma, E.J. Classes and methods for spatial data in r. *R News*, 5 (2), 2005.

[24] R Core Team. R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria.*, ISBN 3-900051-07-0, URL http://www.R-project.org/, 2016.

[25] C.J. Schwarz. Analysis of BACI experiments. *In Course Notes for Beginning and Intermediate Statistics, Available at http://www.stat.sfu.ca/ cschwarz/CourseNotes.*, Chapter 12, 2014.

[26] E. P Smith. *BACI Design, Ecological statistics*. John Wiley & Sons, Ltd, 2002.

[27] L. Thomas, S. T. Buckland, K. P. Burnham, D. R. Anderson, J. L. Laake, D. L. Borchers, and S. Strindberg. *Distance Sampling*. John Wiley & Sons, Ltd, Chichester, 2002.

[28] Len Thomas, Stephen T. Buckland, Eric A. Rexstad, Jeff L. Laake, Samantha Strindberg, Sharon L. Hedley, Jon Rb Bishop, Tiago A. Marques, and Kenneth P. Burnham. Distance software: design and analysis of distance sampling surveys for estimating population size. *J Appl Ecol*, 47(1):5–14, Feb 2010.

[29] W. N. Venables and B. D. Ripley, editors. *Modern Applied Statistics with S. Fourth Edition*. Springer, New York, 2002.

[30] Seth J. Wenger and Mary C. Freeman. Estimating species occurrence, abundance, and detection probability using zero-inflated distributions. *Ecology*, 89(10):2953–2959, Oct 2008.

[31] H. Wickham. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009.

[32] Wood. *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC., 2006.

[33] A. Zuur, E. Ieno, and G.M. Smith. *Analysing Ecological Data*. Springer Science+Business Media, LLC, 2007.

[34] A. F. Zuur. *A beginner's guide to generalized additive models with R*. Highland Statistics Ltd., 2012.

[35] A. F. Zuur, E. N. Ieno, and A. A. Saveliev. *Spatial, Temporal and Spatial-Temporal Ecological Data Analysis with R-INLA*. Highland Statistics Ltd, 2017.

[36] A.F. Zuur, E.N. Ieno, and C.S. Elphick. A protocol for data exploration to avoid common statistical problems. *Methos in Ecology and Evolution*, 1:3–14, 2010.

[37] A.F. Zuur, E.N. Ieno, N.J. Walker, A.A. Saveliev, and G.M. Smith. *Mixed Effect Models and Extensions in Ecology with R*. Springer Science+Business Media, LLC, 2009.

[38] A.F. Zuur, A.A. Saveliev, and E.N. Ieno. *Zero inflated models and gerneralized linear mixed models withh R*. Highland Statistics Ltd., 2012.