Statistical Modelling of Habitat Selection

Shaykhah Aldossari[1], Jason Matthiopoulos[2]
and Dirk Husmeier[1]

[1] School of Mathematics & Statistics, University of Glasgow, Scotland, UK
[2] Institute of Biodiversity, Animal Health and Comparative Medicine, University of Glasgow, Scotland, UK

E-mail for correspondence: `s.aldossari.1@research.gla.ac.uk`

Abstract: To understand the impact of habitat destruction or modification on biodiversity there is increasing demand on predictive models that reliably forecast future changes in species distributions. In the present paper, we build on an existing model, the Generalized Functional Response, whose predictions about habitat preferences and species distribution are robust to changes in habitat availability. We improve upon this model in two distinct ways by using Gaussian mixtures to approximate habitat availability and Gaussian basis functions to describe habitat preferences. The proposed model is found to improve descriptive and predictive performance when applied to realistic simulated data and real species abundance data.

Keywords: Biodiversity, habitat selection function, basis functions, Gaussian mixture model

1 Introduction

The need to understand the ecological impact of land management, building construction and urban expansion on biodiversity is driving demand for new statistical models that can reliably forecast future changes in animal population distribution. Conventional approaches in ecological modelling aim to draw inferences about the importance and direction of the relationship between habitat preference $h(\mathbf{x})$ and environmental covariates $\mathbf{x} = (x_1, \ldots, x_I)$:

$$h(\mathbf{x}) \;=\; exp\left(\sum_{i=1}^{I} \beta_i x_i\right) \tag{1}$$

for fixed coefficients $\beta_i \in \mathbb{R}$. This can works well if the habitat availability does not change. However, Matthiopoulos et al. (2011) discuss the limita-

---

tions of this approach, and argue that it is essential to model the change in animal's habitat selection as well, by allowing the coefficients to vary as functions of habitat availability. Their derivation leads to the following expression for the habitat selection coefficients:

$$\beta_i \ = \ \mathbb{E}\Big(\gamma_i(\mathbf{x})\Big) + \varepsilon_i \ = \ \int \gamma_i(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} + \varepsilon_i \tag{2}$$

where $f(\mathbf{x})$ is a probability density function for habitat availability where each point $\mathbf{x}$ in environmental space represents a habitat, ε_i represents measurement or observation noise, and $\gamma_i(\mathbf{x})$ is a polynomial function in environmental covariate $\mathbf{x}$ which describes how the selection coefficient β adapts to changes in habitat availability $f(\mathbf{x})$ (introducing an integer order parameter M_j):

$$\gamma_i(\mathbf{x}) \ = \ \sum_{j=1}^{I} \sum_{m=0}^{M_j} \delta_{i,j}^{(m)} x_j^m \tag{3}$$

2 Methological Innovation

Matthiopoulos et al. (2011) demonstrate that modelling habitat selection in this way leads to a significant improvement in models of species distributions over the conventional model based on (1). However, the model still suffers from the following limitations: (1) The degree of nonlinear complexity and smoothness is restricted in advance: the functions have only M_j non-zero derivatives. (2) A complex function with a high degree of non-trivial differentiability requires a large number of parameters. (3) While the degree of smoothness is allowed to vary with respect to the choice of environmental variable, it is assumed to be global with respect to its entire range. (4) The expectation value in (2) is approximated by an empirical observed frequency, as the habitat availability is not explicitly modelled. The objective of the present paper is to propose a new statistical model that addresses these limitations. We start by replacing the polynomial with the following basis function approach:

$$\gamma_i(\mathbf{x}) \ = \ \sum_{j} \sum_{m=0}^{M_j} \delta_{i,j}^{(m)} \phi(x_j, \boldsymbol{\theta}_{j,m}) \ = \ \sum_{j} \sum_{m=0}^{M} \delta_{i,j}^{(m)} \phi(x_j, \boldsymbol{\theta}_{j,m}) \tag{4}$$

where ϕ is a basis function (e.g. splines, wavelets, basis functions of a reproducing kernel Hilbert space etc.) with parameters $\boldsymbol{\theta}_{j,m}$, chosen to represent known functional characteristics. Note that on the right-hand side we have simplified the notation by defining $M = \max\{M_j\}$, given that we have the freedom to set $\delta_{i,j} = 0$. Next, we follow Matthiopoulos et al. (2015) and model the probability distribution $f(\mathbf{x})$ with a Gaussian mixture model:

$$f(\mathbf{x}) \ = \ \sum_{k} \pi_k N(\mathbf{x}|\boldsymbol{\mu}_k, \mathbf{C}_k) \tag{5}$$

Inserting this into (2) and making use of (4) gives:

$$\begin{aligned}\beta_i &= \gamma_{i,0} + \int \gamma_i(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} + \varepsilon_i \\ &= \gamma_{i,0} + \int \left[\sum_j \sum_m \delta_{i,j}^{(m)} \phi(x_j, \boldsymbol{\theta}_{j,m}) \right] \left[\sum_k \pi_k N(\mathbf{x}|\boldsymbol{\mu}_k, \mathbf{C}_k) \right] d\mathbf{x} + \varepsilon_i \\ &= \gamma_{i,0} + \sum_j \sum_m \sum_k \delta_{i,j}^{(m)} \pi_k \left[\int \phi(x_j, \boldsymbol{\theta}_{j,m}) N(\mathbf{x}|\boldsymbol{\mu}_k, \mathbf{C}_k) d\mathbf{x} \right] + \varepsilon_i \quad (6)\end{aligned}$$

If we choose an RBF basis function for $\phi(x_j, \boldsymbol{\theta}_{j,m})$:

$$\phi(x_j, \boldsymbol{\theta}_{j,m}) = \exp\left(-\frac{1}{2} \frac{(x_j - \xi_{j,m})^2}{\sigma_{j,m}^2} \right) \quad (7)$$

with parameter vector $\boldsymbol{\theta}_{j,m} = (\xi_{j,m}, \sigma_{j,m})$, then the integral

$$\psi(\boldsymbol{\theta}_{j,m}, \boldsymbol{\mu}_k, \mathbf{C}_k) = \int \phi(x_j, \boldsymbol{\theta}_{j,m}) N(\mathbf{x}|\boldsymbol{\mu}_k, \mathbf{C}_k) d\mathbf{x} \quad (8)$$

has a closed-form solution (see e.g. Bishop, Section 2.3) and we get:

$$\beta_i = \gamma_{i,0} + \sum_j \sum_m \sum_k \delta_{i,j}^{(m)} \pi_k \psi(\boldsymbol{\theta}_{j,m}, \boldsymbol{\mu}_k, \mathbf{C}_k) + \varepsilon_i \quad (9)$$

TABLE 1. Comparison of the proposed method with the approach of Matthiopoulos et al. (2011) on simulated habitat data. Smaller values indicate better performance.

Method	AIC	BIC	RMSE
Matthiopoulos et al. (2011)	907217.4	907860.4	3.97
Method proposed here	907206	907818.2	3.94

TABLE 2. Comparison of the proposed method with the approach of Matthiopoulos et al. (2011) on the sparrow population data from Matthiopoulos et al. (2018).

Method	AIC	BIC	RMSE
Matthiopoulos (2011)	1696.024	1902.209	23.52
Method proposed here	1683.656	1889.841	11.44

3 Empirical Evaluation

We evaluate the performance of the proposed model on the simulated data described in Matthiopoulos et al. (2011). The simulation was an individual-based model of the dependence of species abundance on two habitat variables: food and cover (the converse of predation risk).

We set the polynomial order for the model in Matthiopoulos et al. (2011) and the number of basis functions in the model proposed here equal to 10 based on model selection scores. We evaluate the predictive performance in terms of root mean square error (RMSE) on out-of sample test data that have not been used for parameter estimation, and compare the accuracy of the model proposed in Matthiopoulos et al. (2011) with our proposed model. The results are shown in Table 1 and suggest that a noticable improvement in terms of model selection scores (AIC, BIC) and out-of-sample (RMSE) over the state-of-the-art Generalized Function Response (GFR) model can be achieved.

4 Real-World Application

We have applied our model to the sparrow population data described in Matthiopoulos et al. (2018). This habitat use model consists of three habitat variables (the percentage of grass, bush and roof in each cell). The best polynomial order for the model in Matthiopoulos et al. (2011) and the number of basis functions in the model proposed here is 3 based on the model selection scores. The results of model selection scores and the evaluation of predictive performance on out-of-sample test data are shown in Table 2. Our model outperforms the GFR in Matthiopoulos et al. (2011), with an improvement of the model selection scores and out-of-sample RMSE.

5 Conclusions

We have modelled habitat preference with a flexible approach that extends the model proposed in Matthiopoulos et al. (2011) in two distinct ways, by using Gaussian mixtures to approximate habitat availability and Gaussian basis functions to describe habitat preferences. We have tested the new model on both simulated data and real survey data, using the sparrow population data from Matthiopoulos et al. (2018). Our results suggest that a noticeable improvement can be obtained in terms of AIC, BIC and RMSE.

References

Bishop (2006). Pattern Recognition and Machine Learning. *Springer*.

Matthiopoulos, Hebblewhite, Aarts and Fieberg (2011). Generalized functional responses for species distributions. *Ecology*, **92**, 583–589.

Matthiopoulos et al. (2015). Establishing the link between habitat selection and animal population dynamics *Ecol. Monographs*, **85**, 413–436.

Matthiopoulos, Field and MacLeod (2018). Predicting population change from models based on habitat availability and utilization. *Proceedings of the Royal Society B*, **286**, 20182911.