Proceedings of the 3rd International Conference on Statistics: Theory and Applications (ICSTA'21)
~~Prague, Czech Republic~~ Virtual Conference – July, 2021
Paper No. 125
DOI: 10.11159/icsta21.125

Generalized Functional Responses in Habitat Selection Fitted by Decision Trees and Random Forests

Shaykhah Aldossari[1], Dirk Husmeier[1], Jason Matthiopoulos[2]
[1] School of Mathematics and Statistics, University of Glasgow
Glasgow G12 8SQ, UK
s.aldossari.1@research.gla.ac.uk; dirk.husmeier@glasgow.ac.uk
[2] Institute of Biodiversity, Animal Health and Comparative Medicine, University of Glasgow
Glasgow G12 8SQ, UK
jason.matthiopoulos@glasgow.ac.uk

Abstract – Species Distribution Models (SDMs) are important regression tools in the ecological sciences that can support distribution predictions using different environmental variables. Most of the research in the area of SDMs has assumed that regression coefficients in these models are fixed. However, species respond differently to different habitats depending on the habitat availability, meaning that regression coefficients change as functions of habitat availability, a phenomenon known as a functional response in habitat selection. The generalized functional response (GFR) is a varying-coefficient extension of the basic SDM framework, designed for more robust forecasts of species distributions in a rapidly changing world. The original GFR model formulated the varying regression coefficients using a polynomial function approach, which led to improvements of forecasting performance in many applications. The purpose of this paper is to improve the out-of-sample performance of the GFR model using a decision tree and Breiman's random forest algorithm. We compare the original GFR model with a decision tree and random forests using the GFR model by applying both models to a real population dataset on house sparrows. The results revealed a noticeable improvement in terms of out-of-sample R^2 in the decision tree and the random forest approaches over the original GFR model.

***Keywords*:** Decision tree, generalized function response, habitat selection function, random forest, space use model

1. Introduction

Species distributions in space is a topic of considerable importance in ecological research, particularly in terms of the environmental drivers of species distribution [1]. Species Distribution Models (SDMs) aim to connect animal behaviour with availability of habitat resources and estimate habitat usage by comparing different samples taken from different locations, where each sample has different environmental features [2]. Predicting habitat usage using standard regression-based SDMs is very challenging particularly for animals [3] because the same organisms behave differently depending on the environmental context they find themselves in. This non-linear phenomenon is known as a functional response in habitat selection [4] and can be modelled by writing the coefficients of the SDMs as a function of the availability of all habitats within reach of the organisms [5] [6]. The generalized functional response (GFR) approach [7] was developed as a broad class of varying-coefficient models, formulated originally as polynomial functions of habitat availability. The GFR model has been shown to display better out-of-sample predictions than the fixed coefficient SDM approaches. However, as seen from Eq. (4) below, the GFR of [7] is based on a global polynomial of fixed maximum order. The limitations of such models for learning non-linear functions from data have been widely discussed in the Machine Learning and Statistics literature (see e.g. [11]). The aim of the present paper is to try a more advanced and flexible modelling approach based on classification and regression trees (CART) and random forests (RFs), and to quantify the improvement in out-of-sample performance that can be achieved on a real-world application related to habitat usage in a sparrow population.

2. Methods

The preference, $h(x)$, of a certain habitat x by a given species is the ratio of the probability density of usage $f_u(x)$ of that habitat, over the availability of the same habitat, $f_a(x)$ [1] (also defined as a PDF over the space of environmental dimensions):

$$h(x) = \frac{f_u(x)}{f_a(x)} \tag{1}$$

Estimating the significance and direction of the relationship between habitat preference and environmental covariates often forms the goal of SDM analyses. Habitat preference is routinely represented as an exponential transformation of a predictor function of environmental covariates $x = (x_1, \ldots, x_I)$ [7]:

$$h(x) = \exp(\sum_{i=1}^{I} \beta_i x_i) \tag{2}$$

where the $\beta_i's$ are fixed coefficients of the habitat selection model. Habitat selection models that are predominantly formalized in this way are unable to capture species distribution when habitat availability changes. There is no biological reason that the proportionality assumption between use and availability in Eq. (1) should remain when the availability of other habitat options changes. Consequently, there is no reason to expect that the beta coefficients estimated via a regression approach based on Eq. (2) should be fixed when broad availability profiles change [7]. In the GFR model, Matthiopoulos et al. [7] allow the $\beta_i's$ to vary as functions of habitat availability, as follows [7]:

$$\beta_i = \int \gamma_i(x) f_a(x) dx + \varepsilon_i \tag{3}$$

where ε_i is measurement noise and $\gamma_i(x)$ is a flexible real function of habitat variable x. Matthiopoulos et al. use a polynomial function to represent the $\gamma_i(x)$ for each environmental variable [7]. Their derivation results in the following relation of the β_i 's:

$$\beta_{i,b} = \gamma_{i,0} + \sum_{j=1}^{I} \sum_{m=0}^{M_j} \delta_{i,j}^{(m)} \left[X_j^m\right]_b + \varepsilon_{i,b} \tag{4}$$

where M_j is an integer order parameter, $\left[X_j^m\right]_b$ is the mth moment of the covariate j in the bth sampling instance where each sample instance could be a sample taken in different years for the same population or sample taken from different sub-distributions and $\delta_{i,j}^{(m)}$ is the coefficient of $\gamma_i(x)$ for the power m of the jth variable.

Classification and regression trees (CART) are widely applied for modelling nonlinear functions learned from data. The idea of the CART algorithm is building a tree, which consists of nodes and leaves where each node is a variable that can be split into two leaves or parts [9]. The node x_i and the threshold or split value h , which is a value from x_i, is chosen based on some criteria. We use Breiman's criteria in [10] to choose the best variable in each node and the best split of that variable. A locally optimal maximum likelihood estimator is used for the split function as follows:

$$(j^*, h^*) = \underset{j \in \{1,\ldots,D\}}{\min} \underset{h \in h_j}{\min} \; cost(\{x_i, y_i : x_{ij} \leq h\}) + cost(\{x_i, y_i : x_{ij} > h\}) \tag{5}$$

where j^* is the best feature to split , h^* is the best split of that feature, y_i is the response variable and D is the data in each leaf. Misclassification rate, entropy and Gini index are common classification costs. The classification cost using Gini index is the probability of a randomly chosen element to be classified as the incorrect class if it was randomly classified based on the distribution of classes in the data D. In a piecewise constant function case, the regression cost for a given node is the residual error after fitting the model in each leaf using the variables in the path from the root to that node as follows [9]:

$$cost\ (D) = \sum_{i \epsilon D} (y_i - \bar{y}\ )^2 \tag{6}$$

where $\bar{y}$ is the response variable mean of D, which is the data assigned to the corresponding leaf node. We use a pruning scheme, whereby the minimum cross-validation cost, based on a 10-fold cross-validation scheme on the training data (but excluding the test data) determines the best number of terminal nodes [9]. We use the CART model in combination with the GFR model to quantify the out-of-sample forecasting performance. The idea is to fit the GFR model in each leaf of a training data by using the best feature and the best split, which are the explanatory variable and the threshold with the lowest cost,

and use the cost to choose the variable and the split value in each node of the tree, then do10-fold cross-validation in the training data to prune the tree. To measure the forecasting performance of the CART model using the GFR model, we predicted the target variable in a different test set.

Matthiopoulos et al. make use of Eq. (4) in Eq. (2) to estimate the relationship between habitat preference and environmental covariates using OLS by minimizing residual sum of squares, which means that the cost function of the CART model using the GFR model in each leaf is:

$$cost_{GFR}(D) = \sum_{i\epsilon D}(y_i - \bar{y}_{GFR})^2 \quad (7)$$

where y_i is the value of the variable to be predicted in D, which is the data assigned to the corresponding leaf node, and $\bar{y}_{GFR}$ is prediction generated by using the new $\beta_i{}'s$ as specified in Eq. (4). For instance, in the root node and the leaf node, we fit the GFR model using all explanatory variables and the possible thresholds of the variables using OLS approach in Eq. (7) to estimate the GFR coefficients and start the tree using the variable and split with the lowest cost. Random forests (RFs) are widely used for regression and classification to improve the out-of-sample generalization performance. An RF is a type of ensemble method containing multiple trees where each tree has a random sample of features. For each tree in RF, 63.2% of the observations are chosen randomly with replacement from the dataset to apply the GFR model [11]. The explanatory variables are also selected randomly for each tree where the number of explanatory variable in each tree in a classification tree is equal to the square root of the total number of explanatory variable in the dataset, $\sqrt{p}$ where p is the total number of the explanatory variables in the dataset, and for regressions, the number of explanatory variables in each tree is the total number of the variables in the dataset divided by 3, $p/3$ [11]. We apply the RF using the GFR model in each node to improve the forecasting performance for out-of-sample data. We use the method in the training data and the prediction in the test data to measure the performance of the models using the sparrow population dataset. We use out-of-sample R^2 to measure how well the model predicts the response variable from the training data for new observations in the testing data.

3. Materials

We used the same sparrow population data that was used in [8] to predict population change based on habitat availability using the GFR model. The set contains data for 32 colonies in the region of Glasgow, each colony containing instantaneous abundance data for 40 spatial cells. The sparrow response variable was analysed by a use-availability approach [8]. We use three main variables in the dataset: the estimated percentage of grass, bush and roof for each cell as captured by Google earth. To increase the ability to account for the variability in both models, an additional explanatory variable that represents values applying uniformly to an entire sampling instance was included in both models. The size of each colony was determined as the maximum number of males counted in each colony and was included in both models because larger colonies behave differently and use space differently.

4. Results

Using the sparrow population dataset, we fit the GFR model using different polynomial orders from the 1st order to the 10th. We chose the 3rd order to fit the polynomial model based on the model selection score: AIC and BIC. We applied the GFR, the CART and the RF models to the sparrow population dataset. The out-of-sample forecast performance was used for evaluation. We did 32-fold cross validation, leave-one-colony-out, to calculate the out-of-sample R^2 score since the sparrow population data covers 32 colonies. Each time we fit the GFR or RF model to 31 colonies, we predict the habitat use of sparrows in each cell for the missing colony to calculate R^2. To get the results from GFR model, we fit the third polynomial GFR model to the training data, then predict on the test data. We fit the CART model to the data by using the GFR model in each leaf with the best features and split in the path. The feature in the sparrow dataset are the variables grass, bush , roof, colony size, the moments of the variables and the mixed terms, which are the combinations of the variables with the moments. This follows Eq. (8) in [7]. For the RF model, we fit the GFR model in each leaf of 500 trees using the training set to predict the habitat use of sparrows in each cell for all the colonies. The mean out-of-sample R^2 score over the 32 colonies is shown in Table 1. We also include the median R^2 score, which is more robust to outliers than the mean. Table 1 suggests that the CART and the RF methods using the GFR model outperform the original GFR in [7], with a noticeable improvement for out-of-sample R^2.

Table 1: Mean and median of out-of-sample performance scores of 32 colonies for GFR, CART and RF models. The GFR model was used in each leaf in the tree of CART and RF.

Model	Mean of R^2's	Median of R^2's
GFR	0.160	0.250
CART	0.255	0.545
RF	0.420	0.765

5. Conclusion

The generalized functional response model is a useful way of producing out-of-sample predictions for habitat preference by species. The GFR model uses a polynomial function to model the habitat selection function. The mth moment of the habitat covariate in a certain sampling instance is used to represent the habitat selection coefficients β's after allowing the $\beta_i's$ to vary as functions of habitat availability instead of being fixed. The GFR model is a non-linear model because of the many interactions it has and we hypothesised that this level of non-linearity will be dealt with by the CART approach. A decision tree (CART) is an algorithm used to predict values in a target variable based on decision rules obtained from training data. We used the CART model to fit the GFR model to each leaf. The GFR and CART models use the OLS approach to estimate the β's. The RF model is an ensemble approach used for regression and classification which contains a multitude of trees to aggregate the results obtained from each tree. RFs have been shown to improve the out-of-sample performance over single trees [9]. We have compared the forecasting performance of the GFR, CART and RF models using the out-of-sample performance score. A real-world application was used to measure the performance of the models. We used the sparrow population dataset that was used in [8] and applied the GFR model. We used leave-one-colony-out cross validation to make predictions for each colony in the GFR model. For RF, we applied the GFR model in each node of 500 trees. Based on the findings in Table 1, the CART and the RF models using the GFR model are more effective at making predictions than the original GFR based on the mean and the median of the out-of-sample performance scores for each colony.

References

[1] J. Matthiopoulos, J. Fieberg and G. Aarts, Species-Habitat Associations: Spatial data, predictive models, and ecological insights, 2020.

[2] D. H. Johnson, "The comparison of usage and availability measurements for evaluating resource preference," Ecology, vol. 61, no. 1, pp. 65--71, 1980.

[3] H.L. Beyer, D.T. Haydon, J.M. Morales, J.L. Frair, M. Hebblewhite, M. Mitchell and J. Matthiopoulos, "The interpretation of habitat preference metrics under use--availability designs," Philosophical Transactions of the Royal Society B: Biological Sciences, vol. 365, no. 1550, pp. 2245--2254, 2010.

[4] A. Mysterud and R.A. Ims, "Functional responses in habitat use: Availability influences relative use in trade-off situations," Ecology, vol. 79, no. 4, pp. 1435--1441, 1998.

[5] M.S. Boyce and L.L. McDonald, "Relating populations to habitats using resource selection functions," Trends in ecology & evolution, vol. 14, no. 7, pp. 268--272, 1999.

[6] M. Mauritzen, S.E. Belikov, A.N. Boltunov, A.E. Derocher, E. Hansen, R.A. Ims, Ø. Wiig and N. Yoccoz, "Functional responses in polar bear habitat selection," Oikos, vol. 100, no. 1, pp. 112--124, 2003.

[7] J. Matthiopoulos, M. Hebblewhite, G. Aarts, and J. Fieberg, "Generalized functional responses for species distributions," Ecology, vol. 92, no. 3, pp. 583--589, 2011.

[8] J. Matthiopoulos, C. Field and R. MacLeod, "Predicting population change from models based on habitat availability and utilization," Proceedings of the Royal Society B, vol. 286, no. 1901, p. 20182911, 2019.

[9] K. P. Murphy, Machine learning: a probabilistic perspective, MIT press, 2012.

[10] L. Breiman, " Random forests," Machine learning, vol. 45, no. 1, pp. 5--32, 2001.

[11] S. RColorBrewer and M.A Liaw, "Package 'randomForest'," University of California, Berkeley: CA, USA, 2018.