Transferable species distribution modelling: Comparative performance evaluation and interpretation of novel Generalized Functional Response models

Shaykhah Aldossari

Submitted in fulfilment of the requirements for the
Degree of Doctor of Philosophy

School of Mathematics & Statistics
University of Glasgow

December 2022

Abstract

Predictive species distribution models (SDMs) are becoming increasingly important in ecology, in the light of rapid environmental change. The predictions of most current SDMs are specific to the habitat composition of the environments in which such models were fitted. However, species respond differently to a given habitat depending on the availability of all habitats in their environment, a phenomenon known as a functional response in resource selection. The Generalised Functional Response (GFR) framework captures this dependence by formulating the SDM coefficients as functions of habitat availability in the broader environment. The original GFR implementation used global polynomial functions of habitat availability to describe functional responses. In the present thesis, I develop several refinements of this approach and compare their explanatory and predictive performance using two simulated and three real datasets.

I use local radial basis functions (RBF), a more flexible approach than global polynomials, to represent the habitat selection coefficients and regularization to balance bias and variance and prevent over-fitting. Second, I use the RBF-GFR and GFR models in combination with the classification and regression tree (CART), which has more flexibility and better predictive powers for non-linear modelling. As further extensions, I use random forests (RF) and extreme gradient boosting (XGBoost) ensemble approaches that consistently lead to variance reduction in generalization error.

After applying the original and extended models to four different datasets, I find that the different methods perform consistently across the datasets, such that their approximate ranking for out-of-data prediction is preserved. The traditional stationary approach to SDMs, excluding the GFR model, consistently performs at the bottom of the ranking. The best methods in my list provide non-negligible improvements in predictive performance,

in some cases taking the out-of-sample R^2 score from 0.3 up to 0.7 across datasets.

At times of rapid environmental change and spatial non-stationarity ignoring the effects of functional responses on SDMs, results in two different types of prediction bias (under-prediction or mis-positioning of distribution hotspots). However, not all functional response models are created equal. The more volatile GFR models may fall foul of similar biases. My results indicate that there are consistently robust GFR approaches that achieve transferability consistently across very different datasets.

In addition to these improvements in predictive performance resulting from the GFR, RBF-GFR and their extensions, it is also essential to know whether these models can offer insights into the mechanisms mediating species distributions. I use one of the simulated datasets to interpret two of the models that provide the best predictive power for this dataset. The resulting selection coefficients from the two models are similar, which explains why the two models are able to explain the observed data in similar ways. In addition, the behaviour of the availability-filtered selectivity coefficients is consistent with the known mechanisms generating the data. These findings indicate that despite their purely statistical nature these fundamentally different models show convergent and realistic behaviour.

To test the transferability of the improved versions of the GFR model in a large-scale and multi-species dataset, I use the challenging large-scale North American Breeding Bird Survey BBS dataset. I discuss how the information in the dataset affects the predictive ability of each species abundance. My recent extensions of the GFR model double the biodiversity prediction accuracy compared to the standard generalised linear model (GLM) and the original GFR model.

Contents

List of Tables

List of Figures

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisors, Prof. Dirk Husmeier and Prof. Jason Matthiopoulos, for their scientific guidance, support, kindness, patience, invaluable suggestions and encouragement throughout my PhD. Without their input and help over the years, I doubt I would ever have been able to undertake this work. I genuinely could not have hoped for better supervisors.

I would like to thank the Ministry of Higher Education and King Faisal University for funding my study. I would also like to thank the Saudi Arabian Cultural Bureau SACB for their help and support throughout the whole process.

A huge thank goes to my honourable parents, Wasmyah and Abdullah, for their love, prayers, support and encouragement. I would like to thank my brothers and sisters for being a part of my journey, their support and wishes were unlimited. A special thanks to my brother Mesfer, who travelled with me and shared my PhD journey, for his unconditional support and encouragement.

I would also like to extend my thanks to my friends Bashayr, Riham, Shuhrah, and Hanadi for their love, help, support and kindness. Thanks for making the past four years much more enjoyable and keeping me sane throughout my study.

Declaration

I hereby declare that this thesis has been written by me under the supervision of Prof. Dirk Husmeier and Prof. Jason Matthiopoulos, and has not been submitted previously as part of any application for a degree. I have acknowledged all sources used and have cited these in the bibliography section.

Part of the work in Chapters 2 and 3 is published in Proceedings of the 35th International Workshop on Statistical Modelling with the title "Statistical Modelling of Habitat Selection" in 2020 with ISBN 9788413192673 (Aldossari et al., 2020).

I presented some parts of Chapters 3 and 5 at the 3rd International Conference on Statistics: Theory and Applications (ICSTA'21) in 2021, which was published and received the Best Paper Award in proceedings of the conference under the title "Generalized functional responses in habitat selection fitted by decision trees and random forests" with ISBN 9781927877913 (Aldossari et al., 2021).

The work presented in Chapters 3, 4 and 5 has been published in the Ecological Informatics journal titled "Transferable species distribution modelling: Comparative performance of Generalised Functional Response models", volume: 71, pages: 101803, year: 2022, publisher: Elsevier (Aldossari et al., 2022).

I will present Chapter 7 at the 15th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics 2022), which King's College London will host on 17-19 December 2022.

Chapter 1

Introduction

The need to understand the relationship between organisms and their physical environment drives demand for new statistical models that can reliably forecast future changes in species distributions, in response to changes in habitat availability. A *habitat* is a specific point in environmental space (Paton and Matthiopoulos, 2016), the combination of particular values in different environmental dimensions. *Habitat use* is the proportion of time that an individual, population or species spend at a particular habitat, and *habitat selection* is the behavioural process by which an organism chooses its habitat, which results in habitat use and influences the species' survival (Block and Brennan, 1993). Habitats are not used randomly, and therefore habitat use can differ from habitat avaiability. *Habitat preference* refers to the disproportionality between the use of a habitat and its availability in the environment (Aarts et al., 2008). The models used to model habitat preferences by using a quantitative comparison between habitat use and availability are *habitat models* (Paton and Matthiopoulos, 2016).

As the complexity of questions related to conservation and ecosystem management begins to outstrip our ability to collect detailed spatial and temporal data (Fordham et al., 2016; Kindsvater et al., 2018), we have come to rely on more sophisticated statistical methodologies for interpolating between locations, times and taxonomic groups and for predicting into the future. Predictive models of species distributions, in particular, play an increasingly important role as organisms respond to accelerating changes in climate and land use (Evans et al., 2012; Houlahan et al., 2017; Maris et al., 2018; Mouquet et al.,

2015; Sequeira et al., 2018; Travers et al., 2019; Yates et al., 2018).

The demand for transferable models (i.e., models that can predict accurately in environments very different to those used for model fitting - Yates et al., 2018) has led to the realisation that statistical species distribution models (SDMs) are currently not fit-for-purpose, particularly in the case of animal species (Austin, 2002; Bahn and McGill, 2013; Barbet-Massin et al., 2018; Barbosa et al., 2009; Dormann, 2007; Ehrlén and Morris, 2015; Randin et al., 2006; Tessarolo et al., 2021; Torres et al., 2015; Zurell et al., 2009).

A key challenge with the transferability of SDMs is that species, particularly animals, respond differently to a particular habitat depending on the availability of other habitats in their environment (Boyce and McDonald, 1999; Mysterud and Ims, 1999). Fig. 1.1, which has been taken from Matthiopoulos et al. (2011), illustrated this problem when a generalized linear model was applied using a simple animal in a particular environment. The model fits well in the same environment (comparing Fig. 1.1 c and d) but provides poor predictions of habitat use when placed in a different environment (comparing Fig. 1.1 g and h). This process, termed a functional response in habitat selection, is the result of complex mechanistic interactions between habitat availability and animal behaviour. These two factors interact with each other to produce a response, rather than either influence acting on its own (Matthiopoulos et al., 2011; Mauritzen et al., 2003; Mysterud and Ims, 1998). This is difficult to capture with standard statistical models because the estimated parameters of SDMs are specific to the environmental settings where these models were fitted. The consequence of functional responses is that unless the environmental context is explicitly taken into account, spatial predictions can be increasingly inaccurate as the prediction settings diverge from the model fitting environmental profiles (Paton and Matthiopoulos, 2016). The standard species distribution models do not take into account the effects of habitat availability functions, where these models give poor predictions if they are used in extrapolation scenarios (i.e. prediction in different environments). This process is not new in statistics, but it appears to have been somewhat unclear in the ecology literature. Functional responses in habitat selection are detectable in real datasets. Fig. 1.2 shows how moose in Norway used nine different habitat types based on relative habitat availability (Fig. 6 in Bjørneraas et al. (2012)). Two species of birds show positive functional responses to three treatments of habitat (see Fig. 2 and Table 3 in Gillies and

St. Clair (2010)). Functional responses to pastures were detected in a telemetry dataset containing 62 red deer in Norway (Godvik et al., 2009). When pasture was rare, the selection of pastures was increased, but the selection of pasture decreased with increasing pastures availability (Godvik et al., 2009). Functional responses are detectable by many different methods, but the exact nature of the response depends on the statistical methodology that is used to capture it. For example, two species of conservation concern, Canada lynx in the United States and woodland caribou in Canada, were used to evaluate four different functional response approaches (Holbrook et al., 2019). Habitat use in the additive scale, habitat use model in the multiplicative scale, habitat selection with resource selection function RSF, and habitat selection with the interaction of RSF. There was a variation among these approaches with regard to evaluating the functional response (See Fig. 3 in Holbrook et al. (2019)). Some approaches show increases in habitat use by Canada lynx with increasing advanced regenerating forest availability while other approaches show the opposite effect. Some approaches demonstrated no functional response. The same variation in results occurred when testing woodland caribou habitat use in response to linear features. Differences resulting from different implementations highlight the importance of investigating the robustness of functional response models, my initial objective in this thesis.

The functional response model is a model in which the model coefficients are functions and the response is a scalar. This model is a functional extension of linear regression and a type of functional data analysis, where such models are usually called generalised regression of scalars on functions (For more details, see Chapter 15 in Ramsay and Silverman, 2005). Different approaches have been proposed to model functional responses in habitat selection, ranging from single-habitat models of usage as a function of availability (Mysterud and Ims, 1998) to writing SDM coefficients as functions of the availability of all habitats (Boyce and McDonald, 1999). The need to account for functional responses is clearly demonstrated by the efficacy of approaches that do not use any model of functional response but simply recognise the distinction between different environmental scenarios by means of random effects (Gillies et al., 2006). The generalized functional response GFR approach (Matthiopoulos et al., 2011) uses a function of availability to represent the SDM's coefficients. The coefficients of the GFR are modeled by functions of lo-

cal habitat availability using a polynomial function approach (Matthiopoulos et al., 2011; Matthiopoulos et al., 2019). The GFR model is ultimately structured using the local value of the habitat covariates, moments (e.g., the means) from the distribution of the habitat covariates, and the pairwise interactions between these terms (Matthiopoulos et al., 2011). The GFR is an example of a varying-coefficient model, an extension of the generalized linear model with coefficients written as functions of other variables (Hastie and Tibshirani, 1993). For example, Fig. 1.3 shows a similar simulated process of Fig. 1.1, where an animal whose priorities are feeding and hiding is observed in a habitat where a simple SDM is fitted and provided a good fit in the same habitat (comparing Fig. 1.3 b and c) and poor predictions when placed in a different habitat (comparing Fig. 1.3 a and f). Fig. 1.4 shows that the predictions from the GFR model are in good agreement with the ground truth of the second environment used in Fig. 1.3.

The approach taken in Matthiopoulos et al. (2011) was to model each of the SDM coefficients as a global polynomial, motivated by the fact that under fairly general regularity conditions, any smooth function can be approximated by a Taylor series. The practical problem, however, is that this power series expansion with its polynomial coefficients has to be learned from data. Taking a high polynomial order leads - for limited and noisy data - to potential over-fitting (and poor transferability). Standard approaches, therefore, aim to find the adequate degree of model complexity, e.g., via cross-validation or based on information criteria, such that for small datasets and high noise levels, less complex models are preferred.

However, for a global polynomial function, controlling model complexity e.g., by restricting the number of adjustable model parameters, implies a truncation of the polynomial order and a limitation of the degree of non-trivial differentiability. This is methodologically inconsistent: the highest polynomial order and the degree of non-trivial differentiability are an intrinsic feature of the systems under investigation and must not be dictated by the quantity and quality of the available data. The first aim of the thesis is to build on the GFR approach proposed by Matthiopoulos et al. (2011), by replacing the global polynomial expansion by several more recent methods from multivariate statistics and machine learning such that the logical inconsistency outlined above is avoided.

To address the limitations of the original GFR model proposed in Matthiopoulos et al.

(2011), I adapt three state-of-the-art flexible regression paradigms to model the habitat selection coefficients. The first approach is based on a radial basis function (RBF) expansion, as e.g., reviewed in Chapter 5 of Bishop, 1995, and I refer to this model as the RBF-GFR model. I ensure that the RBF-GFR model flexibility is not deployed indiscriminately by using regularization to limit that flexibility so that, under out-of-sample predictions, the model can behave well. I refer to this model as the regularized RBF-GFR model. Next, I combine classification and regression trees (CART), reviewed e.g., in Chapter 9 of Hastie et al., 2008 or Section 16.2 in Murphy, 2012, with both the original GFR model and RBF-GFR model. I refer to these models as GFR-CART and RBF-GFR-CART, respectively. I finally create model ensembles, based on random forests (RFs) trained with bagging (see Chapter 15 in Hastie et al., 2008) or boosting (see e.g., Chapter 16 in Hastie et al., 2008 or Section 16.4 in Murphy, 2012). I refer to these ensembles with the suffix "RF" or "XGBoost".

I assess these models on two different levels. First, I explore the out-of-sample spatial predictions by looking at scenario-specific plots of predicted usage over geographical space for four small-scale and single-species datasets: two individual-based simulated datasets and two real-life applications. The two simulated datasets are species abundance levels, whereas the real applications are binary species use/availability datasets. I compare the test set accuracies, which have been quantified in terms of out-of-sample R^2 scores, and split the presentation of the results by dataset. Looking at the predictive level is an insufficient assessment, and an ecological realism assessment is needed to explore species abundance models (Austin 2007). Thus, looking at space- and scenario-independent plots is the second level of model assessment I use. I look for the spurious effects in the graphical diagnostic selectivity coefficients γ_i and try to use these coefficients to see if these models biologically make sense. It is important to offer some explanatory power of the mechanisms mediating species distributions and know whether these models allow us to look beyond the predictions by visualising the changes in the regression coefficients. I use one of the simulated datasets for this assessment because the mechanisms generating this dataset are known. Hence, the models can be assessed by their ability to infer these mechanisms. The mechanisms generating data can be explored by looking at the patterns illustrated in images, where these patterns result from the statistical model being fitted to

data.

It is better to use larger scale and multiple species datasets to apply varying coefficient models because prevailing conditions are likely to vary a lot across a large map. After looking at how the models behave under relatively small scale or simulated datasets, it is essential to test those models in large-scale conditions. I use the large-scale North American Breeding Bird Survey (BBS) dataset to predict the abundance of ten different species in the data by applying the GFR model and various recent extensions using land-cover types and the temperature of each segment as covariates. Furthermore, I use the Shannon entropy score, the most frequently used measure of biodiversity in ecology, to investigate three different things. First, I use the entropy score to assess the transferability of the generalized function response (GFR) model and its extensions by measuring the information content in the dataset. Second, since biodiversity is widely used to describe the variation and there are emergent trends in how biodiversity increases or declines with ecological context, I observe the relationship between biodiversity and land cover types using the GFR model and various recent extensions. Finally, I quantify the legacy effect arising from extinction debts and colonisation credits (Haddou et al., 2022), using the GFR models of land cover types on biodiversity.

The present thesis is structured as follows: I review the methodology, statistical background and related literature that are used throughout this thesis in Chapter 2. Details of my new models and outlines of the relationship between all the models proposed in the present study can be found in Chapter 3. Chapter 4 provides an overview of two simulated and three real-world datasets used for a comparative evaluation of the various methods. The results of the predictive evaluation of four datasets are presented in Chapter 5. Explanatory quantification and some ecological interpretations of selectivity coefficients of one of the simulated datasets can be found in Chapter 6. Chapter 7 provides discussions of individual species distribution predictions in the large-scale North American Breeding Bird Survey (BBS) dataset, the relationship between biodiversity and land cover types using the GFR models, and legacy effects on biodiversity. I finish with my general dissection and conclusions in Chapter 8. To keep the main text sufficiently concise, I have relegated some methodological details along with more comprehensive simulation results to Appendix A.

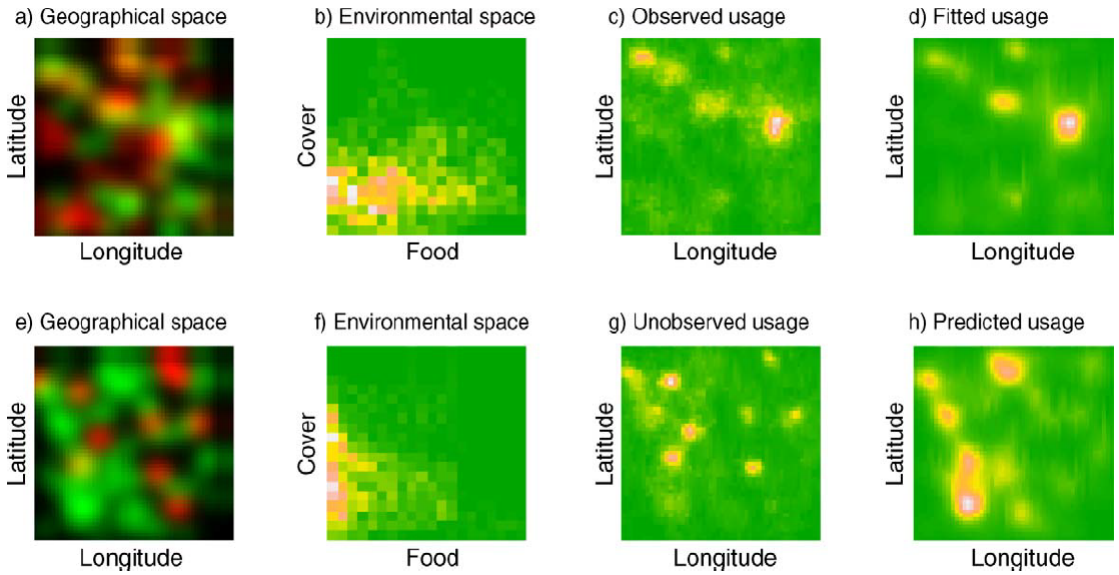


Figure 1.1: Illustration of the key challenge with the transferability of SDMs. The two rows represent two different environments with different geographical (a and e) and environmental (b and f) spaces. The local densities of the two resources in a and e are represented by the intensity of the two colors (red and green). The colors in the environmental space plots b and f represent the prevalence of a particular combination of values for the two resources going from green (low) to white (high). The colors in the observed usage plots c and represent species abundance in the two environments in terms of latitude and longitude for the ground truth where green indicates low abundance levels and white represents a high abundance levels. The generalized linear model fits well in the first environment by comparing c and d. However, the same model provides poor predictions in the second row when applied to a different environment using the same animal by comparing g and h. This figure has been taken from Matthiopoulos et al. (2011).

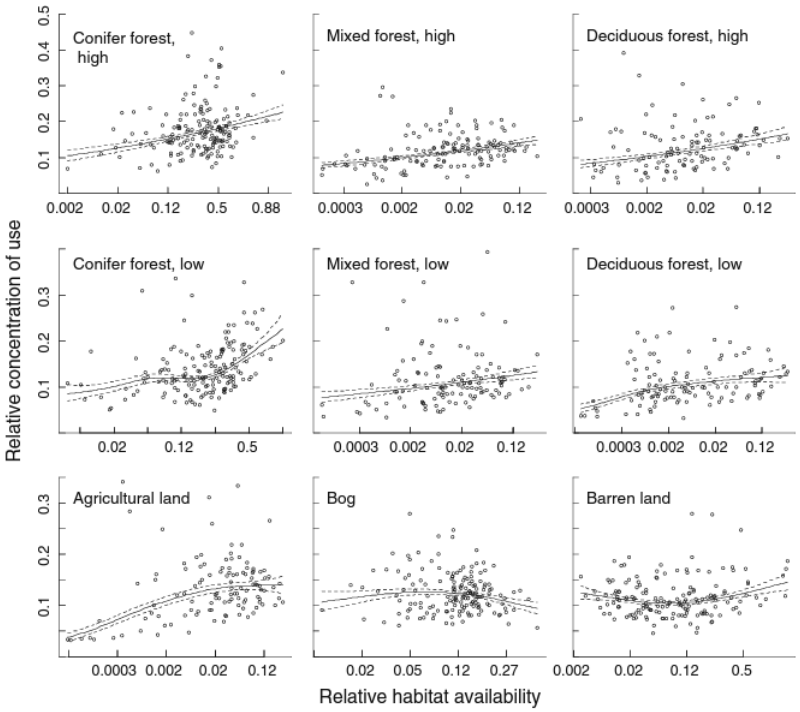


Figure 1.2: Evidence of functional responses in habitat selection taken from Bjørneraas et al. (2012). The plots represent the relationship between the proportion available of nine different habitat types and the concentration of moose use, which increased relative to the availability of several habitat types except for bog and barren.

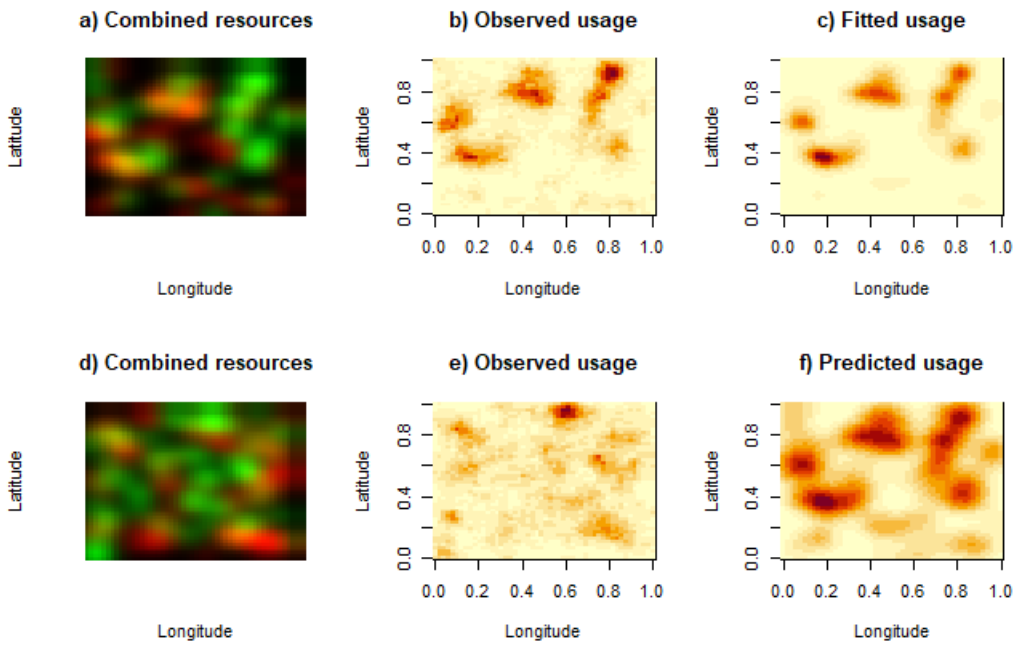


Figure 1.3: The two rows represent two different environments with different geographical (a and d) where the availability of two resources (food and cover) is represented by the intensity of the two colours (red and green). Panels d and e are heat maps of species abundance in the two environments in terms of latitude and longitude for the ground truth where light colours indicate low abundance levels, so the abundance levels increase as the colour shading gets darker. The generalized linear model fits well in the first environment by comparing b and c. However, the same model provides poor predictions in the second row when applied to a different environment using the same animal by comparing e and f.

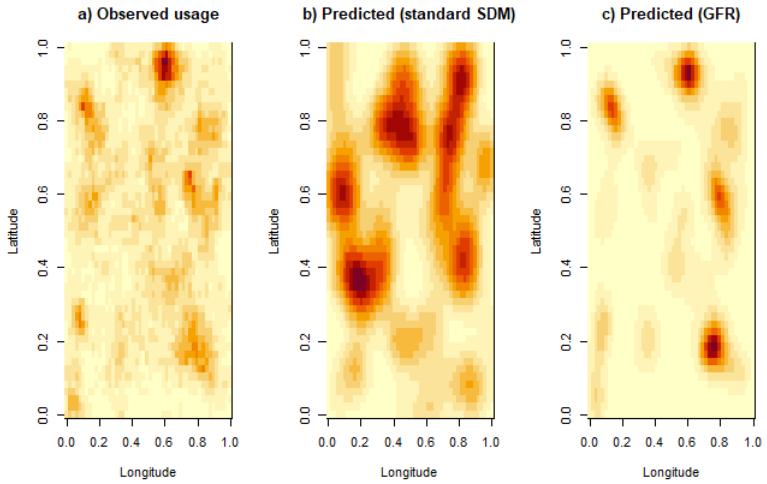


Figure 1.4: Heat maps of species abundance in the second environment (second row of Fig. 1.3) in terms of latitude and longitude for the ground truth where light colours indicate low abundance levels, so the abundance levels increase as the colour shading gets darker. The generalized linear model fits in the first environment (first row of Fig. 1.3) and provides poor predictions when applied to environment a (i.e., shows stronger deviations from the ground truth). The predictions from the GFR model show a very good agreement with the ground truth (by comparing a and c).

Chapter 2

Statistical Background and Related Literature

2.1 Introduction

This chapter provides an overview of the methodology, statistical background and related literature that are used throughout this thesis. Section 2.2 introduces the mathematical details of the generalized function response (GFR) model, which is the model motivating this thesis. The maximum likelihood estimation method which is used to estimate the GFR model parameters is outlined in section 2.3. Section 2.4 introduces the basis function approach, focusing on the radial basis function which is the foundation of the first extension model. Section 2.5 provides an overview of mixture models, with a particular focus on Gaussian mixture models (GMMs) and outlines the method that was used to estimate the GMM parameters. Section 2.6 introduces the main concept of the classification and regression tree, which is used in the second extension model. Section 2.7 gives a brief outline of model ensemble, which was used as a basis of the final extension models, especially the random forest model in Section 2.7.1 and the extreme gradient boosting model in Section 2.7.2. Section 2.8 covers the techniques used for making decisions about selecting the best number of parameters or models. The measures of the out-of-sample performance (and hence, the transferability) and total variability explained by models used in this thesis

are described in Section 2.9.

2.2 The Generalized Functional Response (GFR) Model

To explain GFRs more precisely, I need to introduce this class of models mathematically, starting from the basic form of modern species distribution models (SDMs). Both of the dominant methods for modelling species distributions (i.e., Maxent and Resource Selection Functions) use the following formulation of a predictor function of environmental covariates $\mathbf{x} = (x_1, \ldots, x_I)$ to predict the distribution of a species in space as follows:

$$h(\mathbf{x}) \;=\; \exp\left(\beta_0 + \sum_{i=1}^{I} \beta_i x_i\right) \tag{2.1}$$

where $h(\mathbf{x})$ is the habitat preference and the coefficient β_i for the ith covariate is fixed (i.e, the coefficients are scalars and not functions, Phillips et al., 2006; Boyce et al., 2002). The SDM here is written in terms of the environmental variable x, and the SDMs literature uses the exponential function (non-negative valued function), thereby avoiding negative usage by using some type of environmental variable, such as environmental risks. This exponential transformation of the linear predictors could be a generalized linear model (GLM) using an appropriate link function depending on the distribution of data (Matthiopoulos et al., 2020a). If the response variable is a binary species use/availability indicator, then the data can be modelled as a Bernoulli process and the logit transformation is used. Alternatively, for species abundance levels, the Poisson distribution is used with log transformation. Although the summand shown above is the prototypical expression for a linear predictor and $h(\mathbf{x})$ is a generalized linear model, the approach is augmented with customary extensions such as higher-order polynomial terms, interactions, or generalized additive terms. The coefficients of the linear predictor are either estimated by likelihood (Václavík and Meentemeyer, 2009) or entropy criteria (Phillips et al., 2006), and they are fitted to different types of data (e.g., telemetry, survey etc) by different link functions. Modern inference has unified all the existing approaches to different types of data under the framework of Inhomoheneous Point Processes (IPP), and has therefore tended to interpret the

quantity in Eq. (2.1) as the intensity function of the IPP (reviewed in Matthiopoulos et al., 2020a).

Habitat selection models formalized in this way are known to have low transferability to new environmental scenarios (Boyce and McDonald, 1999; Mysterud and Ims, 1999). The predictions from Eq. (2.1) rely on the assumption that animals use habitats in proportion to their preference and that preference does not change when habitat availability changes. This assumption is born of statistical convenience, not biological reality, where any changes in habitat availability lead to disproportionate changes in species' response, a phenomenon called functional response in resource selection (Mysterud and Ims, 1998).

Boyce and McDonald (1999) argued that functional responses could be captured by relaxing the stationarity of the fixed coefficients $\beta_i \in \mathbb{R}$. Hence, in their original GFR model, Matthiopoulos et al. (2011) allowed the β_i to vary as functions of habitat availability in the case of continuous environmental space:

$$\beta_{i,b} = \int \gamma_i(\mathbf{x}) f_b(\mathbf{x}) d\mathbf{x}, \tag{2.2}$$

where $f_b(\mathbf{x})$ is a probability density function for habitat availability in the b^{th} sampling instance for the i^{th} variable. A sampling instance represents an environmental scenario defined in a biological way as the environment experienced by the study animals during an appropriate spatiotemporal frame of accessibility (Matthiopoulos et al., 2020b). For example, a sampling instance could represent the spatial domain of a well-mixing subpopulation during a given year. A sampling instance could represent different years for the same population or different sub-populations in the same year (i.e., a space-for-time substitution in sampling effort is possible). An approximate, discretised version of this formulation, uses summation:

$$\beta_{i,b} = \sum_{n=1}^{N} \gamma_i(\mathbf{x}_n) f_b(\mathbf{x}_n), \tag{2.3}$$

where n encodes for a specific habitat. Intuitively, the function $\gamma_i(\mathbf{x})$ describes the change in the SDMs slope for the i^{th} covariate, generated by a unitary increase in the availability of the n^{th} habitat type.

Matthiopoulos et al. (2011) used a polynomial function to formulate the $\gamma_i(\mathbf{x})$ for each environmental variable $(\mathbf{x})$:

$$\gamma_i(\mathbf{x}) \;=\; \sum_{j=1}^{I} \sum_{m=0}^{M_j} \delta_{i,j}^{(m)} x_j^m \tag{2.4}$$

where the coefficient of $\gamma_i(\mathbf{x})$ for the m^{th} power of the j^{th} variable is $\delta_{i,j}^{(m)}$. This derivation leads to the following expression for the β's:

$$\beta_{i,b} \;=\; \gamma_{i,0} + \sum_{j=1}^{I} \sum_{m=0}^{M_j} \delta_{i,j}^{(m)} E[X_j^m]_b \tag{2.5}$$

where M_j is an integer order parameter and $E[X_j^m]_b$ is the mth moment of the covariate j calculated for the conditions prevailing in the bth sampling instance, which assumed to be normally distributed. Furthermore, $\gamma_{i,0}$ is an intercept corresponding to the scenario of zero expectations. If at least the first two moments of X are zero (corresponding to zero mean and variance for X), this implies that the environmental variable has its baseline value, uniformly across accessible space. Let $\boldsymbol{z}$ denote a vector composed of all elements $\{x_j^m\}$ and $\{E[X_j^m]_b\}$. Using the polynomial function approach described above, habitat preference $h(\boldsymbol{z})$ can be expressed as a function of the fixed effects of covariates and pairwise interactions between covariates and their moments:

$$h(\boldsymbol{z};\boldsymbol{\theta}) \;=\; \exp\left\{ \gamma_{0,0} + \sum_{i=1}^{I} \left(\sum_{m=0}^{M_i} \delta_{0,i}^{(m)} E[X_i^m]_b + \gamma_{i,0} x_i + x_i \sum_{j=1}^{I} \sum_{m=0}^{M_j} \delta_{i,j}^{(m)} E[X_j^m]_b \right) \right\} \tag{2.6}$$

where $\boldsymbol{\theta}$ is a parameter vector composed of the parameters γ_i and δ_i and $\boldsymbol{z}$ is a vector combining habitat variables x_i and their expectation values $E[X_i^m]$, as well as their product terms.

2.3 Maximum Likelihood (ML)

Seeking a good method to estimate a parameter θ of a population described by a function $p(x|\boldsymbol{\theta})$ is very important because it properly represents a population (Casella and Berger, 2021). A good estimation method is a method that provides a consistent, efficient, and unbiased estimation (Murphy, 2012). An estimator is unbiased when the sampling distribution of the estimated parameter is centred around the true parameter (i.e, $bias(\hat{\theta}) = E(\hat{\theta}) - \theta = 0$; Murphy, 2012). An unbiased estimator is efficient if it provides an estimate with the smallest theoretically possible variance (i.e., it achieves the Cramer-Rao lower bound; Murphy, 2012; Miura, 2011). A consistent estimator is when the estimated parameter converges in probability to the true parameter as the sample size of the data goes to infinity. There are different methods used for parameter estimating. For example, the method of moments is, perhaps, the oldest method used for estimating parameters (Casella and Berger, 2021; Tallis and Light, 1968). The method of moments is an easy method to use for parameter estimation. However, the estimates of the method of moments could be inaccurate in some cases (Marchisio and Fox, 2005). Furthermore, the method of moments is sometimes biased and inefficient estimators (Pearson, 1936). Bayesian inference is a popular approach to estimating parameters of interest depending on a prior distribution of the parameters, which are not fixed quantities (Casella and Berger, 2021). However, unless the posterior distribution is available in closed form, Bayesian inference is an expensive computational approach (Sunnåker et al., 2013).

Maximum likelihood (ML) is a widely used parameter estimation method (Casella and Berger, 2021). Under certain regularity conditions, maximum likelihood estimators (MLEs) are consistent, asymptotically unbiased and asymptotically efficient (Murphy, 2012). In general, the goal of MLE is to maximize the likelihood function $L(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{y})$; find $\hat{\boldsymbol{\theta}}$ that maximizes the function $L(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{y})$ over the parameter space. It is often easier to work with the log likelihood function $l(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{y}) = \frac{1}{N} \log L(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{y})$. If the response variable for a dataset is a binary species use/availability indicator, that is, $y_n \in \{0, 1\}$, then the Bernoulli model is used in each site with probability of use $p = Pr(\boldsymbol{y} = 1|\boldsymbol{z}; \boldsymbol{\theta})$ and

probability of availability $1-p = Pr(\boldsymbol{y}=0|\boldsymbol{z};\boldsymbol{\theta})$ where:

$$p = Pr(y_n = 1|\boldsymbol{z}_n;\boldsymbol{\theta}) = \frac{1}{1+e^{-h(\boldsymbol{z}_n;\boldsymbol{\theta})}}; \tag{2.7}$$

$$Pr(y_n = 0|\boldsymbol{z}_n;\boldsymbol{\theta}) = 1-p \tag{2.8}$$

where the subscript n denotes a geographical patch or plot where species counts are taken. The likelihood function is:

$$L(\boldsymbol{\theta},\boldsymbol{z},\boldsymbol{y}) = \prod_{n=1}^{N} Pr(y_n|\boldsymbol{z}_n;\boldsymbol{\theta}). \tag{2.9}$$

The log likelihood is:

$$l(\boldsymbol{\theta},\boldsymbol{z},\boldsymbol{y}) = \sum_{n=1}^{N} \log Pr(y_n|\boldsymbol{z}_n;\boldsymbol{\theta}). \tag{2.10}$$

For species abundance levels, the Poisson distribution is used with mean parameter equals to $h(\boldsymbol{z};\boldsymbol{\theta})$ in Eq. (2.6) for the GFR model. If y_n is the number of species in cell n, then

$$p(y_n|\boldsymbol{z}_n;\boldsymbol{\theta}) = \frac{h(\boldsymbol{z}_n;\boldsymbol{\theta})^{y_n} e^{-h(\boldsymbol{z}_n;\boldsymbol{\theta})}}{y_n!}. \tag{2.11}$$

The likelihood function using the maximum likelihood algorithm to optimize $\boldsymbol{\theta}$ is:

$$L(\boldsymbol{\theta},\boldsymbol{z},\boldsymbol{y}) = \prod_{n=1}^{N} \frac{h(\boldsymbol{z}_n;\boldsymbol{\theta})^{y_n} e^{-h(\boldsymbol{z}_n;\boldsymbol{\theta})}}{y_n!} \tag{2.12}$$

The log likelihood is:

$$
\begin{aligned}
l(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{y}) &= \sum_{n=1}^{N} \log \frac{h(\boldsymbol{z}_n;\boldsymbol{\theta})^{y_n} e^{-h(\boldsymbol{z}_n;\boldsymbol{\theta})}}{y_n!} \\
&= \sum_{n=1}^{N} [\log(h(\boldsymbol{z}_n;\boldsymbol{\theta})^{y_n}) + \log(e^{-h(\boldsymbol{z}_n;\boldsymbol{\theta})}) - \log(y_n!)] \\
&= \sum_{n=1}^{N} [y_n \log(h(\boldsymbol{z}_n;\boldsymbol{\theta})) - h(\boldsymbol{z}_n;\boldsymbol{\theta}) - \log(y_n!)]
\end{aligned}
\tag{2.13}
$$

Further note that for GLM-type models, the ML equations have no closed-form solution, and the iteratively reweighted least squares optimization algorithm is therefore applied (see e.g., Section 4.3.3 in Bishop, 2006; McCullagh and Nelder, 1983; Faraway, 2016; Chapter 3 in Wood, 2017 for details).

2.4 Basis Functions

The basis function approach is widely used as a flexible regression and an extension of the linear regression model (Bishop, 2006) where the linear regression model is in the form:

$$
y(\mathbf{x}, \boldsymbol{\beta}) \; = \; \sum_{j=0}^{I} \beta_j x_j \tag{2.14}
$$

for $\mathbf{x} = (x_1, \ldots, x_I)$ and the linear coefficients $\boldsymbol{\beta}$. The basis function represents the target variable y as a linear combination of fixed non-linear functions instead of a linear combination of input variables as used in linear regression as follows:

$$
y(\mathbf{x}, \boldsymbol{\delta}) \; = \; \sum_{j} \sum_{m=0}^{M_j} \delta_j^{(m)} \phi(x_j, \boldsymbol{\theta}_{j,m}) \tag{2.15}
$$

where $\phi(x_j, \boldsymbol{\theta}_{j,m})$ is a basis function of input covariate x_j with parameters $\boldsymbol{\theta}_{j,m}$ for the j^{th} varaible and m^{th} basis function.

There are many possible choices of ϕ. The polynomial basis function is one possibility

of ϕ, as follows:

$$y(\mathbf{x}, \boldsymbol{\delta}) \;=\; \sum_{j} \sum_{m=0}^{M_j} \delta_j^{(m)} x_j^m. \tag{2.16}$$

Matthiopoulos et al. (2011) used the same method to formulate $\gamma_i(\mathbf{x})$ in Eq. (2.4). A spline basis function of degree K addresses the limitation of global functions using the polynomial approach discussed in Section 2.2. A spline basis function of order K is achieved by dividing the input domain into regions and then fitting the polynomial function of degree $K-1$ in each region.

A radial basis function (RBF) expansion is a widely used flexible basis function approach, which is a local approach that disentangles the number of basis functions from the degree of differentiability:

$$\phi(x_j, \boldsymbol{\theta}_{j,m}) \;=\; \exp\left(-\frac{1}{2}\frac{(x_j - \xi_{j,m})^2}{\sigma_{j,m}^2}\right) \tag{2.17}$$

where $\xi_{j,m}$ is the center of the m^{th} basis function for the j^{th} covariate and $\sigma_{j,m}$ is its bandwidth parameter.

A sigmoidal basis function is another way of formulating ϕ, as follows :

$$\phi(x_j, \boldsymbol{\theta}_{j,m}) \;=\; \frac{1}{1+\exp\left(\frac{-(x_j-\xi_{j,m})}{\sigma_{j,m}}\right)} \tag{2.18}$$

A Fourier basis function is an extension of a sigmoidal basis function using the sinusoidal function as the basis function, which is not local where the spatial range is infinite, from positive to negative infinity, leading to poor sharp approximation (Donald et al., 2009). In contrast, a wavelet basis function is a flexible basis function approach that is localized in finite ranges, leading to better sharp approximation.

2.5 Gaussian Mixture Model (GMM)

Finite mixture models are statistical/machine learning models commonly used to estimate an unknown distribution or cluster a data into k clusters (McLachlan and Rathnayake, 2014). The importance of finite mixture models increased after it was proven that, under certain regularity conditions, a target probability distribution can be arbitrarily closely approximated by a mixture model (provided there is no limit on the number of mixture components) (Nguyen et al., 2020; McLachlan et al., 2019). The model in the mixture is one of the exponential families such as Gaussian, Poisson, Binomial, Bernoulli, Beta or Gamma as follows:

$$f(\mathbf{x}) \;=\; \sum_{k=1}^{K} w_k p(\mathbf{x}|\boldsymbol{\theta}_k) \tag{2.19}$$

where K is the total number of mixture components, w_k is the mixing weight of the k^{th} component, which satisfies $\sum_{k=1}^{K} w_k = 1$ and $\boldsymbol{\theta}_k$ is the vector of the distribution's unknown parameters of component k. The Gaussian mixture model (GMM) is a finite mixture probability distribution model which assumes that observations in the data are generated from a finite number of Gaussian distributions with unknown parameters (Boiarov and Granichin, 2019). The GMM traditionally uses the maximum likelihood estimator to estimate the GMM parameters via an expectation–maximisation algorithm (Xuan et al., 2001). The GMM assumes that the data point in $\mathbf{x} = (x_1, \ldots, x_I)$ is generated from a Gaussian distribution mixture with unknown parameters as follows:

$$f(\mathbf{x}) \;=\; \sum_{k=1}^{K} w_k N(\mathbf{x}|\boldsymbol{\mu}_k, \mathbf{C}_k) \tag{2.20}$$

where K is the total number of mixture components, $\boldsymbol{\mu}_k$ defines its centre and $\mathbf{C}_k$ is the covariance matrix. $N(\mathbf{x}|\boldsymbol{\mu}_k, \mathbf{C}_k)$ is Gaussian distribution for each component defined as follows:

$$N(\mathbf{x}|\boldsymbol{\mu}_k, \mathbf{C}_k) \;=\; \frac{1}{(2\pi)^{\frac{I}{2}}|\mathbf{C}_k|^{\frac{1}{2}}} \cdot \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)'\mathbf{C}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right) \tag{2.21}$$

The covariance matrix $\mathbf{C}_k$ can be a diagonal matrix, as follows:

$$N(\mathbf{x}|\boldsymbol{\mu}_k, \mathbf{C}_k) \;=\; \prod_{j=1}^{I} \frac{1}{\sqrt{2\pi}\sigma_{j,j}} \cdot \exp\left(-\frac{1}{2}\frac{(x_j-\mu_{j,k})^2}{\sigma_{j,j}^2}\right) \tag{2.22}$$

I used the GMM to model the probability distribution $f_b(\mathbf{x})$ in Eq. 2.2 (i.e., the habitat availability characterising a sampling instance). There are different choices for the components' orientations, shapes and volumes. These components' characteristics can be the same or vary between clusters. The covariance matrix $\mathbf{C}_k$ determines the shape, volume and orientation of each component. The covariance matrix $\mathbf{C}_k$ can be written as follows:

$$\mathbf{C}_k \;=\; \lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T \tag{2.23}$$

where λ_k is a constant that controls the component volume, $\mathbf{D}_k$ is an orthogonal matrix of eigenvectors control the component orientation and $\mathbf{A}_k$ is a diagonal matrix which controls the component shape (Russell et al., 2014). If each component has its own general covariance matrix, then the components are fully flexible and may independently adopt any volume, orientation and shape. If all components have the same covariance matrix, then they have the same shape, but this may be any shape. The contour axes are oriented along the coordinate axes if the model has diagonal covariance matrices. Spherical components are the components that have their own single variance.

For example, EII refers to components that are spherical with the same size, VII refers to components that are spherical with different sizes, EEI refers to components that are diagonal with the same size and shape, VEI refers to components that are diagonal with different sizes and the same shape, and EEE refers to components that are ellipsoidal with the same size, shape, and orientation (Fraley et al., 2012).

2.5.1 Expectation-Maximization for the GMM Parameters

The GMM parameters $\boldsymbol{\theta} = (w_1, w_2, ..., w_K, \boldsymbol{\mu}_1, , \boldsymbol{\mu}_2, ..., \boldsymbol{\mu}_K, \mathbf{C}_1, \mathbf{C}_2, ..., \mathbf{C}_K)$ are estimated using the maximum likelihood approach. An expectation-maximization (EM) algorithm is a broadly used for latent variable models to iteratively compute maximum likelihood estimates (Gentle et al., 2012). The EM algorithm attempts to find maximum likelihood

estimates for models with latent variables $\boldsymbol{z} = (z_1, ..., z_n)$ that determine the component from which the observation originates such that $p(z_i = k) = w_k$. The EM algorithm is an iterative algorithm which consists of an E-step followed by an M-step. The iteration starts with initial estimates of the parameters to run the E-Step and the M-Step and maximize the log-likelihood function until it converges; no further significant changes occur between the log-likelihood of the last iteration and the previous iteration, as explained in Algorithm 1.

Algorithm 1 Expectation–Maximization Algorithm

1: **procedure** ESTIMATE GMM PARAMETERS(*state*)
2: Start with initial values of $\boldsymbol{\theta}$.
3: **for** each data instance x_i **do**
4: $\mathbf{E} - \mathbf{step}$: compute the responsibility value r_{ik} (the probability that x_i belongs to k): $r_{ik} = \frac{w_k N(x_i|\boldsymbol{\mu}_k, \mathbf{C}_k)}{\sum_{j=1}^{K} w_j N(x_i|\boldsymbol{\mu}_j, \mathbf{C}_j)}$
5: **for** each component k **do**
6: $\mathbf{M} - \mathbf{step}$: use r_{ik} to re-estimate $\boldsymbol{\theta}$ as follows: $w_k = \frac{\sum_{i=1}^{n} r_{ik}}{n}$, $\boldsymbol{\mu}_k = \frac{\sum_{i=1}^{n} r_{ik} x_i}{\sum_{i=1}^{n} r_{ik}}$ and $\mathbf{C}_k = \frac{\sum_{i=1}^{n} r_{ik}(x_i - \boldsymbol{\mu}_k)(x_i - \boldsymbol{\mu}_k)^T}{\sum_{i=1}^{n} r_{ik}}$
7: Evaluate the log likelihood: $\sum_{i=1}^{n} \log \sum_{k=1}^{K} w_k N(x_i|\boldsymbol{\mu}_k, \mathbf{C}_k)$
8: **Repeat** steps 3 to 7 until the log likelihood convergence.

2.5.2 Number of Gaussian Components

Including a few Gaussian mixture components loses discrimination between the Gaussian components, while including many components may ignore the relevance among samples (Gao and Dai, 2014). Most studies set the number of components either as a fixed number or based on some information criteria (Gao and Dai, 2014). Some studies set the number of components based on cross-validated log likelihood, where the data is divided into training and validation sets to try different numbers of components and select the best based on log likelihood values (McLachlan and Rathnayake, 2014). In this thesis, the number of components k is selected based on the Bayesian information criterion (BIC).

2.6 Classification and Regression Trees (CARTs)

Classification and regression trees (CARTs) are widely used in regression and classification to predict variables in a new dataset using training data. CART can be represented by a binary tree, leading to one leaf per region, as illustrated in the top panel of Figure 2.1. For a comprehensive review of tree-based methods, refer to Breiman et al. (1984). A cost function is used to choose the root node variable, which is the best candidate variable to start the tree, and the best split value, which is the threshold of the candidate variable (Murphy, 2012). Consider the bottom panel of Figure 2.1. The left panel shows a hypothetical domain defined by two habitat variables, x_1 and x_2, and a population of species observations, reported by a team of ecologists. Green triangles indicate that a species has been found and reported, while red circles indicate the absence of a species. The histograms on the right show the distribution of presence/absence labels for two alternative candidate splits, one at habitat variable x_1, the other at habitat variable x_2. Which split is better? Intuitively, the bottom split is better, in that it gives a clearer separation of the regions in habitat space where species can be found and where they are absent, whereas the former split still shows a high degree of uncertainty. The node x_i and the threshold or split value h, which is a value from x_i, is chosen based on criteria such as the Breiman criteria, as in Breiman et al. (1984). In the Breiman criteria, a locally optimal maximum likelihood estimator is used for the split function. The regression cost for a certain node is the residual error left after fitting the model in each leaf using the variables in the path from the root to that node (Murphy, 2012). Entropy and Gini index are common classification costs; the split with the lower Gini or entropy scores would be preferred (Murphy, 2012). The classification cost using the Gini index refers to the probability of a randomly chosen element being classified in an incorrect class if it was randomly classified based on the distribution of classes in the data as follows:

$$Gini \;=\; 1 - \sum_{i=1}^{k} p_i^2 \tag{2.24}$$

where p_i is the probability of a randomly chosen element in the leaf being classified in class i and k is the total number of classes in this leaf. Entropy is defined as follows:

$$H(p) \;=\; -\sum_{i=1}^{k} p_i \log(p_i) \tag{2.25}$$

Fig. 2.2 shows the entropy measure for a two-class problem where $H(p)$ is maximized for $p = 0.5$, i.e., for maximal uncertainty, and minimized for $p \in \{0,1\}$, i.e., when there is no uncertainty. The minimum value of $H(p)$ is when $p = \; 0 \; or \; 1$, as follows:

$$\lim_{p_i \to 1} p_i \log(p_i) \;=\; 1 \;\times\; \log(1) = \; 0 \tag{2.26}$$

For $p_i \to 0$, L'Hopital's rule is used as follows (Lopez, 1994):

$$\lim_{p_i \to 0} p_i \log(p_i) \;=\; \lim_{p_i \to 0} \frac{\log(p_i)}{\frac{1}{p_i}} \;=\; \lim_{p_i \to 0} \frac{[\log(p_i)]'}{[\frac{1}{p_i}]'}$$

$$= \; - \lim_{p_i \to 0} \frac{\frac{1}{p_i}}{\frac{1}{p_i^2}} \;=\; - \lim_{p_i \to 0} p_i \;=\; 0 \tag{2.27}$$

It can be shown that the maximum value of $H(p)$ for a two-class problem is when $p = \frac{1}{k} = 0.5$. Before taking the derivation of $H(p)$ with respect to p and setting it equal to zero in order to find the maximum value, the constraint $C(p) = \sum_{i=1}^{k} pi = 1$ and the Lagrange multiplier technique is used as follows:

$$\nabla_p H(p) - \lambda \nabla_p C(p) = 0 \tag{2.28}$$

The derivation of Eq. (2.28) is:

$$-[\log(p_i) + 1] - \lambda = 0 \tag{2.29}$$

By solving Eq. (2.29) for p_i, I get:

$$p_i = exp(1 - \lambda) = c$$

$$\Longrightarrow \; p_i = c \tag{2.30}$$

Thus, using the constraint, I get:

$$1 = \sum_{i=1}^{k} p_i = \sum_{i=1}^{k} c = kc$$

$$\implies c = \frac{1}{k} \tag{2.31}$$

Inserting Eq. (2.31) into Eq. (2.30) leads to the following:

$$p_i = \frac{1}{k} \tag{2.32}$$

Because the original GFR model is in the generalized linear model (GLM) family, which limits its modelling flexibility, the CART model is used here to address this by recursively partitioning the input space into subregions, each modelled with a separate GLM. A pruning scheme is used whereby the minimum cross-validation, based on a 10-fold cross-validation scheme on the training data (but excluding the test data) determines the best number of terminal nodes.

2.7 Ensemble Models, Bagging and Boosting

The model ensemble is a widely used machine learning method, especially in recent decades (Polikar, 2012; Zhang and Ma, 2012; Pintelas and Livieris, 2020). The growth of ensemble use comes after it was proven effective in real applications and problem-solving (Zhou, 2012). A model ensemble trains collections of base models that are supervised, which learns from training data to predict in a new dataset (Oza and Russell, 2001), by combining multiple models to make a final decision (Sagi and Rokach, 2018). The main idea of a model ensemble is to provide a final model by combining multiple models for solving machine learning tasks and yielding better results than any individual model used in the model ensemble. The model ensemble reduces the problem of over-fitting that can occur in a single model, especially using small datasets. The over-fitting problem is reduced by averaging the outcomes that result from the ensemble members, which reduces the risk of predicting a wrong outcome using a single model (Sagi and Rokach, 2018).

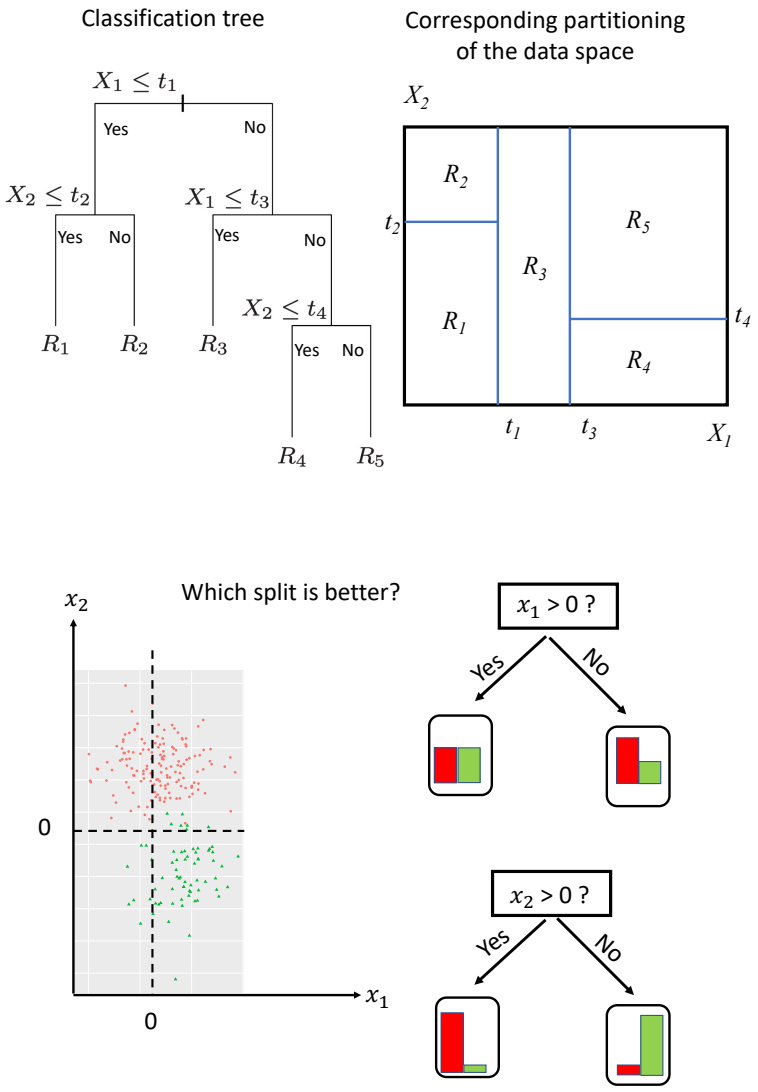


Figure 2.1: Illustration of CART. *Top panel:* Classification tree (left) and corresponding partitioning of the data space (right). *Bottom panel:* Two alternative splits at an internal node. The left panel shows a hypothetical domain defined by two habitat variables x_1 and x_2, and a population of species observations. Green triangles indicate that a species has been found and reported, red circles indicate the absence of a species. The histograms on the right show the distribution of presence/absence labels for two alternative candidate splits, one at habitat variable x_1, the other at habitat variable x_2. Which split is better?

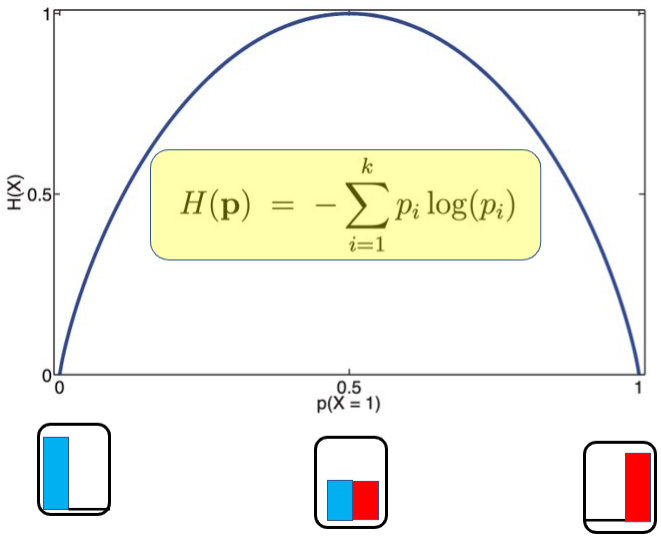


Figure 2.2: The entropy measure for binary classification $(k = 2)$. The horizontal axis corresponds to the probability of class 1 $(p(X = 1))$ and the vertical axis is the corresponding entropy score.

The ensemble approach helps to extend the search space, which helps make the optimal hypothesis inside the model space, by combining different models because the optimal hypothesis could be outside any single model space (Sagi and Rokach, 2018). Furthermore, the model ensemble performs very well when dealing with class imbalance issues, where the presence of one class is determined to be very high compared to other classes by using some techniques, such as training each of the members using a balanced subsample of the data (Nikulin et al., 2009; Sagi and Rokach, 2018). Moreover, this approach deals with high-dimensional datasets (Sagi and Rokach, 2018). This high-dimension problem can be improved with a model ensemble using bagging, where each ensemble member is trained using a random subset of features (Bryll et al., 2003) or using a different algorithm for feature selection (Rokach, 2008; Huang et al., 2010). A single output $\hat{y}_i$ is predicted using the model ensemble as follows:

$$\hat{y}_i = G(f_1, f_2, ..., f_k) \tag{2.33}$$

where G is an the aggregation function that combines the outputs from k ensemble members f_i, $i = 1,\ldots,k$. There are different techniques to train the model ensemble. The aggregation function is different for classification and regression problems. In classification problems, the predicted class is the one that has the most votes from the members. The average of the predictions across all members is the predicted value in a regression problem. There are two frameworks of the model ensemble; independent and dependent. An independent model ensemble is where all members of the ensembles are built independently from each other, such as bagging, random forests and extremely randomized trees (Sagi and Rokach, 2018). Bagging and random forests are similar techniques depending on bootstrapping and aggregating but the members of a random forest are trees and each node of a tree is chosen based on a subset of features while the members of bagging are any model and the nodes are selected from all the features. The extremely randomized trees method is different from bagging and random forests in the way that the split is chosen: randomly for extremely randomized trees and based on some criteria such as the Gini index or entropy for bagging and random forest (Geurts et al., 2006). Dependent frameworks, such as adaptive boosting, gradient boosting and extreme gradient boosting, are when each member of the model ensemble is built based on the other members. All the dependent methods use weak learners, which are models that perform slightly better than random guessing, in each step to create a better model in the subsequent step and combine these weak learners to create a strong learner, which performs much better than random guessing and achieves arbitrarily good accuracy (Zhou, 2012). In each iteration of the adaptive model, different weights are given to the observations based on the previous iteration: greater weight is awarded to previously misclassified observations, and weight is also assigned to each weak learner based on their overall predictive performance to combine them to create a final strong learner. The gradient boosting approach adds weak learners using gradient descent to minimize the loss function (Sagi and Rokach, 2018). First-order iterative optimisation algorithm, which is a gradient descent (first derivative) to find a local minimum, is used to minimize the loss function in the gradient boosting approach and the second-order iterative optimization, which uses Hessian (second derivative) to find a local minimum, is used in extreme gradient boosting. Random forests from the independent framework and extreme gradient boosting from the dependent framework

were used in this thesis, where each leaf node of the trees of the RF model and each iteration of the XGBoost model is a GFR model. These ensemble approaches were used because of the increased attention being paid to them due to the excellent results in classification and regression problems (Du et al., 2015; Rodriguez-Galiano et al., 2012; Rakhra et al., 2021; Sheridan et al., 2016).

2.7.1 Random Forest (RF)

The method of random forests (RFs) is well known for classification and regression problems. In general, the RF approach is an ensemble approach that contains a collection of trees, and each tree is constructed independently from the other trees by randomly selecting the tree inputs and features. The RF approach has been successfully implemented in different fields such as data science, ecology, chemoinformatics and bioinformatics (Biau and Scornet, 2016). The RF method has proven to be successful for prediction purposes since its innovation 21 years ago by L. Breiman (Breiman, 2001). Before that in 1996, L. Breiman built different trees using a bagging approach, a shortcut of "bootstrap aggregating", without replacement to create the forest (Breiman, 1996). The bagging approach was followed by the idea of choosing the split of each node randomly from the 20 best candidate splits based on information gain (Dietterich, 1998). The L. Breiman's RF approach used here was motivated by Amit and Geman's approach where each split is randomly selected from a large number of features defined geometrically (Breiman, 2001).

In addition to the RF's ability to predict well in new datasets, it has few tuning parameters, that show the ability to handle small-size datasets, a large number of features and real applications (Biau and Scornet, 2016). The expected loss can be explained by the bias and the variance of a model, which is called the bias-variance decomposition (Bishop, 2006). The expected out-of-sample prediction error of y_i can be decomposed into a bias and a variance component as follows:

$$
\begin{aligned}
E[(\hat{y}_i - y_i)^2] &= E[[(\hat{y}_i - \bar{y}) + (\bar{y} - y_i)]^2] \\
&= E[(\hat{y}_i - \bar{y})^2] + 2(\bar{y} - y_i)E[\hat{y}_i - \bar{y}] + (\bar{y} - y_i)^2 \\
&= E[(\hat{y}_i - \bar{y})^2] + (\bar{y} - y_i)^2 \\
&= var[\hat{y}_i] + bias^2[\hat{y}_i] = variance + bias^2
\end{aligned} \tag{2.34}
$$

where y_i is the observation in the test set, $\hat{y}_i$ is the model prediction and $\bar{y}$ is the mean of the predictions. For a flexible model, such as CART, the main contribution to this error comes from the variance term, that can be reduced in a model ensemble. In RF, the expected out-of-sample quadratic prediction error for y_i using K independent trees is:

$$
\begin{aligned}
E[\frac{1}{K}\sum_{k=1}^{K}(\hat{y}_i^{(k)} - y_i)]^2 &= var[\frac{1}{K}\sum_{k=1}^{K}\hat{y}_i^{(k)}] + [E[\frac{1}{K}\sum_{k=1}^{K}\hat{y}_i^{(k)}] - y_i]^2 \\
&= \frac{1}{K^2}\sum_{k=1}^{K}var[\hat{y}_i^{(k)}] + [\frac{1}{K}\sum_{k=1}^{K}E[\hat{y}_i^{(k)}] - y_i]^2 \\
&= \frac{1}{K^2}Kvar[\hat{y}_i] + [\frac{1}{K}\,K\,E[\hat{y}_i] - y_i]^2 \\
&= \frac{1}{K}var[\hat{y}_i] + [\bar{y}_i - y_i]^2 \\
&= \frac{variance}{K} + bias^2
\end{aligned} \tag{2.35}
$$

where $\hat{y}_i^{(k)}$ is the prediction from tree k. The variance contribution to the expected out-of-sample prediction error of a model ensemble can be reduced by a factor of K for K independent models, as seen in Eq. (2.35). Thus, bagging and subsets of candidate features for split nodes help make the models more independent. The idea behind the RF approach is to divide and conquer (Biau and Scornet, 2016). In this case, 'divide' refers to the partitioning of the configuration space, 'conquer' refers to the ability to learn complex functions, and 'strategy' refers to the approach to learn a complex function by breaking up a complex task into several simpler subproblems, each of which can be tackled with a simpler model, such as a GLM. The piecewise linear model is a very simple example in

which the data are divided into subsets and every subset is modelled with a linear model. The RF approach follows a divide-and-conquer strategy by recursively partitioning the input space into subregions and combining the predictions of multiple individual decision trees, fit in all subregions, to make a final prediction. This helps to reduce overfitting and improves the overall accuracy of the model. In machine learning, Breiman's RF consists of different trees where each tree uses a random subset of the dataset with replacement to insures each tree does not depend on the other trees to reduce the overall variance. The best prediction of a certain input x is the class receiving the highest vote from the trees in classification problems, and the average of the outputs across the trees in regression random forests (Breiman, 2001). The members of a random forest are trees and each node of a tree is chosen based on a subset of features while the members of bagging are any model and the nodes are selected from all the features, as seen in Algorithm 2 and Fig. 2.3 that have been taken from Raschka et al. (2022).

Algorithm 2 Bagging

1: **procedure** BAGGING OF m BOOTSTRAP SAMPLES ($state$)
2: **for** i=1 to m **do**
3: Draw bootstrap sample of size n, D_i .
4: Train base classifier h_i on D_i.
5: $\hat{y} = comb\{h_1(\mathbf{x}), ..., h_m(\mathbf{x})\}$: $comp$ is a function of classifiers (e.g., the most popular class in classification and the average of the outputs in regression).

In RF, each tree is created using a random sample of features from the dataset as described in Algorithm 3. I use bagging to obtain an individual member or tree of the forest. Each observation has a probability of $1/n$ of being selected in a bag and a probability of $1-1/n$ of not being chosen. The dataset is sampled n times with replacement and resulting in a bootstrap sample of n samples. Therefore, the probability of an observation not being selected in all the draws for a bag is $(1-1/n)^n$. As the value of n increases, $\lim_{n\to\infty}((1-1/n)^n) = e^{-1}$ and the probability of an observation being selected in a bag with replacement at least once = $1-1/e$ which is about 63.2% . Therefore, an individual member or tree of the forest D_n contains about 63.2% (the probability of an observation being selected in a bag with replacement at least once = $1-1/e$) of the observations, which are chosen randomly with replacement from the dataset . By contrast, the unused about

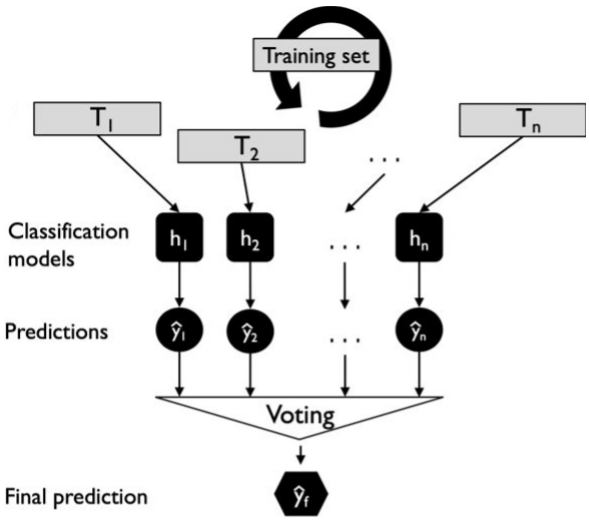


Figure 2.3: The process of a model ensemble using the bagging approach for a classification problem, as described in Algorithm 2 (Raschka et al., 2022). $T_1 ..., T_n$ are bootstrap samples used to predict $\hat{y}_i$'s form training different classifiers h_i. The final prediction $\hat{y}_f$ is the most popular class.

36.8 % of the observations are called out-of-bag data and used for validation purposes (Breiman, 1996; Han et al., 2022; Liaw and Wiener, 2018). Each node has p random candidate variables, which are the $\sqrt{v}$ for classification and $v/3$ for regression, where v is the total number of features in D dataset (Breiman, 2001). The optimal number of candidate variables achieves a trade-off between the correlation between the trees and the strength of the individual trees; increasing the candidate variables number leads to an increase in the strength of each tree, but a higher correlation between the trees and hence increases the generalization error (Breiman, 2001). Each tree is stopped growing if each terminal node consists of 1 point in classification or 5 points in regression based on the R package 'randomForest' (Liaw and Wiener, 2018).

2.7.1.1 Variable Importance

In general, machine learning methods are black-box methods that are hard to interpret (Rudin, 2019). In machine learning methods, variable importance refers to the relative

Algorithm 3 L. Breiman's Random Forest

1: **procedure** RANDOM FOREST OF $m = (1, ..., M)$ TREES $(state)$
2: **for** $m = 1$ to M **do**
3: Draw a bootstrap sample D_n from the training dataset D.
4: Build a full tree using the CART model described in section 2.6, but each node has p random candidate variables to be used.
5: Stop growing the tree if each terminal node consists of 1 point in classification or 5 points in regression.
6: **return** the best prediction of each input x: the most popular class in classification and the average of the outputs in regression.

importance of each variable in predicting a target variable. These measurements are vital because they help provide model improvement insight so we can better understand the model process (Lundberg and Lee, 2017). Each predictor in a random forest has a score of how important that variable in the forest is computed by using mean decrease in accuracy. The most important variable is the variable with the highest score, and the other variables are ranked accordingly. The importance score of each variable is based on an out-of-bag dataset. The out-of-bag dataset is the unused data in each tree (about 36.8%); about 63.2% of the data were used for training in each tree. To calculate the importance score for variable j (Han et al., 2016):

1: For out-of-bag data in each tree, compute the error rate for classification and mean squared error (MSE) for regression.

2: Randomly permute the j^{th} variable and calculate the error rate or MSE after permuting the variable.

3: Find the difference between the error rate and MSE from step 1 and the error rate and MSE from step 2 for each tree.

4: Average the difference from step 3 over all trees.

5: Normalize using the standard deviation of the difference from step 3.

2.7.2 Extreme Gradient Boosting (XGBoost)

In general, boosting is a machine learning method that is based on the principle that finding multiple moderate accurate rules is not as difficult as coming up with one highly accurate prediction rule (Schapire, 2003). The main goal of the boosting method is to improve the accuracy of any method or algorithm (Schapire, 2003). Boosting combines ensemble members (classifiers) to generate predictions that are better than any single ensemble member's predictions (Bishop, 2006). The boosting variant used in this work is extreme gradient boosting (XGBoost), proposed by Chen and Guestrin, 2016. XGBoost is another ensemble approach that was used to make predictions from regression and classification models using multiple trees, which shines in many files. XGBoost has recently come to dominate the machine learning field and has won many competitions in Kaggle (Poongodi et al., 2022). This considerable attention resulted from XGBoost's speed and the ability to perform well in an unseen dataset (Poongodi et al., 2022; Barnwal et al., 2022). XGBoost is an ensemble dependence model that focuses on difficult cases in the training set, which are hard to predict. It is a fast implementation of the gradient tree boosting approach (Chen et al., 2015), where each iteration aims to minimize the loss function and learn from the previous tree by computing the first partial derivatives of the loss function, whereas XGBoost uses the second partial derivatives of the loss function. The loss function for the classification problem is different from the loss function for the regression problem. The objective of XGBoost is to minimize the following function at iteration t:

$$L(s_j)^{(t)} = \sum_{j=1}^{k} \sum_{i \in I_j} [l(y_i, \hat{y_i}^{(t-1)}) + g_i s_j + \frac{1}{2} h_i s_j^2 + \frac{1}{2} \lambda s_j^2] \tag{2.36}$$

where s_j is the weight that one wants to optimize of the j^{th} node, k is the total number of leaf nodes in the tree and I_j is a set of instances belongs to the j^{th} node. The regularization parameter λ helps avoid the over-fitting of a tree by penalizing the leaf weights. $l(y_i, \hat{y_i}^{(t-1)})$ is the loss function of the previous prediction and equals $\frac{1}{2}(y_i - \hat{y_i}^{(t-1)})^2$ for regression and $l(y_i, \hat{y_i}^{(t-1)}) = -y_i log(\hat{y_i}^{(t-1)}) + (1 - y_i) log(1 - \hat{y_i}^{(t-1)})$ for classification. $g_i = \frac{\partial l(y_i, \hat{y_i}^{(t-1)})}{\partial \hat{y_i}^{(t-1)}}$ represents the first-order gradient statistics of the loss function and h_i is the second-order $\frac{\partial^2 l(y_i, \hat{y_i}^{(t-1)})}{\partial [\hat{y_i}^{(t-1)}]^2}$ resulting from using a second-order Taylor expansion to quickly

optimize $l(s_j)^{(t)}$. By omitting $l(y_i, \hat{y}_i{}^{(t-1)})$ from the optimization because it does not contain s_j, taking the derivative of Eq. (2.36) with respect s_j and setting it to 0 in order to minimize $l(s_j)^{(t)}$ (see Chen and Guestrin, 2016 for more details), we get:

$$s_j^* = \frac{-\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \tag{2.37}$$

Since the loss function for regression is different from the classification problem, s_j^* for regression is:

$$s_j^* = \frac{\sum_{i \in I_j} (y_i - \hat{y}_i{}^{(t-1)})}{\sum_{i \in I_j} 1 + \lambda} \tag{2.38}$$

s_j^* for classification is:

$$s_j^* = \frac{\sum_{i \in I_j} (y_i - \hat{y}_i{}^{(t-1)})}{\sum_{i \in I_j} [\hat{y}_i{}^{(t-1)}(1 - \hat{y}_i{}^{(t-1)})] + \lambda} \tag{2.39}$$

The corresponding $L(s_j^*)^{(t)}$ using Eq. (2.37) is:

$$L(s_j^*)^{(t)} = -\frac{1}{2} \sum_{j=1}^{k} \frac{(\sum_{i \in I_j} g_i)^2}{\sum_{i \in I_j} h_i + \lambda} \tag{2.40}$$

The first step to build the first tree is choosing an initial prediction $\hat{y}_i^{(0)}$ (usually $\hat{y}_i^{(0)} = \frac{\sum_{i=1}^{n} y_i}{n}$). The root node is the residuals $y_i - \hat{y}_i^{(0)}$, and to choose the best split of this node, the gain is calculated as follows:

$$Gain = L(s^*_{j(left)})^{(t)} + L(s^*_{j(right)})^{(t)} - L(s^*_{j(root)})^{(t)}$$

where $L(s^*_{j(left)})^{(t)}$ and $L(s^*_{j(right)})^{(t)}$ are the corresponding $L(s_j^*)^{(t)}$ in Eq. (2.40) for the left and right nodes of the j^{th} node and $L(s^*_{j(root)})^{(t)}$ is the corresponding $L(s_j^*)^{(t)}$ for the root. The branch that has the maximum gain score is chosen. The first tree is completed by choosing the split that has the maximum gain at each node. The prediction from the

first tree is:

$$\hat{y_i}^{(1)} = \hat{y_i}^{(0)} + \eta \times residual_{(i)}^{(1)}$$

where η is learning rate (= 0.3 by default) and $residual_{(i)}^{(1)}$ are the residuals from the first tree. The second tree is built based on the residuals from the first tree, and the tree is built using the gain score as in the first tree; continually build trees so that each tree predicts smaller residuals than the previous tree. The final prediction is:

$$\hat{y_i}^{(t)} = \hat{y_i}^{(0)} + \eta \times residual_{(i)}^{(1)} + \eta \times residual_{(i)}^{(2)} + ... + \eta \times residual_{(i)}^{(t)}$$

The ensemble size, which is the number of iterations t in XGBoost, matters much more for boosting than for the RF model. In XGBoost, large ensemble sizes can cause over-fitting because the gradient technique focuses on the most difficult cases, which can be noise cases and cause over-fitting. To avoid over-fitting in XGBoost, the validation set scheme is used to optimize the number of training iterations.

2.7.2.1 SHAP Feature Importance

The feature importance is computed in XGBoost using SHapley Additive exPlanations (SHAP) introduced by Lundberg and Lee, 2017. It is a model-agnostic local explanation technique that aims to explain each prediction of the model. Each feature value in the dataset has a SHAP value once a model is trained, and that value shows how much this feature value contributed to the difference between the prediction and the average prediction across all observations (Molnar, 2020). The sum of the SHAP values of feature values in a dataset row is the difference between the prediction of the row of features and the average of the predictions across the data, as follows:

$$\hat{y_i} = E(y) + \sum_{j=1}^{J} SHAP_j^{(i)} \tag{2.41}$$

where $E(y)$ is the predictions average, J is the total number of features in the i^{th} row of the dataset and $SHAP_j^{(i)}$ is the SHAP value of the j^{th} variable in row i of the dataset computed

as follows:

$$SHAP_j^{(i)} = \sum_{R \subseteq A \setminus \{j\}} \frac{|\,R\,|!\,(|\,A\,| - |\,R\,| - 1)!}{|\,A\,|} [f_{R\cup\{j\}}(x^{(i)}_{R\cup\{j\}}) - f_R(x^{(i)}_R)] \tag{2.42}$$

where A is the set that contains all the features in $x^{(i)}$: the observations in the i^{th} row of the dataset. R is all feature subsets of A, while $|\,A\,|$ and $|\,R\,|$ are the total number of features in the full set A and the subset R, respectively. $f_{R\cup\{j\}}(x^{(i)}_{R\cup\{j\}})$ is the model's output with the feature j included in the model, and $f_R(x^{(i)}_R)$ is the output without feature j included. Excluding a feature in a row from the model is accomplished by putting a random value from the train dataset instead of the actual value, in the row. Eq. (2.42) demonstrates the SHAP value for each feature value in the dataset after training a model. The SHAP feature importance P_j for feature j can be obtained using:

$$P_j = \frac{1}{N} \sum_{i=1}^{N} |\, SHAP_j^{(i)} \,| \tag{2.43}$$

where P_j is the average of the absolute SHAP values per feature j across the data (Molnar, 2020). This SHAP variable importance measure is an alternative measure to the mean decrease in accuracy used in the RF approach. The difference between the two measures is that the mean decrease in accuracy is computed based on decreases in model performance while the SHAP variable importance measure is computed using feature contributions of the output (Molnar, 2020).

2.8 Model Comparison

In a statistical model, selecting an adequate number of parameters that will fit data well is an issue that statisticians have always faced (Schwarz, 1978). There are some criteria for checking, evaluating and identifying different statistical models applied to the same observed dataset. A good model selection criterion attempts to achieve the trade-off between a model's goodness of fit and complexity (Myung and Pitt, 2004). Increasing the number of parameters, which is the case of maximizing likelihood function L, leads to

over-fitting. Adding a penalty term that controls the model complexity is necessary and was the primary motivation for model selection criteria. Akaike Information Criterion (AIC), Deviance information criterion (DIC), Watanabe-Akaike or widely available information criterion (WAIC), and Bayesian Information Criterion (BIC) are examples of model selection criteria that penalize the goodness of fit of a model (Gelman et al., 1995). In the present thesis, I use AIC and BIC because they are computationally affordable. DIC and WAIC are used for Bayesian model selection, where the posterior distributions of the models have been obtained by Markov chain Monte Carlo (MCMC) simulation which is computationally expensive.

2.8.1 Akaike Information Criterion (AIC)

AIC (Akaike, 1974) is a model selection criterion used to compare different models and determine the best fit for the data. AIC penalizes the maximum likelihood using the number of independent variables used to build the model, as follows:

$$AIC = -2\log(L(\hat{\boldsymbol{\theta}})) + 2\ p \tag{2.44}$$

where $L(\hat{\boldsymbol{\theta}})$ is the maximized value of the likelihood function represents the goodness of fit term, and p is the number of parameters used to build the model, which captures the complexity of the model. The model with the minimum AIC value would be preferred when comparing two or more models.

2.8.2 Bayesian Information Criterion (BIC)

BIC (Schwarz, 1978) is another common model selection criterion. BIC is derived using the log marginal likelihood, where the marginal likelihood is defined as follows:

$$p(\mathbf{x}) = \int p(\mathbf{x}|\boldsymbol{\theta})\ p(\boldsymbol{\theta})\ d\boldsymbol{\theta} \tag{2.45}$$

where $p(\boldsymbol{\theta})$ is the prior for the random variable $\boldsymbol{\theta}$ and $p(\mathbf{x})$ is the posterior density. The Gaussian approximation to the second-order around $\boldsymbol{\theta}^*$ by Taylor expansion is used to

approximate the log marginal likelihood (Murphy, 2012), as follows:

$$\log(p(\mathbf{x})) \approx \log(p(\mathbf{x}|\boldsymbol{\theta}^*)) + \log(p(\boldsymbol{\theta}^*)) - \frac{1}{2}\log(|H|) \tag{2.46}$$

where H is the Hessian matrix of second derivatives of the log posterior given as $H = \nabla^2 ln(p(\mathbf{x}|\boldsymbol{\theta}^*))$ and $\log|H| = \log|n\hat{H}| = \log(n^p|\hat{H}|) = p\log(n) + \log|\hat{H}|$ for a fixed matrix $\hat{H}$. Assuming that H is a full rank matrix, having uniform prior ($ln(p(\boldsymbol{\theta}) \propto 1$), dropping $\log|\hat{H}|$ because it is independent of n and using MLE to estimate $\boldsymbol{\theta}$, we get:

$$\log(p(\mathbf{x})) \approx \log(p(\mathbf{x}|\hat{\boldsymbol{\theta}})) - \frac{p}{2}\log(n) \tag{2.47}$$

where p is the number of parameters and n is the number of observations in the dataset. Maximizing the function $\log(p(\mathbf{x}))$ is equivalent to minimizing BIC, as follows:

$$BIC = -2\log(L(\hat{\boldsymbol{\theta}})) + p\ \log(n) \tag{2.48}$$

The size of the data is included in the penalty term of BIC. The model with the minimum BIC value would be preferred when comparing two or more models. AIC and BIC are conceptually different, and they are both asymptotic measures. AIC is used for predictive model selection, and BIC is used as an explanatory model selection (Shmueli, 2010; Sober, 2002). BIC prefers simpler models than AIC because the penalty term of BIC is larger than the penalty term of AIC, especially for large datasets. Therefore, in the case of having different models chosen as best based on the values of BIC and AIC, I choose the best model based on BIC because I want to have a less complex model and a more parsimonious model.

2.8.3 Effective Number of Parameters

AIC and BIC depend on the number of parameters p, where the simplest approach is to take the actual number of parameters to calculate AIC and BIC. The actual number of parameters for a model is the total number of parameters p in the model. The linear model

is defined as follows:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} \tag{2.49}$$

where $\mathbf{y}$ is the response variable and $\mathbf{X}$ is the inputs matrix of size $n \times p$ defined as follows:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

where n in this thesis denotes the number of units or geographical patches and p is the number of habitat variables. $\boldsymbol{\beta}$ is the vector of the model's coefficients. The method of ordinary least squares estimation to estimate $\boldsymbol{\beta}$ is minimizing the following quadratic prediction error:

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{2.50}$$

$\hat{\boldsymbol{\beta}}$ has the following closed-form solution (Goldberger et al., 1964):

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\,\mathbf{y} \tag{2.51}$$

The effective number of parameters (p_e) is $tr(\mathbf{S})$, where $\mathbf{S}$ is a smoothing matrix, that can be written by analogy with linear models as follows:

$$\mathbf{S} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \tag{2.52}$$

The smoothing matrix $\mathbf{S}$ is obtained by, formally, writing the vector of estimates $\hat{\mathbf{y}}$ as:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} \tag{2.53}$$

Then,

$$\begin{aligned} \hat{\mathbf{y}} &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\,\mathbf{y} \\ \hat{\mathbf{y}} &= \mathbf{S}\mathbf{y} \end{aligned} \tag{2.54}$$

Without the regularization approach, the effective number of parameters (p_e) reduces to the actual number of parameters p, as follows:

$$\begin{aligned} p_e &= tr(\mathbf{S}) \\ &= tr[\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T] \\ &= tr[\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}] \\ &= tr[\mathbf{I}] = dim(\boldsymbol{\beta}) = p \end{aligned} \tag{2.55}$$

However, if the matrix $(\mathbf{X}^T\mathbf{X})^{-1}$ in Eq. (2.51) is singular or to prevent over-fitting, the regularized approach can be used. Ridge regression is a particular form of regularization, where the quadratic prediction error in Eq. (2.50) becomes:

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^T\boldsymbol{\beta} \tag{2.56}$$

where λ is a regularization parameter. The corresponding solution for $\hat{\boldsymbol{\beta}}$ is given by (Murphy, 2012):

$$\hat{\boldsymbol{\beta}}_{ridge} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\,\mathbf{y} \tag{2.57}$$

Thus, the smoothing matrix for ridge regression can be written as:

$$\mathbf{S}_{ridge} = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T \tag{2.58}$$

The effective number of parameters for ridge regression is calculated as:

$$\begin{aligned} p_e = tr[\mathbf{S}_{ridge}] = \; & tr[\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T] \\ & = tr[\mathbf{H}_\lambda] \\ & = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda} \end{aligned} \tag{2.59}$$

where d_j are the singular values of $\mathbf{X}$; the square roots of the eigenvalues of $\mathbf{X}^T\mathbf{X}$ (the details of calculating the singular values can be found in Stewart, 1993). If λ is zero (i.e, without using the regularization approach), $p_e = \sum_{j=1}^{p} 1 = p$, which is the same result obtained using Eq. (2.55). The concept of the effective number of parameters is illustrated in Fig. 2.4

Increasing λ reduces the effective number of parameters in the model. I have explained the concept to the effective number of parameters for linear models. To obtain the effective number of parameters for a generalized linear model, a linearization via a first-order Taylor series expansion of the model output around the mode is carried out, where the mode is the parameter vector at the optimum of the regulazised objective function in Eq. (2.56). The matrix $\mathbf{X}$ in the previous equations is now replaced by the Jacobian matrix, i.e., the matrix of partial derivatives of the function outputs with respect to the model parameters. For more details and the exact mathematical expressions I refer the reader to the literature, e.g., Chapter 10 in Bishop (1995) (especially Eqs. (10.30), (10.31), (10.39), (10.68), (10.70)) or (MacKay, 1992). The effective number of parameters will be needed for the model selection discussed in Section 5.2.2.

Fig. 2.4 in my thesis is inspired by Fig. 10.12 in Bishop (1995), which approaches the topic of the efficient number of parameters from a Bayesian perspective. The two figures can be related as follows.

Take a zero mean multivariate normal distribution with isotropic covariance matrix (which is commonly done) as the prior distribution in parameter space. The log of this prior distribution is equal to the penalty term in Eq. (2.56), second term, except for an uninformative additive constant. See Eq. (10.9) in Bishop (1995), apply a log transformation and allow for the difference in notation.

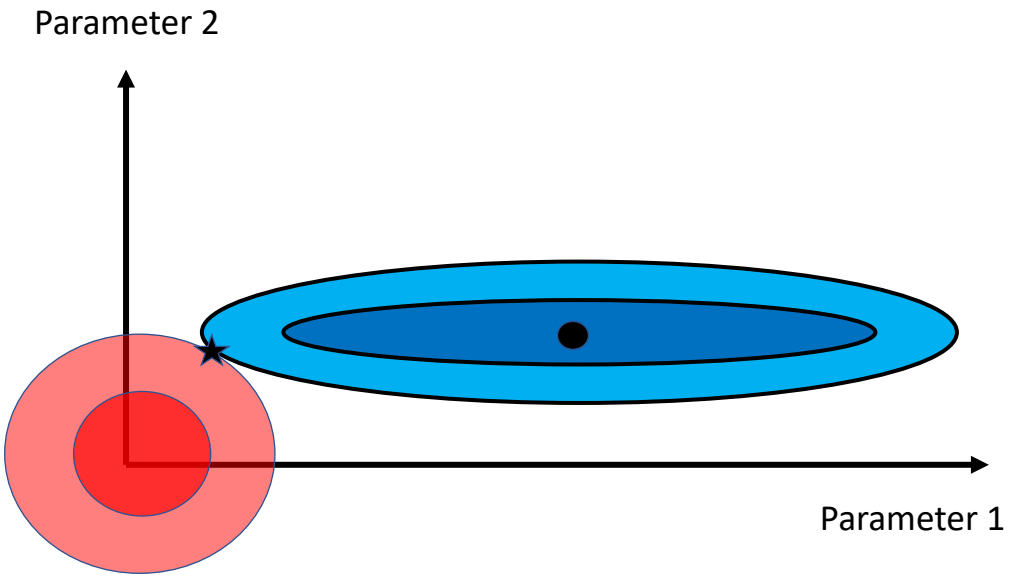


Figure 2.4: Illustration of the effective number of parameters. The figure shows contour lines of the unregularized objective function in Eq. (2.50) (in blue) and the L2 (ridge regression) regularization term in Eq. (2.56), second term, (in red) for a hypothetical model with two parameters. Without regularization, the optimum of the objective function is given by the black dot, and the effective number of parameters is equal to the total number of parameters. With regularization, the optimum of the regularized objective function is given by the black asterisk. The eigenvalues d_1^2 and d_2^2 (see Eq. (2.59)) for parameters 1 and 2 are inversely proportional to the curvature of the unregularized objective function, indicated by the blue contour lines; the regularization parameter λ is inversely proportional to the radius of the red contour lines. In the direction of parameter 1, the eigenvalue d_1^2 is small compared with λ and so the quantity $d_1^2/(d_1^2+\lambda)$ is close to zero. This implies that the first parameter does not make a significant contribution to the effective number of parameters and is effectively ignored. In the direction of parameter 2, the eigenvalue d_2^2 is large compared with λ and so the quantity $d_2^2/(d_2^2+\lambda)$ is close to 1. This implies that the second parameter makes a significant contribution to the effective number of parameters.

For iid Gaussian additive noise (which is also commonly assumed), the log likelihood is equal to the standard residual sum of squares objective function - Eq. (2.50) - except for an uninformative additive constant. See Eq. (10.15) in Bishop (1995), apply a log transformation, allow for the difference in notation as well as the difference between vector notation (Eq. (2.50)) and scalar notation (Bishop (1995)'s Eq. (10.15)).

2.9 Median, and Median-Absolute-Deviation

To measure the out-of-sample performance (and hence, the transferrability) of different models, the out-of-sample R^2 was used, derived by splitting the dataset in two parts, for training and testing. The metric is defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{2.60}$$

where y_i are the observations in the test set and $\bar{y}$ is the mean of the observations in the test set, $\hat{y}_i$ are the predictions. The k-fold cross-validation scheme was used to calculate the R^2; then the median of the out-of-sample R^2 scores obtained from each fold was calculated because it is more robust than the mean (Leys et al., 2013). I used k-fold cross-validation, where each fold is a block or sample instance. The block cross-validation approach is used to address the spatial autocorrelation of dataset structures because ignoring structure dependence in data increases the susceptibility to overfitting. I have implemented the block cross-validation approach to account for the autocorrelation of dataset structures and to reduce the overfitting issues of the models that are used to predict in different sample instances. For species abundance levels, $R^2_{DEV,P}$ is generally a better behaved statistic based on deviance residuals for count data regression models (Cameron and Windmeijer, 1996). $R^2_{DEV,P}$ is defined as:

$$R^2_{DEV,P} = 1 - \frac{\sum_{i=1}^{n}\{y_i log(y_i/\hat{y}_i) - (y_i - \hat{y}_i)\}}{\sum_{i=1}^{n} y_i log(y_i/\bar{y})}, \tag{2.61}$$

In addition, the median-absolute deviation (MAD) of the out-of-sample R^2 scores was used to measure the variability of the R^2 scores obtained from each fold. MAD is a more

robust quantity than standard deviation, being less sensitive to outliers. In this study, MAD was thus defined as

$$MAD = median(\mid R^2 - \widetilde{R^2} \mid) \times c$$

where $\widetilde{R^2}$ is the median of the out-of-sample R^2 scores and c = 1.4826, with the latter acting as a factor that converts MAD to the standard deviation, based on an assumption that the data has a Gaussian distribution (Leys et al., 2013).

Pseudo R^2 score was used to measure the proportion of the total variability explained by the model as follows:

$$R^2 = 1 - \frac{Residual\ Deviance}{Null\ Deviance} = 1 - \frac{2l(\boldsymbol{\theta}_c, \boldsymbol{z}, \boldsymbol{y}) - 2l(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{y})}{2l(\boldsymbol{\theta}_c, \boldsymbol{z}, \boldsymbol{y}) - 2l(\boldsymbol{\theta}_0, \boldsymbol{z}, \boldsymbol{y})}, \tag{2.62}$$

where Residual Deviance is the difference between the log-likelihood for the full model $l(\boldsymbol{\theta}_c, \boldsymbol{z}, \boldsymbol{y})$ and the proposed model $l(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{y})$ whereas Null Deviance is the difference between the log-likelihood for the full model and the null model $l(\boldsymbol{\theta}_0, \boldsymbol{z}, \boldsymbol{y})$. The null model consists of the intercept only and the full model is where the number of parameters is equal to the number of data points (Smith and McKenna, 2013).

Chapter 3

Methodological Innovation

3.1 Introduction

To address the limitations of the original GFR model proposed in Matthiopoulos et al. (2011), I have adapted three state-of-the-art flexible regression paradigms to model the habitat selection coefficients. The first approach is based on a radial basis function (RBF) expansion, as reviewed in Section 2.4, and I refer to this model as the RBF-GFR model in Section 3.2. Section 3.2.2 gives a brief outline of some methods I have applied to select the RBF-GFR model parameters. I provide details of how regularization reduces over-fitting in Section 3.3. Next, I combine classification and regression trees (CART), which have been reviewed in Section 2.6, with both the original GFR model and RBF-GFR model. I refer to these models as GFR-CART and RBF-GFR-CART, respectively, in Section 3.4. I finally create model ensembles based on random forests (RFs) trained with bagging (see Section 2.7.1) or boosting (see Section 2.7.2). I refer to these ensembles with the suffix 'RF' or 'XGBoost' in Sections 3.5 and 3.6, respectively. Table (3.1) and Fig. 3.5 in Section 3.7 provide outlines of the relationship between all the models proposed in the present study.

3.2 A Radial Basis Function (RBF-GFR) Model

Although the GFR model has been shown to achieve better generalization performance than the conventional GLM model of Eq. (2.1), it suffers from various limitations. The degree of nonlinear complexity and smoothness is restricted in advance: the functions in Eq. (2.6) are only M_j times differentiable, where the $M_j's$ are the highest polynomial orders. A complex function with a high degree of differentiability thus requires a large number of parameters, which renders the approach susceptible to over-fitting. Restricting the maximum polynomial order commensurately with the training set size leads to the paradox situation that the functional complexity of the habitat preference coefficients, which is an inherent property of the species and the habitat under investigation, becomes contingent on the arbitrariness of the data acquisition process. Moreover, while the degree of smoothness and model complexity is allowed to vary with respect to the choice of environmental variable, it is assumed to apply globally to the entire input domain. These shortcomings are well-known in the statistics and machine learning communities, and various flexible regression methods have been developed to address them (see e.g., Hastie et al., 2008). To address these limitations, I use a basis function expansion to model habitat preference $\gamma_i(\mathbf{x})$ instead of the polynomial function that was used in the original GFR model:

$$\gamma_i(\mathbf{x}) \;=\; \sum_j \sum_{m=0}^{M_j} \delta_{i,j}^{(m)} \phi(x_j, \boldsymbol{\theta}_{j,m}) \;=\; \sum_j \sum_m \delta_{i,j}^{(m)} \phi(x_j, \boldsymbol{\theta}_{j,m}) \tag{3.1}$$

where $\delta_{i,j}^{(m)}$ is the coefficient of $\gamma_i(\mathbf{x})$ for the m^{th} basis function of the j^{th} variable and 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, and the sum over m going from 0 to M_j. It can be shown that this is equivalent to a Gaussian process (see e.g., Bishop, 2006, Section 6.4), with the form of the covariance function determined by $\phi(.)$, by writing Eq. (3.1) in matrix form:

$$\boldsymbol{\gamma} \;=\; \boldsymbol{\Phi}\boldsymbol{\delta} \tag{3.2}$$

where $\mathbf{\Phi}$ is a design matrix composed of the basis functions $\phi(x_j, \boldsymbol{\theta}_{j,m})$, and placing a multivariate normal distribution as a prior on $\boldsymbol{\delta}$, $\boldsymbol{\delta} \sim N(\boldsymbol{\delta}|\mathbf{0}, \mathbf{C})$. This implies that $\boldsymbol{\gamma}$ is a zero-mean Gaussian process with covariance matrix $\mathbf{\Phi C \Phi}^{\mathsf{T}}$.

I choose a radial basis function (RBF) for $\gamma_i(\mathbf{x})$ because it is computationally easier to get a closed-form integral of the Gaussian distribution (see Bishop, 1995):

$$\gamma_i(\mathbf{x}) \;=\; \sum_j \sum_m \delta_{i,j}^{(m)} \exp\left(-\frac{1}{2} \frac{(x_j - \xi_{j,m})^2}{\sigma_{j,m}^2} \right) \tag{3.3}$$

where $\xi_{j,m}$ is the center of the mth basis function for the jth covariate and $\sigma_{j,m}$ is its bandwidth parameter.

I follow Matthiopoulos et al. (2015) and model the probability distribution $f_b(\mathbf{x})$ (i.e., the habitat availability characterising a sampling instance) with a Gaussian mixture model:

$$f_b(\mathbf{x}) \;=\; \sum_{k=1}^{K} [w_k]_b N(\mathbf{x}|[\boldsymbol{\mu}_k]_b, [\mathbf{C}_k]_b) \tag{3.4}$$

where K is the number of mixture components, $[w_k]_b$ is the mixing weight of the k^{th} component and b^{th} sample instance, $\boldsymbol{\mu}_k$ defines its centre and $\mathbf{C}_k$ is the covariance matrix. I assume that it is implied that f, w, $\boldsymbol{\mu}$, $\mathbf{C}$ are all specific to a sampling instance, and therefore that the subscript b is implied for all these quantities. Inserting Eq.(3.3) into Eq.(2.2) and making use of Eq.(3.4) gives:

$$\begin{aligned} \beta_{i,b} &= \gamma_{i,0} + \int \gamma_i(\mathbf{x}) f_b(\mathbf{x}) d\mathbf{x} \\ &= \gamma_{i,0} + \sum_j \sum_m \sum_k \delta_{i,j}^{(m)} [w_k]_b \left[\int \phi(x_j, \boldsymbol{\theta}_{j,m}) N(\mathbf{x}|[\boldsymbol{\mu}_k]_b, [\mathbf{C}_k]_b) d\mathbf{x} \right] \end{aligned} \tag{3.5}$$

where

$$\phi(x_j, \boldsymbol{\theta}_{j,m}) \;=\; \exp\left(-\frac{1}{2} \frac{(x_j - \xi_{j,m})^2}{\sigma_{j,m}^2} \right) \tag{3.6}$$

The coefficient $\beta_{i,b}$ has closed-form solution:

$$\beta_{i,b} \;=\; \gamma_{i,0} + \sum_{j}\sum_{m}\sum_{k} \delta_{i,j}^{(m)} [w_k]_b \psi(\boldsymbol{\theta}_{j,m}, [\boldsymbol{\mu}_k]_b, [\mathbf{C}_k]_b) \tag{3.7}$$

where

$$\psi(\boldsymbol{\theta}_{j,m}, [\boldsymbol{\mu}_k]_b, [\mathbf{C}_k]_b) \;=\; \int \phi(x_j, \boldsymbol{\theta}_{j,m}) N(\mathbf{x}|[\boldsymbol{\mu}_k]_b, [\mathbf{C}_k]_b) dx \tag{3.8}$$

The idea of the following derivation steps is to simplify Eq. (3.8) by getting a closed-form integral of the Gaussian distribution to take it out of the calculation as follows:

$$= \int \exp\left(-\frac{1}{2}\frac{(x_j - \xi_{j,m})^2}{\sigma_{j,m}^2}\right) \times \frac{1}{\sqrt{2\pi}[\sigma_{j,j}]_b} \cdot \exp\left(-\frac{1}{2}\frac{(x_j - [\mu_{j,k}]_b)^2}{[\sigma_{j,j}^2]_b}\right) dx$$

$$= \int \frac{1}{\sqrt{2\pi}[\sigma_{j,j}]_b} \cdot \exp\left[-\frac{1}{2}\left(\frac{(x_j - \xi_{j,m})^2 \cdot [\sigma_{j,j}^2]_b + (x_j - [\mu_{j,k}]_b)^2 \cdot \sigma_{j,m}^2}{[\sigma_{j,j}^2]_b \cdot \sigma_{j,m}^2}\right)\right] dx$$

By multiplying the exponential part by $\frac{[\sigma_{j,j}^2]_b + \sigma_{j,m}^2}{[\sigma_{j,j}^2]_b + \sigma_{j,m}^2}$, I get:

$$= \int \frac{1}{\sqrt{2\pi}[\sigma_{j,j}]_b} \cdot \exp\left[-\frac{1}{2}\left(\frac{(x_j - \xi_{j,m})^2 \cdot \frac{[\sigma_{j,j}^2]_b}{[\sigma_{j,j}^2]_b + \sigma_{j,m}^2} + (x_j - [\mu_{j,k}]_b)^2 \cdot \frac{\sigma_{j,m}^2}{[\sigma_{j,j}^2]_b + \sigma_{j,m}^2}}{\frac{[\sigma_{j,j}^2]_b \cdot \sigma_{j,m}^2}{[\sigma_{j,j}^2]_b + \sigma_{j,m}^2}}\right)\right] dx$$

By multiplying the denominator by $\sqrt{\frac{[\sigma_{j,j}^2]_b + \sigma_{j,m}^2}{[\sigma_{j,j}^2]_b + \sigma_{j,m}^2}}$, I get:

$$= \int \frac{1}{\sqrt{2\pi[\sigma_{j,j}^2]_b \cdot \frac{\sigma_{j,j}^2 + \sigma_{j,m}^2}{[\sigma_{j,j}^2]_b + \sigma_{j,m}^2}}} \cdot \exp\left[-\frac{1}{2}\left(\frac{(x_j - \xi_{j,m})^2 \cdot \frac{[\sigma_{j,j}^2]_b}{[\sigma_{j,j}^2]_b + \sigma_{j,m}^2} + (x_j - [\mu_{j,k}]_b)^2 \cdot \frac{\sigma_{j,m}^2}{[\sigma_{j,j}^2]_b + \sigma_{j,m}^2}}{\frac{[\sigma_{j,j}^2]_b \cdot \sigma_{j,m}^2}{[\sigma_{j,j}^2]_b + \sigma_{j,m}^2}}\right)\right] dx$$

$$= \int \frac{1}{\sqrt{2\pi \cdot [\frac{[\sigma^2_{j,j}]_b \cdot [\sigma^2_{j,j}]_b}{[\sigma^2_{j,j}]_b + \sigma^2_{j,m}} + \frac{[\sigma^2_{j,j}]_b \cdot \sigma^2_{j,m}}{[\sigma^2_{j,j}]_b + \sigma^2_{j,m}}]}} \cdot \exp\left[-\frac{1}{2}\left(\frac{(x_j - \xi_{j,m})^2 \cdot \frac{[\sigma^2_{j,j}]_b}{[\sigma^2_{j,j}]_b + \sigma^2_{j,m}} + (x_j - [\mu_{j,k}]_b)^2 \cdot \frac{\sigma^2_{j,m}}{[\sigma^2_{j,j}]_b + \sigma^2_{j,m}}}{\frac{[\sigma^2_{j,j}]_b \cdot \sigma^2_{j,m}}{[\sigma^2_{j,j}]_b + \sigma^2_{j,m}}}\right)\right] dx$$

$$= \int \frac{1}{\sqrt{2\pi \cdot \frac{[\sigma^2_{j,j}]_b}{[\sigma^2_{j,j}]_b + \sigma^2_{j,m}} \cdot [[\sigma^2_{j,j}]_b + \sigma^2_{j,m}]}} \cdot$$

$$\exp\left[-\frac{1}{2}\left(\frac{(x_j - \xi_{j,m})^2 \cdot \frac{[\sigma^2_{j,j}]_b}{[\sigma^2_{j,j}]_b + \sigma^2_{j,m}} + (x_j - [\mu_{j,k}]_b)^2 \cdot \frac{\sigma^2_{j,m}}{[\sigma^2_{j,j}]_b + \sigma^2_{j,m}}}{\frac{[\sigma^2_{j,j}]_b \cdot \sigma^2_{j,m}}{[\sigma^2_{j,j}]_b + \sigma^2_{j,m}}}\right)\right] dx \tag{3.9}$$

By moving the part $\sqrt{[\sigma^2_{j,j}]_b + \sigma^2_{j,m}}$ from the denominator to the exponential part, I get:

$$= \int \frac{1}{\sqrt{2\pi \cdot \frac{[\sigma^2_{j,j}]_b}{[\sigma^2_{j,j}]_b + \sigma^2_{j,m}}}} \cdot \zeta dx \tag{3.10}$$

where ζ is:

$$\zeta = \exp\left[\left(\ln \frac{1}{\sqrt{[\sigma^2_{j,j}]_b + \sigma^2_{j,m}}}\right) - \frac{1}{2}\left(\frac{(x_j - \xi_{j,m})^2 \cdot \frac{[\sigma^2_{j,j}]_b}{[\sigma^2_{j,j}]_b + \sigma^2_{j,m}} + (x_j - [\mu_{j,k}]_b)^2 \cdot \frac{\sigma^2_{j,m}}{[\sigma^2_{j,j}]_b + \sigma^2_{j,m}}}{\frac{[\sigma^2_{j,j}]_b \cdot \sigma^2_{j,m}}{[\sigma^2_{j,j}]_b + \sigma^2_{j,m}}}\right)\right] \tag{3.11}$$

After some algebra (the details of the following derivation are in Appendix A.1), I obtain the following for ζ:

$$\zeta = \exp\left[-\frac{1}{2}\left(\frac{(x_j - A)^2 + C}{\frac{[\sigma^2_{j,j}]_b \cdot \sigma^2_{j,m}}{[\sigma^2_{j,j}]_b + \sigma^2_{j,m}}}\right)\right] \tag{3.12}$$

where $A = \left(\frac{\xi_{j,m}[\sigma^2_{j,j}]_b + \sigma^2_{j,m}[\mu_{j,k}]_b}{\sigma^2_{j,m} + [\sigma^2_{j,j}]_b}\right)$, and

$$C = \frac{[\sigma^2_{j,j}]_b \sigma^2_{j,m}}{\left([\sigma^2_{j,j}]_b + \sigma^2_{j,m}\right)^2} \cdot \left([\mu_{j,k}]_b - \xi_{j,m}\right)^2 + \frac{[\sigma^2_{j,j}]_b \sigma^2_{j,m}}{\left([\sigma^2_{j,j}]_b + \sigma^2_{j,m}\right)} \cdot \left(\ln\left([\sigma^2_{j,j}]_b + \sigma^2_{j,m}\right)\right)$$

By using Eq. (3.12), $\psi(\boldsymbol{\theta}_{j,m}, [\boldsymbol{\mu}_k]_b, [\mathbf{C}_k]_b)$ is as follows:

$$\psi(\boldsymbol{\theta}_{j,m}, [\boldsymbol{\mu}_k]_b, [\mathbf{C}_k]_b) = \int \frac{1}{\sqrt{2\pi \cdot \frac{[\sigma^2_{j,j}]_b}{[\sigma^2_{j,j}]_b + \sigma^2_{j,m}}}} \cdot \exp\left[-\frac{1}{2}\left(\frac{(x_j - A)^2}{\frac{[\sigma^2_{j,j}]_b \cdot \sigma^2_{j,m}}{[\sigma^2_{j,j}]_b + \sigma^2_{j,m}}}\right)\right] \cdot \exp\left[-\frac{1}{2}\left(\frac{C}{\frac{[\sigma^2_{j,j}]_b \cdot \sigma^2_{j,m}}{[\sigma^2_{j,j}]_b + \sigma^2_{j,m}}}\right)\right] dx \tag{3.13}$$

By multiplying the first fraction of Eq. (3.13) by $\frac{\sigma_{j,m}}{\sigma_{j,m}}$, I get:

$$= \int \frac{\sigma_{j,m}}{\sqrt{2\pi \cdot \frac{[\sigma^2_{j,j}]_b \cdot \sigma^2_{j,m}}{[\sigma^2_{j,j}]_b + \sigma^2_{j,m}}}} \cdot \exp\left[-\frac{1}{2}\left(\frac{(x_j - A)^2}{\frac{[\sigma^2_{j,j}]_b \cdot \sigma^2_{j,m}}{[\sigma^2_{j,j}]_b + \sigma^2_{j,m}}}\right)\right] \cdot \exp\left[-\frac{1}{2}\left(\frac{C}{\frac{[\sigma^2_{j,j}]_b \cdot \sigma^2_{j,m}}{[\sigma^2_{j,j}]_b + \sigma^2_{j,m}}}\right)\right] dx$$

$$= \sigma_{j,m} \cdot \exp\left[-\frac{1}{2}\left(\frac{C}{\frac{[\sigma^2_{j,j}]_b \cdot \sigma^2_{j,m}}{[\sigma^2_{j,j}]_b + \sigma^2_{j,m}}}\right)\right] \cdot \int \frac{1}{\sqrt{2\pi \cdot \frac{[[\sigma^2_{j,j}]_b]_b \cdot \sigma^2_{j,m}}{[\sigma^2_{j,j}]_b + \sigma^2_{j,m}}}} \cdot \exp\left[-\frac{1}{2}\left(\frac{(x_j - A)^2}{\frac{[\sigma^2_{j,j}]_b \cdot \sigma^2_{j,m}}{[\sigma^2_{j,j}]_b + \sigma^2_{j,m}}}\right)\right] dx \tag{3.14}$$

The integral part of Eq. (3.14) is the integral of a Gaussian distribution, which is equal

to 1 with mean A and variance $\frac{[\sigma^2_{j,j}]_b \cdot \sigma^2_{j,m}}{[\sigma^2_{j,j}]_b + \sigma^2_{j,m}}$. So,

$$\psi(\boldsymbol{\theta}_{j,m}, [\boldsymbol{\mu}_k]_b, [\mathbf{C}_k]_b) \;=\; \sigma_{j,m} \cdot \exp\left[-\frac{1}{2}\left(\frac{C}{\frac{[\sigma^2_{j,j}]_b \cdot \sigma^2_{j,m}}{[\sigma^2_{j,j}]_b + \sigma^2_{j,m}}}\right)\right]$$

$$= \sigma_{j,m} \cdot \exp\left[-\frac{1}{2}\left(\frac{\frac{[\sigma^2_{j,j}]_b \sigma^2_{j,m}}{\left([\sigma^2_{j,j}]_b + \sigma^2_{j,m}\right)^2} \cdot \left([\mu_{j,k}]_b - \xi_{j,m}\right)^2 + \frac{[\sigma^2_{j,j}]_b \sigma^2_{j,m}}{\left([\sigma^2_{j,j}]_b + \sigma^2_{j,m}\right)} \cdot \left(\ln\left([\sigma^2_{j,j}]_b + \sigma^2_{j,m}\right)\right)}{\frac{[\sigma^2_{j,j}]_b \cdot \sigma^2_{j,m}}{[\sigma^2_{j,j}]_b + \sigma^2_{j,m}}}\right)\right]$$

$$= \sigma_{j,m} \cdot \exp\left[-\frac{1}{2}\left(\frac{\frac{[\sigma^2_{j,j}]_b \sigma^2_{j,m}}{\left([\sigma^2_{j,j}]_b + \sigma^2_{j,m}\right)^2} \cdot \left([\mu_{j,k}]_b - \xi_{j,m}\right)^2}{\frac{[\sigma^2_{j,j}]_b \cdot \sigma^2_{j,m}}{[\sigma^2_{j,j}]_b + \sigma^2_{j,m}}}\right)\right] \cdot \exp\left[-\frac{1}{2}\left(\ln\left([\sigma^2_{j,j}]_b + \sigma^2_{j,m}\right)\right)\right]$$

$$= \frac{\sigma_{j,m}}{\sqrt{\left([\sigma^2_{j,j}]_b + \sigma^2_{j,m}\right)}} \cdot \exp\left[-\frac{1}{2}\left(\frac{\frac{[\sigma^2_{j,j}]_b \sigma^2_{j,m}}{\left([\sigma^2_{j,j}]_b + \sigma^2_{j,m}\right)^2} \cdot \left([\mu_{j,k}]_b - \xi_{j,m}\right)^2}{\frac{[\sigma^2_{j,j}]_b \cdot \sigma^2_{j,m}}{[\sigma^2_{j,j}]_b + \sigma^2_{j,m}}}\right)\right]$$

Finally,

$$\psi(\boldsymbol{\theta}_{j,m}, [\boldsymbol{\mu}_k]_b, [\mathbf{C}_k]_b) \;=\; \frac{\sigma_{j,m}}{\sqrt{\left([\sigma^2_{j,j}]_b + \sigma^2_{j,m}\right)}} \cdot \exp\left[-\frac{1}{2}\left(\frac{\left([\mu_{j,k}]_b - \xi_{j,m}\right)^2}{[\sigma^2_{j,j}]_b + \sigma^2_{j,m}}\right)\right] \tag{3.15}$$

Inserting Eq. (3.15) into Eq. (3.7) leads to the following expression for the habitat

selection coefficients of the SDM:

$$\beta_{i,b} \;=\; \gamma_{i,0} + \sum_{j}\sum_{m} \delta_{i,j}^{(m)} [I_{j,m}]_b \tag{3.16}$$

A comparison with Eq. (2.6) shows that the new RBF-GFR model replaces $E\left[\mathbf{x}_j^m\right]_b$ from the original GFR model in (Matthiopoulos et al., 2011) by $[I_{j,m}]_b$:

$$[I_{j,m}]_b \;=\; \sum_k [w_k]_b \frac{\sigma_{j,m}}{\sqrt{\left([\sigma_{j,j}^2]_b + \sigma_{j,m}^2\right)}} \cdot \exp\left[-\frac{1}{2}\left(\frac{\left([\mu_{j,k}]_b - \xi_{j,m}\right)^2}{[\sigma_{j,j}^2]_b + \sigma_{j,m}^2}\right)\right] \tag{3.17}$$

Defining $\boldsymbol{z}$ slightly differently from before, namely composed of all elements in $\{x_i\}$ and $\{[I_{j,m}]_b\}$, and inserting Eq.(3.16) into Eq. (2.1), I get the following model of habitat preference:

$$h(\boldsymbol{z};\boldsymbol{\theta}) \;=\; \exp\left\{ \gamma_{0,0} + \sum_{j=1}^{I}\sum_{m=0}^{M_j} \delta_{0,j}^{(m)} [I_{j,m}]_b + \sum_{i=1}^{I}\left(\gamma_{i,0} + \sum_{j=1}^{I}\sum_{m=0}^{M_j} \delta_{i,j}^{(m)} [I_{j,m}]_b\right) x_i \right\} \tag{3.18}$$

The vector $\boldsymbol{z}$, which characterizes the habitat, is usually given a separate subscript, $\boldsymbol{z}_n$, where n denotes a particular plot or geographical patch where species counts are taken.

3.2.1 Testing the Derivation with Monte Carlo Approximation

Monte Carlo approximation is a technique that uses random numbers to approximate the distribution of a function (Murphy, 2012). The Monte Carlo approximation is used to approximate the expected value of any function by repeating random sampling, which is called Monte Carlo integration (Murphy, 2012). The expression in Eq. (3.15) is the result of solving the integral of Eq. (3.8) analytically. The Monte Carlo approach is used to check if the derivation of the integral of Eq. (3.8) is correct. If the numerical approximation of the integral of Eq. (3.8) using the Monte Carlo approach converges to the expression in Eq. (3.15), the expression in Eq. (3.15) is correct. The numerical approximation of

the integral is obtained by randomly selecting samples drawn from $N(\mathbf{x}|[\boldsymbol{\mu}_k]_b, [\mathbf{C}_k]_b)$ and taking the average of $\phi(x_j, \boldsymbol{\theta}_{j,m})$ of these samples, as follows:

$$\tilde{\psi}_N(\boldsymbol{\theta}_{j,m}, [\boldsymbol{\mu}_k]_b, [\mathbf{C}_k]_b) = \frac{1}{N}\sum_{n=1}^{N} \phi(x_j^{(n)}, \boldsymbol{\theta}_{j,m}) \tag{3.19}$$

where $\{x_j^{(n)}\}$ is a sample of the j^{th} element of N vectors drawn from $N(\mathbf{x}|[\boldsymbol{\mu}_k]_b, [\mathbf{C}_k]_b)$. Fig. 3.1 shows that $\tilde{\psi}_N(\boldsymbol{\theta}_{j,m}, [\boldsymbol{\mu}_k]_b, [\mathbf{C}_k]_b)$ converges to $\psi_N(\boldsymbol{\theta}_{j,m}, [\boldsymbol{\mu}_k]_b, [\mathbf{C}_k]_b)$ in Eq. (3.15) with an increasing number of samples N.

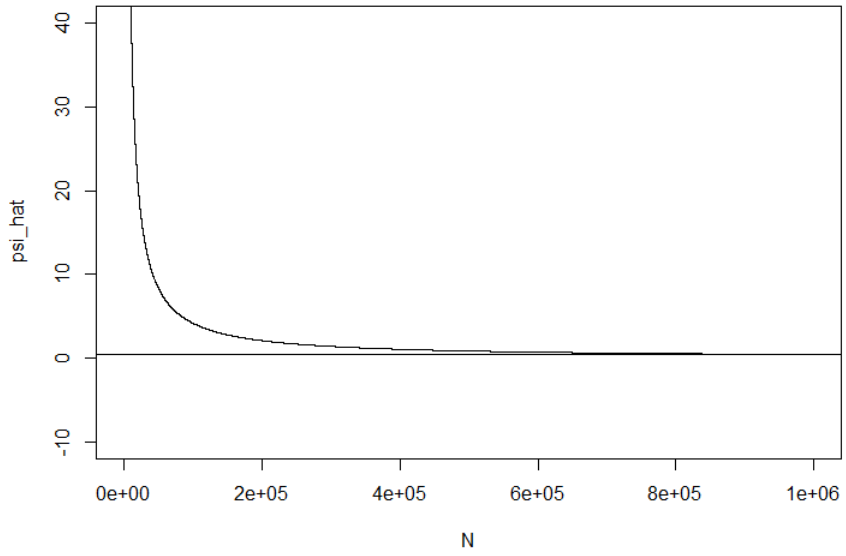


Figure 3.1: Monte Carlo approach to check the integral of Eq. (3.8). The curve is the function $\tilde{\psi}_N(\boldsymbol{\theta}_{j,m}, [\boldsymbol{\mu}_k]_b, [\mathbf{C}_k]_b)$ in Eq. (3.19), which converges to $\psi_N(\boldsymbol{\theta}_{j,m}, [\boldsymbol{\mu}_k]_b, [\mathbf{C}_k]_b) = 0.413$ from Eq. (3.15) (the straight line) with an increasing number of samples N.

3.2.2 RBF-GFR Model Parameters

For the RBF-GFR model, I need to decide on the number of Gaussian mixture components (see Eq. (3.4)) and the number of RBF basis functions, as seen from Eq. (3.3). I repeated the iterative optimization of the objective function from Eq. (3.18) for different choices of the number of RBF basis functions and then picked the one that minimized model selection scores (AIC, BIC). For the number of Gaussian mixture components, I repeated the iterative optimization of the objective function from Eq. (3.4) for different choices, picked the one that minimized model selection scores (AIC, BIC) for each block and then used the average of the number of components of all blocks as the optimal number of Gaussian mixture components. In some cases, the best number of Gaussian components based on AIC or BIC is close to the number of observations in the dataset, and the number of components is not allowed to be close or more than the data points. This is because each component will contain one point and cause singularity issues. Thus, the best number of components can be set to be less than half of the number of data points or by applying other methods which depend on the distance between components, such as k-means and silhouette methods. The parameters of the RBFs, $\xi_{j,m}$ and $\sigma_{j,m}$ in Eq. (3.3), need to be determined in advance to find $[I_{j,m}]_b$ in Eq. (3.17). A general discussion of how to set the RBF parameters can be found in Chapter 5 of Bishop (2006). In my own work, the following methods were applied to select these parameters and the best method was chosen based on AIC or BIC.

3.2.2.1 Histogram Approximation

A histogram approximation was used to approximate the habitat variable and get the parameters of the basis functions. The first method is carried out by constructing a histogram over the observations (cells) and then selecting the midpoints of the histogram bins to the centres $\xi_{j,m}$. Here, the $\sigma_{j,m}$ are determined as the differences between any two consecutive midpoints because the differences are the same and the bins are equally spaced. The number of bins is determined by Sturges' formula, which is based on the range of the data

(Batz et al., 2018; Sturges, 1926), as follows:

$$B = 1 + 3.322 \times log(n) \tag{3.20}$$

where B is the number of bins and n is the number of observations. Another way to determine the number of bins (basis functions) is by varying the number of bins and applying the RBF-GFR model using different numbers of basis functions and then picking the one that minimizes the BIC and AIC scores.

3.2.2.2 Quantile Approach

I set the centres of the basis functions $\xi_{j,m}$ of the jth environmental covariate to be the quantile of the jth environmental covariate. The bandwidth parameter $\sigma_{j,m}$ is the larger of the differences between the m-quantile and $(m-1)$-quantile and the difference between the $(m+1)$-quantile and m-quantile. For example, if the number of basis functions is 3, then the first quantile, median, and third quantile are the centres of the basis functions. Specifically, $\sigma_{j,2}$ is the larger of the difference between the first and second quantiles and the difference between the second and third quantiles. I applied the quantile approach to select the basis function parameters to ensure that most of the observations are included in the basis functions. The larger of the differences between the m-quantile and $(m-1)$-quantile and the difference between the $(m+1)$-quantile and m-quantile is selected as the bandwidth parameter of the basis functions to include most of the data in the basis functions. This process can be used for Gaussian and non-Gaussian datasets.

3.3 Calibration and Regularization

The original GFR and RBF-GFR models are types of GLMs, whose parameters can be estimated via maximum likelihood (ML). In the GFR and RBF-GFR models, the ML approach aims to maximize the likelihood function $L(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{y})$ where $\boldsymbol{\theta}$ is a parameter vector composed of the parameters $\gamma_{i,0}$ and $\delta_{i,j}^{(m)}$ in Eq. (2.6) for the GFR model and Eq. (3.18) for the RBF-GFR model. Note that $\boldsymbol{z}$ is a vector combining habitat variables x_i and either their expectation values, $E[X_i^m]$, or the derived quantities $I_{i,m}$ defined in Eq. (3.17). The

variables z_i are readily available from the observed data. On fitting a Gaussian mixture model to the explanatory data, I obtain the quantities $I_{i,m}$ from Eq. (3.17). The vector $\boldsymbol{y} = (y_1, ..., y_N)$ contains species observations, where N denotes the number of patches where species counts are taken. Depending on the study, the elements of this vector, y_n, can be binary use/availability indicators or count data. The equivalence between grid count, use-availability and point-process data has been demonstrated in the literature, on the basis of their corresponding likelihood functions (Aarts et al., 2012; Renner and Warton, 2013; Warton and Aarts, 2013; Warton and Shepherd, 2010). Although some of the datasets I used in this study took the form of binary (0/1) values, they were nevertheless equivalent to abundance models. A binary dataset is equivalent to this abundance model because it results in similar estimates of habitat preference to those obtained using models fitted to count data in discrete space (Aarts et al., 2012). My analyses stayed firmly in the area of abundance rather than occupancy models. Occupancy models (the recording of the presence of a species, regardless of its abundance) also result in binary data, but involve loss of information on abundance and although they are a widely used type of analysis (MacKenzie et al., 2017), they pose additional analytical challenges that were outside the remit of this thesis.

Parameter estimation with ML can be susceptible to over-fitting, especially for sparse and noisy data. This can be addressed with regularization, where a penalty term that quantifies model complexity is added to the log likelihood; see e.g., Section 3.2 in (Bishop, 2006). A particular form of regularization is ridge regression, where the size of the model's coefficients is penalized by maximizing a combined function that includes the weighted L2 norm of the parameter vector $\boldsymbol{\theta}$:

$$l(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{y}) - \lambda \|\boldsymbol{\theta}\|^2 \tag{3.21}$$

The weighting factor λ is a regularization parameter (see, e.g., Platt et al., 1999) that needs to be optimized. In the present work, I repeated the iterative optimization of the objective function in Eq. (3.21) for 100 discrete candidate values of λ chosen from an equidistant grid (Hastie et al., 2016) and then selected the value that minimizes the model selection scores: AIC and BIC. AIC may possibly disagree with BIC because the two

criteria apply different penalties based on the number of estimated parameters. However, they often agree. AIC always has a chance of choosing too large a model, regardless of the number of observations in the dataset. BIC might disagree with AIC if the number of observations is sufficiently large. If the AIC and BIC scores are in disagreement, I chose λ based on BIC because I want to have a less complex model and a more parsimonious model. For ridge regression, AIC and BIC computed using the actual number of parameters p will only be based on the log likelihood because changing lambda does affect the number of parameters when using L2 regularization. Thus, I used the effective number of parameters calculated in Eq. (2.59) rather than the actual number of parameters when computing AIC and BIC for ridge regression, as described in Section 2.8.3.

3.4 The GFR-CART & RBF-GFR-CART Models

The original GFR and RBF-GFR models are in the generalized linear model (GLM) family, which limits their modelling flexibility. One way to address this is to follow a divide-and-conquer strategy by recursively partitioning the input space into subregions, each modelled with a separate GLM. In what follows, I summarize how I have adapted them to the modelling of habitat preference.

Let $\Phi(\boldsymbol{z}, \boldsymbol{\theta}_k)$ denote one of the previous GLM-type models, Eq. (2.6) for the GFR model and Eq. (3.18) for the RBF-GFR model, with parameter vector $\boldsymbol{\theta}_k$ and input vector $\boldsymbol{z}$ (recall that this is a vector combining habitat variables x_i, expectation variables $\mathbb{E}(X_i)$ or $I_{i,m}$, as well as the product terms). The output of the tree is given by:

$$f(\boldsymbol{z}) \;=\; \sum_{k=1}^{K} \mathbb{I}(\boldsymbol{z} \in R_k)\Phi(\boldsymbol{z}, \boldsymbol{\theta}_k)$$

where $\mathbb{I}()$ is the indicator function, which is 1 if the argument is true and 0 otherwise, R_k is the region defined by the kth leaf node, and K is the number of leaf nodes in the tree. For the example in Figure 2.1, $K = 5$. The output of function $\Phi(\boldsymbol{z}, \boldsymbol{\theta}_k)$ depends on the application. For binary species presence/absence data, $\Phi(\boldsymbol{z}, \boldsymbol{\theta}_k) \in \{0, 1\}$, and we speak of a *classification* tree. For continuous species abundance data, $\Phi(\boldsymbol{z}, \boldsymbol{\theta}_k) \in \mathbb{R}^+$, and we have a

regression tree. The method is usually referred to as CART, described in Section 2.6. Finding the optimal partitioning of the data is an NP-hard problem, and it is therefore common to use a greedy iterative procedure, where in each iteration a new split node is introduced so as to optimize a local optimality criterion. This criterion is different for classification and regression trees. The classification cost is appropriate for species presence/absence data where the response variable determines whether an individual occurs at a location or not, while the regression cost is suitable for species abundance data where the response variable is the number of individuals in a location. I start with the former, i.e., I consider species presence/absence data first. The degree of uncertainty in the presence/absence status can be quantified with the entropy:

$$H(p) \;=\; -p\log(p) - (1-p)\log(1-p)$$

where $p = p(y=1)$ is the observed probability of detecting the species. It can be shown with standard calculus that $H(p)$ is maximized for $p = 0.5$, i.e., for maximal uncertainty, and minimized for $p \in \{0,1\}$, i.e., when there is no uncertainty, as seen in Section 2.6. The best split is then defined as the split that minimizes the uncertainty, maximizing the following difference:

$$H(y) - \frac{N_l}{N}H(y|z_j < t) - \frac{N_r}{N}H(y|z_j \geq t) = -\sum_{c=0}^{1} p(y=c)\log p(y=c)$$
$$+\frac{N_l}{N}\sum_{c=0}^{1} p(y=c|z_j < t)\log p(y=c|z_j < t) + \frac{N_r}{N}\sum_{c=0}^{1} p(y=c|z_j \geq t)\log p(y=c|z_j \geq t) \tag{3.22}$$

where $H(y)$ is the value of the entropy before the split, $H(y|z_j < t)$ is the entropy score of the left branch after the split using the threshold t, $H(y|z_j \geq t)$ is the right branch entropy score after the split, N_l is the number of data points in the left branch and N_r is the number of data points in the right branch, as illustrated in Fig. 3.2. The search is over all habitat variables $\{z_j\}$ and all valid threshold values t. Note that $p(y=c)$ is the probability of detecting the species in Eq. (2.7) or the probability of absence in Eq. (2.8), but the difference is that Eqs. (2.7) and (2.8) depend on the vector $\{\boldsymbol{z}\}$ of all explanatory

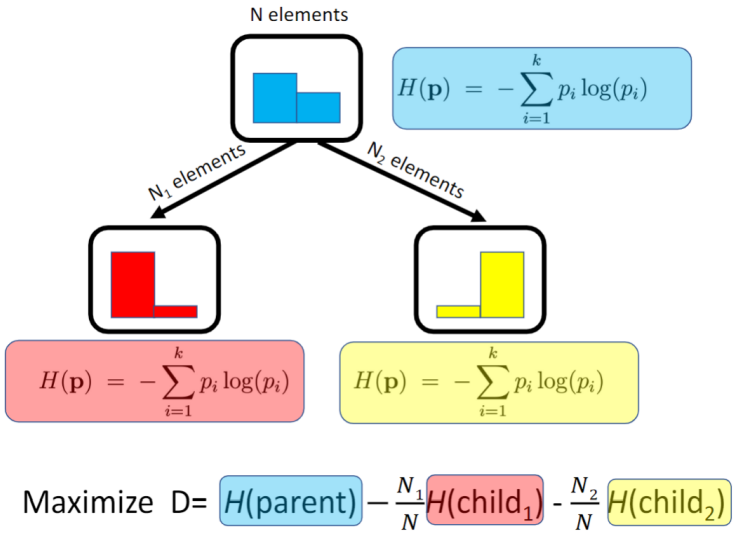


Figure 3.2: Illustration of the entropy criteria in Eq. (3.22) to select the best split of a node in CART. $H(parent)$ refers to the initial value of entropy before the split, whereas $H(child_1)$ and $H(child_2)$ are the left and right branch entropy scores, respectively.

variables and Eq. (3.22) depends on just one coordinate $\{z_j\}$ of the vector $\{\boldsymbol{z}\}$ above or below the threshold t. Conditional on the best split, each region of the habitat space is modelled with a separate GLM type model of one of the forms discussed above (GFR and RBG-GFR), whose parameters are optimized based on the maximum likelihood estimator (MLE), as described in Section 2.3. To prevent over-fitting, I can apply ten-fold cross-validation and stop growing the tree when the average objective function on the validation set starts to increase. In practice, it turns out to be better to grow a “full" tree, and then to use cross-validation to perform pruning. The cross-validation of pruning of the full tree works by allowing the model to grow to its full depth. Once the model grows to its full depth, tree branches are removed to prevent the model from overfitting by splitting the data into ten folds then training the model using nine folds and validating it using the remaining fold and repeating the process for all folds. Finally, the average square error is calculated for the full tree and all sub-trees and the tree that has the minimum average square error is chosen. The 1-SE rule is used to choose the optimal tree depth, i.e., the optimal number of nodes. The standard error is computed during the cross-validation, and

then the tree whose cross-validation error is within one standard error of the minimum error is chosen (Murphy, 2012; Breiman et al., 1984). Fig. 3.3 illustrates the process of pruning a tree using the 1-SE rule (Murphy, 2012). In the case that there are multiple trees within one standard error of the minimum cross-validation error, we will choose the smallest one since that tree would predict as well as the others, but it would also have fewer branches, which leads to a simpler explanatory model and further helps to avoid over-fitting (Therneau et al., 1997).

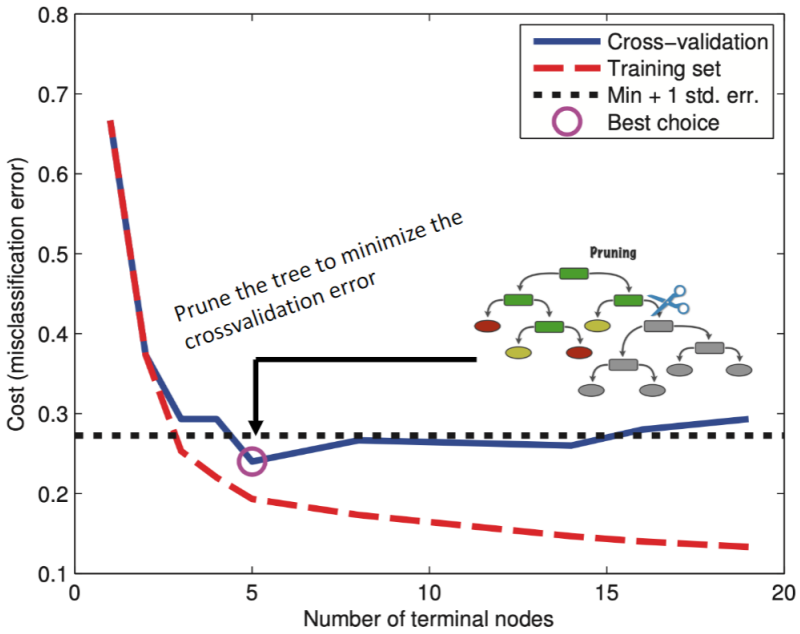


Figure 3.3: Illustration of the cross-validation pruning process. The y-axis is the cross-validation error vs the tree's depth in the x-axis. A model with five terminal nodes gives the least complex tree within 1SE of the minimal cross-validation error. Adapted from Figure 16.5 in Murphy (2012).

For modelling species abundance, the split function chooses the best habitat variable, and the best threshold value for this variable using the gain from before and after the split, as follows:

$$\text{cost}(\boldsymbol{z}_n, y_n, \boldsymbol{\theta}) - \left[\frac{N_l}{N}\text{cost}(z_{ni} < t, y_n, \boldsymbol{\theta}_l) + \frac{N_r}{N}\text{cost}(z_{ni} \geq t, y_n, \boldsymbol{\theta}_r)\right] \tag{3.23}$$

where t is a threshold parameter, i a habitat variable index, z_n are the augmented input vectors (combining habitat variables, their expectations, and cross products between both) and θ_l, θ_r are two separate model parameter vectors for the two separate input space regions (Murphy, 2012; Therneau et al., 1997). The cost function is the Poisson deviance defined as follows:

$$\text{cost}(z_n, y_n, \theta) = 2 \sum_{n=1}^{N} [y_n \log(\frac{y_n}{h(z_n, \hat{\theta})}) - (y_n - h(z_n, \hat{\theta}))] \tag{3.24}$$

where $h(z_n, \hat{\theta})$ is obtained from Eq. (2.6) for the original GFR model or Eq. (3.18) for the RBF-GFR model using the maximum likelihood algorithm to optimize θ using Eq. (2.12).

3.5 The GFR-RF & RBF-GFR-RF Models

We can combine classification and regression trees into a random forest. This is based on the insight that the expected out-of-sample prediction error can be decomposed into a bias and a variance component, that for a flexible model, like CART, the main contribution to this error comes from the variance term, and that this variance term can be reduced in a model ensemble, provided the models are sufficiently uncorrelated (see Section 2.7.1). In general, a model ensemble combines multiple individual models to build a predictive model.

To reduce the correlation between the individual members of the model ensemble, I follow the “bootstrapping and aggregating" procedure proposed by Breiman 2001, also called “bagging", whereby CART models are repeatedly trained on different independent bootstrap replicates, and then aggregated in a model ensemble, via voting (for classification) or averaging (for regression). To further decrease the correlation between the individual CART models, the split rules at the inner nodes of the trees are limited to randomly selected subsets of the features as candidate sets. For further details, see Section 2.7.1. In the present thesis, I propose a new variant of random forests, where each leaf node of the trees in the ensemble is a GFR or GFR-RBF model. Since the number of trees had to be selected, a baseline of 500 trees was set. This parameter is not particularly critical,

provided the number of trees is sufficiently large, and a value of 500 is widely used as a default (see e.g., Kassambara, 2018, Chapter 33). The OOB error variations become very small once the number of trees is sufficiently large. The greater the number of trees in a Random Forest Algorithm, the higher its accuracy. In practice, 500 trees is often a good choice. The disadvantages of having a large number of trees relate more to computational efficiency than to predictive performance. More trees may be needed if the data contain a multitude of features. The random forest is built using Algorithm 3 in Section 2.7.1, where each leaf in each tree was a separate original GFR model or RBF-GFR model; I refer to these models as the GFR-RF and RBF-GFR-RF models.

3.6 The GFR-XGBoost and RBF-GFR-XGBoost Models

An alternative to bagging is boosting, where the models in the ensemble are trained sequentially using a weighted form of the data, in which the weights depend on the previous model such that misclassified or poorly predicted instances receive greater weights (for further details, see Section 2.7.2). In the present thesis, the GFR and RBF-GFR models were used at each iteration of the XGBoost model.

While the ensemble size for bagging is not particularly critical, provided it is sufficiently large (500 is a widely used default value), it does matter for boosting. In XGBoost, large ensemble sizes can cause over-fitting because the gradient technique focuses on the most difficult cases, which can be due to noise. To avoid the over-fitting issue in XGBoost, I used a nested k-fold cross-validation scheme. I split each dataset into 3 subsets: the tuning set (k-2 folds), validation set (1 fold), and test set (1 fold). For each choice of the number of iterations {2, 5, 10, 15, 20, 40, 80, 100, 200, 300, 400, 500} and each fold, I trained the model on the tuning set and monitored the performance on the validation set by calculating the out-of-sample prediction accuracy and taking the median of k-1 folds. This gave me k medians for each number of iterations as explained in Algorithm 4 and shown in Fig. 3.4.

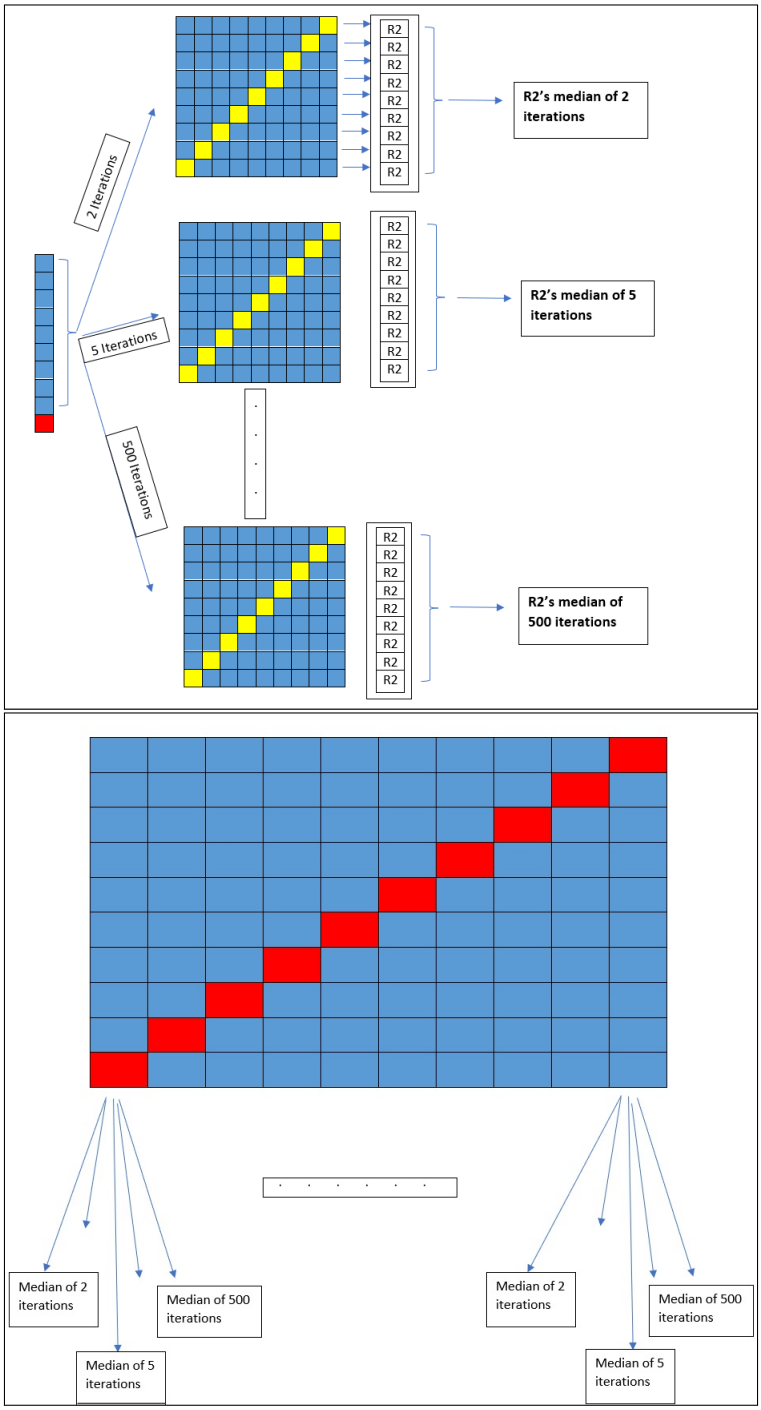


Figure 3.4: For each fold, the training set was split into a tuning set (blue boxes) and a validation set (yellow boxes), with the XGBoost model applied for each number of iterations {2, 5, 10, 15, 20, 40, 80, 100, 200, 300, 400, 500} and each fold, as shown in the top panel. The results after applying Algorithm 4 are given in the bottom panel.

Algorithm 4 Optimize the iteration number

procedure SPLIT THE DATASET INTO K FOLDS(*state*)
    for each k-1 folds **do**
        for each of number of iterations **do**
            Split the dataset to k-2 folds (tuning set), and 1 fold (validation dataset)
            Train the model using the tuning set and the number of iterations.
            Predict using the 1 fold validation set.
            Calculate the out-of-sample R^2.
        Calculate the median of k-1 out-of-sample R^2's for each number of iterations to pick the best number of iterations.
    return a matrix of medians for each fold (k folds) in rows and each number of iterations in columns.

3.7 Models Overview

The overview in Table (3.1) and Fig. 3.5 summarizes and outlines the relationship between all models proposed in the current study. The time to fit the random forest and XGBoost model depends on the dataset. These two model takes minutes to fit in the first simulated dataset but hours to fit in the second simulated dataset. The rest of the models take less than an hour to fit in all datasets.

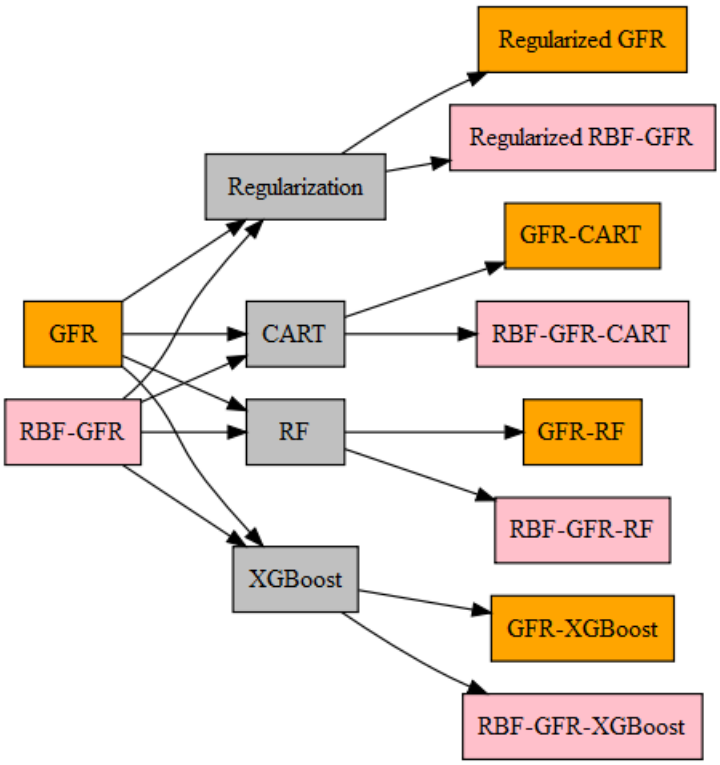


Figure 3.5: A diagram showing the relationship between all models proposed in this study. The orange boxes refer to the GFR model and the extensions of the GFR model while the pink boxes are the RBF-GFR model and its extensions. The gray boxes are the methods that were used to combine to the GFR and RBF-GFR models.

Table 3.1: Description of all models I used.

Models	Description
Standard	In Eq. (2.1), β_i values are fixed.
Original GFR	In Eq. (2.6), the polynomial function in Eq. (2.4) was used to represent β_i values in Eq.(2.1).
RBF-GFR	In Eq. (3.18), the radial basis function in Eq.(3.3) was used to represent β_i values in Eq.(2.1).
Regularized GFR	Ridge regression was used to estimate β_i values in the original GFR model in Eq. (2.6). The size of the GFR model's coefficients is penalized by maximizing a combined function that includes the weighted L2 norm of the parameter vector $\boldsymbol{\theta}$.
Regularized RBF-GFR	ridge regression was used to estimate β_i values in the RBF-GFR model in Eq. (3.18). The size of the RBF-GFR model's coefficients is penalized by maximizing a combined function that includes the weighted L2 norm of the parameter vector $\boldsymbol{\theta}$.
GFR-CART	The combination between the CART model and the GFR modelin Eq. (2.6), where the GFR model was used at each leaf of the tree by recursively partitioning the input space into subregions, each modelled with a separate GFR.
RBF-GFR-CART	The combination between the CART model and the RBF-GFR model in Eq. (3.18), where the RBF-GFR model was used at each leaf of the tree by recursively partitioning the input space into subregions, each modelled with a separate RBF-GFR.
GFR-RF	The combination between the RF algorithm and the GFR model in Eq. (2.6). The GFR model was used at each leaf of the tree in the RF model, where each tree is constructed independently from the other trees by randomly selecting the tree inputs and features.
RBF-GFR-RF	The combination between the RF algorithm and the RBF-GFR model in Eq. (3.18). The RBF-GFR model was used at each leaf of the tree in the RF model, where each tree is constructed independently from the other trees by randomly selecting the tree inputs and features.
GFR-XGBoost	The combination between the XGBoost algorithm and the GFR model in Eq. (2.6). The GFR model was used at each iteration of the XGBoost model, where each iteration aims to minimize the loss function and learn from the previous iteration by computing the second partial derivatives of the loss function.
RBF-GFR-XGBoost	The combination between the XGBoost algorithm and the RBF-GFR model in Eq. (3.18), where the RBF-GFR model was used at each iteration of the XGBoost model.

Chapter 4

Datasets

4.1 Introduction

Five distinct datasets are used in the present thesis. I used the same simulated datasets as in Matthiopoulos et al. (2015) and Matthiopoulos et al. (2011) that were used to apply the original GFR model. These individual-based simulated datasets were generated from multiple simulated instances using the resources and conditions as covariates. The simulated dataset in Matthiopoulos et al. (2015) is a complex version of the simulated dataset in Matthiopoulos et al. (2011), considering the population density of each sample in the simulation. These two simulated datasets are species abundance levels; the response variable is represented by the species abundance of each unit. The response variable in the sparrow and wolf datasets, which are real-life datasets, is a binary species use/availability indicator. I have applied all modelling approaches, as described in Table (3.1) and Fig. 3.5, to the simulated and real-life datasets, as seen in Chapter 5. The key reason for using the simulated datasets is that knowing the mechanisms that generate the data helps infer some of these mechanisms using a statistical model and then assess the model, which is the motivation of Chapter 6. The North American Breeding Bird Survey BBS dataset is another real-life dataset that was used as a count dataset, where each value of the response variable refers to the abundance of 10 different birds. The continental BBS dataset is a large-scale, Spatio-temporal dataset that suffers from the various inherent limitations dis-

cussed in Chapter 7. Still, it consists of multiple species data, which motivated me to use the GFR models as predictive models of biodiversity over this large spatial scale, as seen in Chapter 7.

4.2 Simulated Datasets

4.2.1 Simulated Dataset in Matthiopoulos et al. (2015)

Test data were derived from multiple simulated instances, representing subpopulations of a species living in different landscapes. This simulated data is species abundance levels; the response variable is represented by the species abundance of each unit. Each instance was obtained from a realisation of an individual-based simulation within a small (50x50 cell) spatial arena, where rudimentary energetics gave rise to simple demographic processes and population dynamics. Two spatially autocorrelated environmental covariates (food - a resource - and temperature - a condition) were distributed across the arena. Individuals were programmed to move up gradients of environmental profitability (i.e., food richness moderated by temperature) and their movement was subject to perception error. The population size (N) associated with the entire sampling instance was also included to capture density-dependent effects on the distribution of the animals.

The dataset contains 20 landscape scenarios whose dynamics were modelled for 20 years, yielding 400 different sampling instances, where the dataset has a spatial structure based on these scenarios. Each instance has potentially 2500 spatial observations (50×50 grid). The total sample size of the data is 200,000 spatial cells. Different subsets of this dataset can be used to emulate realistic scenarios of sample size across time and space.

4.2.2 Simulated Dataset from Matthiopoulos et al. (2011)

A simpler version of the above individual-based model (IBM), which simulates populations based on individuals and their properties, looks at two resources (e.g., food and cover) required in alternation, without the effect of demography and population dynamics. The response variable is represented by the species abundance of each unit. The simulated

animals in this version of the individual-based model climb up gradients of food when they are hungry and then climb up gradients of cover when they are sated. Feeding occurs through a Holling type II model of food consumption in which the organism reaches a maximum consumption rate as the food increases because the animal can no longer process more food per unit of time. In trophic ecology, this is called a Holling type II functional response (Matthiopoulos, 2011). These simulated data contain 20 scenarios, and each sampling instance can provide a maximum of 2500 observations, where the total sample size is 50,000 units. The animal simulation step, which generates a usage map, is given certain movement rules in this simulation process. The animal accumulated energy (E) through the consumption of food (u). The rate of food consumption as a function of food abundance was calculated using the Holling type II functional response model (Holling, 1959). The energy balance equation of the animal at time t is calculated as follows:

$$E_t = E_{t-1} + feeding_t - den$$

where den is the metabolic cost; the amount of energy consumed. $feeding_t$ is defined as the amount of food consumed per time unit. The energy at time t depends on the amount of food consumed at time t, the energy that remains from the previous time and the amount of energy consumed at the same time. The feeding rate is a function of the amount of food available calculated using type II functional response that considers saturation as follows:

$$E_t = E_{t-1} + \frac{a \times food_{x(t)}}{b + food_{x(t)}} - den$$

where a is the feeding rate or maximum consumption rate that the organism can reach, b is the half-saturation parameter $a/2$, which describes how fast the value of the function increases, and $food_{x(t)}$ is the amount of available food at time t (Matthiopoulos, 2011). In this simulation, upon satiation ($E > E_1$), the animal stopped feeding and climbed up the gradient of cover until reaching a local maximum. When E fell below a starvation threshold (E_2), the animal climbed up the food gradient until reaching a local maximum.

4.3 Real-Life Datasets

4.3.1 Sparrow Population Dataset

The sparrow population data are used in (Matthiopoulos et al., 2019) when they aimed to prove that population change can be predicted based on habitat availability by using the GFR model. The data were collected by the Royal Society for the Protection of Birds (RSPB) and the University of Glasgow in 2014 during the breeding season from 32 colonies in the United Kingdom. Each colony contains 40 spatial cells, which means that the data contains 1280 different cells. The *Sparrow* variable in the data is the response variable presented by values of 1 and 0 based on the presence of sparrows in each cell. I use three main variables in the dataset, which are the estimated percentages of *grass, bush* and *roof* for each cell captured by Google Earth. The response variables for this data set are binary species presence/absence indicators, the binomial log likelihood in Eq. (2.10) was used when fitting the original GFR model and the RBF-GFR model to the data using the habitat variables *grass, bush*, and *roof* as the main covariates in the models and size of each colony, which is the count of the maximum number of males measured in each colony, as an additional explanatory variable.

4.3.2 Wolf Dataset

The wolf dataset is a telemetry dataset that was used in Mathiopoulos et al. (2011) to fit the original GFR model. The telemetry data are different from the survey datasets because the observations are collected over time (Matthiopoulos et al., 2020a). The wolf data comprise a telemetry dataset incorporating a use-availability approach to determine the response variable, where the response variable is represented by either 0 or 1. These data consist of 11 wolves, which are members of five different packs. The wolf dataset has a grouping structure based on the five packs that the wolves belong to. The dataset consists of continuous and factor variables. There are three continuous variables, which are as follows: *distance to high human use, distance to edge and slope*. The factor levels are the landcover types: *burnt, alpine, shrub, rock and herbaceous*. The sample size of the data is 18,042 spatial units.

4.3.3 BBS Dataset

The North American Breeding Bird Survey (BBS) dataset was used in Haddou et al. (2022) to measure past and current landscape contributions to the current effective number of species. The BBS dataset was instituted in 1966 to monitor the trends in abundance of more than 400 different bird species. Data collection occurs in June during the breeding seasons over more than 3000 routes. Each route is about 40 kilometres long. Making 3min stops every 800 metres (yielding 50 stops per route), the observers record every seen or heard bird. This data is species abundance levels, where the response variable is represented by the species abundance of each stop. The dataset used in the current study is from 2001 to 2019 because the land cover covariates (*urban, forest, grass, crop, wet, water and elevation*) are available for those years in the open-access NLCD CONUS (Yang et al., 2018). The temperature covariate for this period was taken from the PRISM climate dataset (PRISM, 2019). I used the abundance of 10 different birds from more widely distributed species: mourning dove (*Zenaida macroura*), American robin (*Turdus migratorius*), red-winged blackbird (*Agelaius phoeniceus*), American crow (*Corvus brachyrhynchos*), barn swallow (*Hirundo rustica*), brown-headed cowbird (*Molothrus ater*), European starling (*Sturnus vulgaris*), chipping sparrow (*Spizella passerina*), blue jay (*Cyanocitta cristata*), and common yellowthroat (*Geothlypis trichas*). First, to reduce the spatial autocorrelation between the stop points segments and avoid overlaps between these segments in the analysis, across the 50 stops, the 1^{st}, 11^{st}, 21^{st}, 31^{st}, and 41^{st} stop counts for the route were used, where each stop is a representative of about 8 km transects for the landscape surrounding it within a 400 m buffer, as seen in Fig. 4.1. This process is a point process and is represented as points randomly located in space.

The land cover covariates are the percentage of the *forest, grass, urban, crop, wet, water*, and *barren* within a 400 meters buffer around the segment from which bird abundances were taken. These covariates have been used as habitat features and to represent habitat availability with total of 134,275 spatial cells.

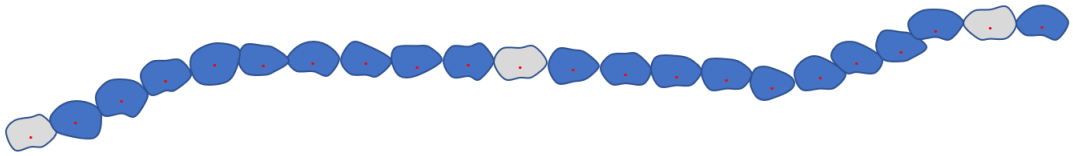


Figure 4.1: A diagram explaining the BBS data pre-processing showing the first 22 stop points of a route in the data. The red points are the stop points and the blue polygons are the 400 m buffer around each stop point. Here, 400 is the size in metres of the radius for which the landscape was sampled around each segment. The grey polygons are the stop points I used in the present study, which are the 1^{st}, 11^{st}, and 21^{st} stop points for the first 22 stop points of the route.

4.4 Datasets Outlines

Table 4.1 outlines the datasets used in the following chapters. The sample instances number refers to the number of sample instances (scenarios) in which a sampling instance represents an environmental scenario defined in a biological way as the environment experienced by the study animals during an appropriate spatiotemporal frame of accessibility. The two simulated, wolf and sparrow datasets are used to apply the original GFR, RBF-GFR and various extension models to assess and compare the predictions of species distributions resulting from the models in Chapter 5. The simulated dataset in Matthiopoulos et al. (2011) is used in Chapter 6 to focus on explanatory modelling for testing the models' transferability. The result of the predictions of species distributions and biodiversity of the GFR models as applied to the BBS dataset are summarized in Chapter 7.

Table 4.1: Overview table of the datasets, the habitat variables, number of sample instances, and data size.

Dataset	Habitat variables	Type of response variable	Sample instances number	Data size
Simulated dataset in Matthiopoulos et al. (2015)	*Food, temperature*, and *population size*	species abundance	400	200,000
Simulated dataset in Matthiopoulos et al. (2011)	*Food*, and *cover*	species abundance	20	50,000
Sparrow	*Grass, bush, roof*, and *colony size*	presence/absence	32	1,280
Wolf	*Distance to high human use, distance to edge, slope, burnt, alpine, shrub, rock*, and *herbaceous*	use-availability	5	18,042
BBS	*urban, forest, grass, crop, wet, water, elevation*, and *temperature*	species abundance	10	134,275

Chapter 5

Using GFRs for Robust Predictions of Species Distributions

5.1 Introduction

In the first part of the comparative model assessment study, the original GFR model, which was reviewed in Section 2.2, was compared with the proposed RBF-GFR model, which was described in Section 3.2, the combination of GFR and RBF-GFR with the CART model, which was described in Section 3.4, the combination of GFR and RBF-GFR with the RF model, which was described in Section 3.5, and the combination of GFR and RBF-GFR with the XGBoost model, which was described in Section 3.6. I compared the test set accuracies, which have been quantified in terms of out-of-sample R^2 scores, and split the presentation of the results by the four datasets (simulated dataset in Matthiopoulos et al. (2015), simulated dataset in Matthiopoulos et al. (2011), sparrow and wolf datasets), as shown in Section 5.2. Section 5.3 shows that model ensembles can perform the same role as regularization. The comparison between the two ensemble models combined with the GFR and RBF-GFR models is discussed in Section 5.4. Section 5.5 shows how regularization is relevant to the second simulated dataset and is not needed for the other datasets. To summarize the comparative model assessments study, the models were then ranked by performance for each of the four datasets in turn, generating a 'league table' of models.

Thus, this table offers the ranks of all the models included in the study, as shown in Section 5.6. Based on the need to assess the quality of the predictions, the predictions of animal habitat usage derived from the models were presented in spatial maps, and further visualisations of species abundance have been generated in Section 5.7. Section 5.8 provides interpretations of which explanatory variables are important in the real-life applications.

5.2 The Datasets' Results on Predicting Species Abundance from Habitat Variables

The GFR and RBF-GFR models depend on different complexity parameters. For the GFR model, it is necessary to define the polynomial order, as seen from Eq. (2.5). For the RBF-GFR model, I had to decide on the number of Gaussian mixture components, as indicated in Eq. (3.4), and the number of RBF basis functions, as seen in Eq. (3.3). I repeated the iterative optimization of the objective function from Eq. (3.18) for different choices of the number of RBF basis functions before, and then picked the one that minimized the model selection score (BIC). For the number of Gaussian mixture components, I found the number of components that minimize the BIC score for each block, and then used the average number of components of all blocks as the optimal number of Gaussian mixture components for the RBF-GFR model and its extensions, as shown in Appendix A.3. The parameters of the RBFs, $\xi_{j,m}$ and $\sigma_{j,m}$, need to be determined in advance to find $[I_{j,m}]_b$ in Eq. (3.17). I used the histogram approximation and quantile approaches, discussed in Section 3.2.2, to select these parameters and the best method (quantile approach) was chosen based on AIC and BIC, as seen in Appendix A.2.

For the best parameters to be selected, I compared the test set accuracies, which have been quantified in terms of out-of-sample R-square scores, for the original GFR and the proposed RBF-GFR model. The CART and RF models were combined with the original GFR and RBF-GFR models, and all models were then compared against the original GFR and RBF-GFR models, as described in Table 3.1. A standard CART algorithm, where each leaf is a separate GFR model or RBF-GFR model, was applied in each case, and the cost function, discussed in Section 3.4, was used to grow the tree and find the best split variable

for each iteration of the optimization algorithm. The tree was then pruned using 10-fold cross-validation based on the training set. For the RF model, the number of trees had to be selected, as described in Section 3.5. In this case, a baseline of 500 trees was set, where each leaf in each tree was a separate original GFR model or RBF-GFR model.

The XGBoost model was used in combination with the original GFR and RBF-GFR models over several different numbers of iterations {2, 5, 10, 15, 20, 40, 80, 100, 200, 300, 400, 500}; Algorithm 4 in Section 3.6 was then used to determine the best number of iterations of XGBoost for use in all subsequent applications.

5.2.1 Results of the First Simulated Dataset

The RBF-GFR model applied to the first simulated dataset using $food$ and $temperature$ as the main covariates and population size N as an additional explanatory variable is as follow:

$$\begin{aligned} use \;=\; \exp\{&\gamma_{0,0}+\gamma_{1,0}food+\gamma_{2,0}temp+\gamma_{3,0}temp2+\gamma_{4,0}N+\sum_{m=1}^{M}\delta_{0,1}^{(m)}I_{food,m}+\\ &\sum_{m=1}^{M}\delta_{0,2}^{(m)}I_{temp,m}+\sum_{m=1}^{M}\delta_{1,1}^{(m)}(food\cdot I_{food,m})+\sum_{m=1}^{M}\delta_{1,2}^{(m)}(food\cdot I_{temp,m})+\\ &\sum_{m=1}^{M}\delta_{2,1}^{(m)}(temp\cdot I_{food,m})+\sum_{m=1}^{M}\delta_{2,2}^{(m)}(temp\cdot I_{temp,m})+\sum_{m=1}^{M}\delta_{3,1}^{(m)}(temp2\cdot I_{food,m})+\\ &\sum_{m=1}^{M}\delta_{3,2}^{(m)}(temp2\cdot I_{temp,m})+\delta_{4,1}(food\cdot N)+\delta_{4,2}(temp\cdot N)+\delta_{4,3}(temp2\cdot N)\} \end{aligned} \tag{5.1}$$

where M is the best number of basis functions in the RBF-GFR model, temp2 is a quadratic main effect for temperature, $\gamma_{i,0}$ is the intercept that does not depend on changes in availability for the ith covariate, and $\delta_{i,j}^{(m)}$ is the coefficient of $\gamma_i(\mathbf{x})$ for the m^{th} basis function of the j^{th} variable. The performance of the original GFR and RBF-GFR models were evaluated using the maximum likelihood function described in Section 2.3. M was varied from 1 to 12, a process was used to determine the best polynomial order for the original GFR, per Eq. (2.5), and the best number of basis functions in the RBF-GFR model in Eq. (3.3) was set to 10 based on model selection scores as shown in Fig. 5.1. On varying

the number of Gaussian mixture components from 1 to 100, 9 Gaussian mixture components was selected as the best number of components for availability approximation based on the model selection score, Bayesian information criterion (BIC) as seen in Appendix A.3. The training set for the first set of simulated data contained 90% of scenarios, with

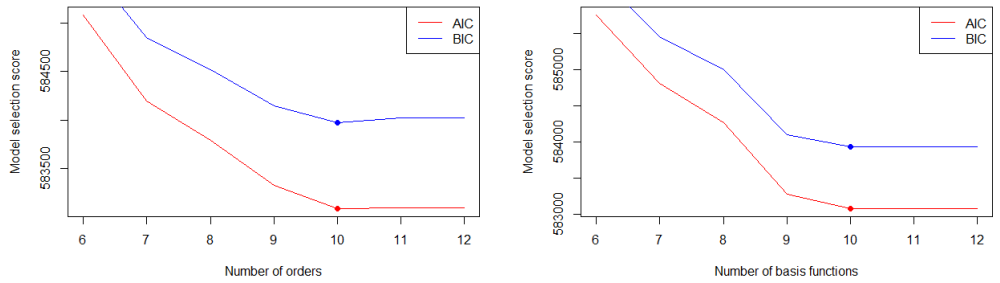


Figure 5.1: Optimization of the polynomial order number using model selection scores for the original GFR model (left panel) and the RBF-GFR model (right panel) as applied to the first simulated dataset. The two points refer to the best number of orders and basis functions based on AIC and BIC. The best polynomial order for the original GFR and the best number of basis functions in the RBF-GFR model is 10, which is the minimum for both the AIC and BIC scores.

360 sample instances, while the test set consisted of 10% of the scenarios, with 40 sample instances, for both models. The out-of-sample performance score was calculated using Eq. (2.60) with a 10-fold cross-validation test for both models, with the standard model in Eq. (2.1) for comparison. The results for the median out-of-sample $R^2 \pm$ MAD are shown in Table 5.1.

The first dataset is a large dataset that contains a large number of sample instances, as discussed in Section 4.2.1. As setting 500 trees for a large-size dataset is computationally expensive, the number of trees, in that case, was set to 15. The habitat usage of the test set was then predicted by the GFR-CART, RBF-GFR-CART, GFR-RF and RBF-GFR-RF models to measure the out-of-sample prediction scores, as presented in Table 5.2.

No over-fitting problems occurred in the first simulated dataset, and the out-of-sample R^2 score in Eq. (2.60) did not decrease when the number of XGBoost iterations increased as shown in Fig. 5.3. Thus, I decided to run the algorithm for 500 iterations.

Table 5.1: Median of out-of-sample performance scores of the standard, original GFR and RBF-GFR models described in Table 3.1 applied to the first simulated dataset.

Orders	Standard ($R^2 \pm$ MAD)	Original GFR ($R^2 \pm$ MAD)	RBF-GFR ($R^2 \pm$ MAD)
1	0.731 ± 0.026	0.768 ± 0.021	0.760 ± 0.080
2	0.731 ± 0.026	0.813 ± 0.011	0.813 ± 0.018
3	0.731 ± 0.026	0.818 ± 0.012	0.822 ± 0.009
4	0.731 ± 0.026	0.825 ± 0.010	0.822 ± 0.010
5	0.731 ± 0.026	0.829 ± 0.012	0.824 ± 0.010
6	0.731 ± 0.026	0.830 ± 0.008	0.828 ± 0.010
7	0.731 ± 0.026	0.831 ± 0.009	0.835 ± 0.013
8	0.731 ± 0.026	0.834 ± 0.012	0.837 ± 0.013
9	0.731 ± 0.026	0.835 ± 0.013	0.837 ± 0.012
10	0.731 ± 0.026	0.837 ± 0.014	0.837 ± 0.014
11	0.731 ± 0.026	0.837 ± 0.014	0.837 ± 0.014
12	0.731 ± 0.026	0.837 ± 0.014	0.837 ± 0.014

Table 5.2: Median of out-of-sample performance scores for the original GFR and RBF-GFR models in combination with the CART, RF and XGBoost models using the first simulated dataset. The scores for the original GFR and RBF-GFR models are provided for comparison.

Models	R^2 (order or basis)	R^2 (CART)	R^2 (RFs)	R^2 (XGBoost)
Original GFR	0.837 ± 0.014	0.821 ± 0.006	0.936 ± 0.008	0.944 ± 0.012
RBF-GFR	0.837 ± 0.014	0.822 ± 0.007	0.937 ± 0.010	0.941 ± 0.011

Both the GFR and RBF-GFR models outperform the simple SDM model when applied to the first simulated dataset in terms of predictive performance. There is no significant difference between the performance resulting from the GFR model and that resulting from the RBF-GFR model. The forecasting performance of the RF and XGBoost in combination with the original GFR and RBF-GFR models outperformed that of the original and RBF-GFR models.

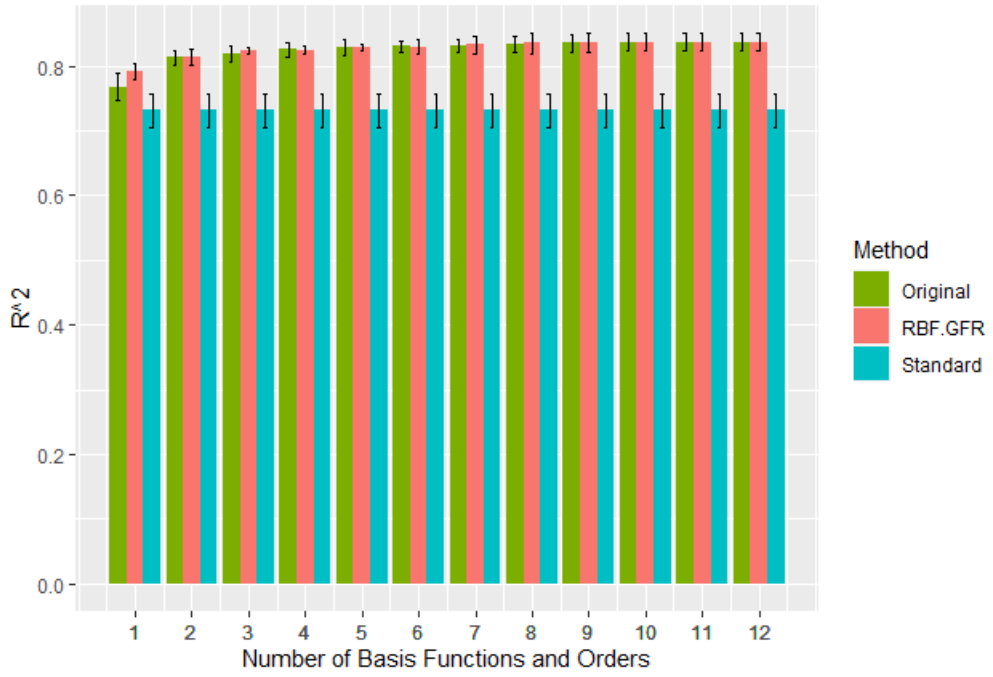


Figure 5.2: Comparison of performance score for the RBF-GFR and original GFR approaches on the first simulated model; bars are the $\pm$ MAD.

5.2.1.1 Model Checking

I assumed that the response variable for this simulated dataset follows the Poisson distribution because it is species abundance levels. In the Poisson distribution, the mean and variance are mathematically exactly the same. However, the mean and variance for the response variable of the dataset are very different (the mean is 3.4 and the variance is 11.6). A chi-squared goodness of fit test can be used to test the hypothesis that observed data follow a particular distribution. So, I used it here to observe if the response variable follows the Poisson distribution. Here, the null and alternative hypotheses for the chi-square goodness of fit test are the following:
Null: The data follow the Poisson distribution.
Alternative: The data do not follow the Poisson distribution.
When the p-value for the chi-square goodness of fit test is less than the significance level, I reject the null hypothesis. I observed that the p-value is 0, so I reject the null hypothesis;

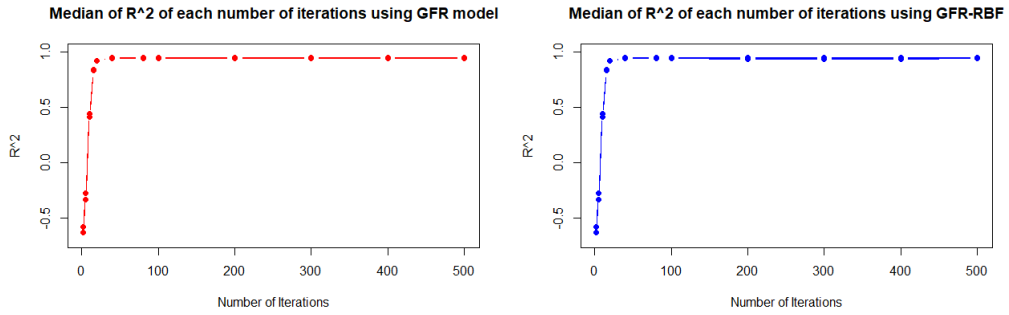


Figure 5.3: Optimization of iteration number using Algorithm 4 on a combination of the XGBoost with the original GFR (left panel) and XGBoost using RBF-GFR (right panel) as applied to the first simulated dataset.

the data do not follow the Poisson process. Therefore, I applied the negative binomial model, which is a more flexible model that is similar to the Poisson model but incorporates an additional term to account for the excess variance. The negative binomial model can be used when the variance is substantially higher than the mean. Unlike the Poisson distribution, the variance of the negative binomial distribution is a function of its mean as follows:

$$\sigma^2 = \mu + \frac{\mu^2}{k} \tag{5.2}$$

where k is the dispersion parameter. Here, there are no substantial differences in either models' predictive performance when assuming the negative binomial distribution or the Poisson distribution using the original GFR, RBF-GFR, regularized GFR and regularized RBF-GFR models, as seen in Table 5.3. The predictive performance of the GFR, RBF-GFR, regularized GFR and regularized RBF-GFR models assuming negative binomial gives a slight improvement, but it's not a massive improvement. Therefore, I assume the predictive performance does not change much for the more flexible models, such as CART, RF and XGBoost. I fitted the model in the R program using the library glm.nb. The glm.nb function uses the ML estimate of the dispersion parameter k. There is a possibility that this standard library does not do an appropriate maximization for the dispersion parameter, and it might be a local optimum. In future work, it is useful to try an alternative library, such

as feglm.nb. No substantial differences were evident between the Poisson and Negative Binomial Distributions assumption; therefore, I continued to use the Poisson distribution for the count datasets. Some diagnostic plots can be found in Appendix A.4.

Table 5.3: Median of out-of-sample performance scores for the original GFR, RBF-GFR regularized GFR and regularized RBF-GFR models using the first simulated dataset assuming Poisson and Negative binomial distributions.

Distribution	R^2 (GFR)	R^2 (RBF-GFR)	R^2 (Reg-GFR)	R^2 (Reg-RBF-GFR)
Poisson	0.837	0.837	0.796	0.796
Negative binomial	0.841	0.841	0.798	0.796

5.2.2 Results of the Second Simulated Dataset

For the second simulated dataset, the effect of the training set size was investigated across various scenarios using the rules developed by Matthiopoulos et al. (2011), as described in Section 4.2.2. The number of scenarios was set to 20 sites. The best number of polynomial orders and basis functions was selected as 10, as seen in Fig. 5.4, with the best number of Gaussian mixture components being 24 based on BIC, as discussed in Appendix A.3.

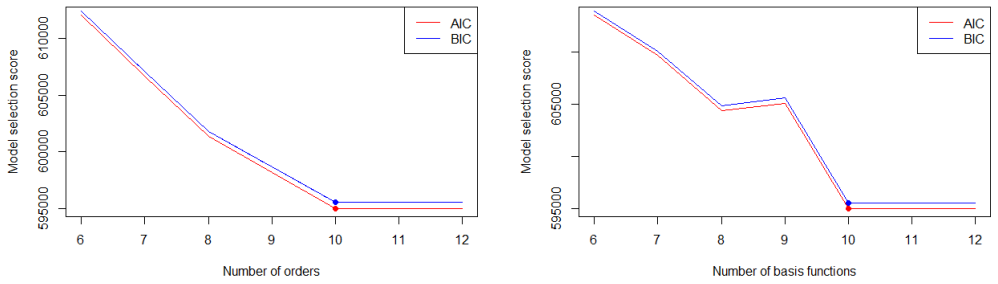


Figure 5.4: Optimization of the polynomial order number using model selection scores for the original GFR model (left panel) and the RBF-GFR model (right panel) as applied to the second simulated dataset. The two points refer to the best number of orders and basis functions based on AIC and BIC. The best number of polynomial orders and basis functions was selected as 10 based on both AIC and BIC.

Estimating the original GFR and RBF-GFR model parameters using a maximum likelihood estimator as described in Section 2.3 did not provide good forecasting performance scores, however, especially when the number of basis functions and orders increased, based on the 20-fold cross-validation test of out-of-sample R^2 for both models as shown in Table 5.4. The out-of-sample R^2 is a model comparison of the prediction model with a baseline prediction model. Here, the negative value means the prediction tends to be less accurate than the average value of the data, which is a result of overfitting and led me to regularize the model using ridge regression.

Table 5.4: Out-of-sample performance scores of the original GFR and RBF-GFR models for the second simulated dataset.

Number of orders	Original GFR ($R^2 \pm$ MAD)	RBF-GFR ($R^2 \pm$ MAD)
1	0.434 $\pm$ 0.213	0.297$\pm$ 0.199
2	0.472 $\pm$ 0.278	0.391 $\pm$ 0.371
3	0.543 $\pm$ 0.316	0.506$\pm$ 0.296
4	0.537 $\pm$ 0.334	0.597$\pm$ 0.173
5	0.593 $\pm$ 0.268	0.521 $\pm$ 0.196
6	0.573 $\pm$ 0.260	0.597 $\pm$ 0.139
7	0.450 $\pm$ 0.417	0.521 $\pm$ 0.386
8	0.300$\pm$ 0.502	0.459$\pm$ 0.453
9	-3.43 $\pm$ 5.770	-1.44$\pm$ 3.03
10	-2.223 $\pm$ 2.212	-2.34 $\pm$ 3.65
11	-1.691 $\pm$ 1.678	-1.68 $\pm$ 2.76
12	-1.691 $\pm$ 1.678	-1.76 $\pm$ 3.39

Ridge regression discussed in Section 3.3 was applied using Eq. (3.21) to both models on the second simulated data to improve the performance of their forecasting ability. The results of R^2 for both models using λ that gave the lowest BIC in the training set are listed in Table 5.5 and Fig. 5.5. Recall that the information criteria BIC is a function of the effective number of parameters, which depend on the regularization parameter λ as explained in Section 2.8.3 and illustrated in Fig. 2.4.

A comparison with Table 5.4 shows that the inclusion of L2 regularization created a general performance improvement.

Table 5.5: Median of R^2 for 20-fold cross-validation for different models using regularization where BIC is used to choose λ for the second simulated dataset (20 scenarios).

Basis or orders	Standard ($R^2 \pm$ MAD)	Original ($R^2 \pm$ MAD)	RBF-GFR ($R^2 \pm$ MAD)
1	0.256 ± 0.179	0.425 ± 0.262	0.201 ± 0.201
2	0.256 ± 0.179	0.426 ± 0.256	0.379 ± 0.213
3	0.256 ± 0.179	0.517 ± 0.290	0.492 ± 0.174
4	0.256 ± 0.179	0.397 ± 0.307	0.542 ± 0.148
5	0.256 ± 0.179	0.389 ± 0.332	0.566 ± 0.121
6	0.256 ± 0.179	0.392 ± 0.316	0.611 ± 0.158
7	0.256 ± 0.179	0.396 ± 0.307	0.617 ± 0.161
8	0.256 ± 0.179	0.358 ± 0.311	0.619 ± 0.136
9	0.256 ± 0.179	0.358 ± 0.155	0.633 ± 0.127
10	0.256 ± 0.179	0.359 ± 0.341	0.635 ± 0.147
11	0.256 ± 0.179	0.359 ± 0.337	0.640 ± 0.148
12	0.256 ± 0.179	0.358 ± 0.337	0.631 ± 0.153

The CART and RF models were used in combination with the RBF-GFR and original GFR models to improve forecasting performance for the second simulated dataset under 20 scenarios. Table 5.6 shows that the CART and RF outperformed the original GFR and RBF-GFR models.

Table 5.6: Median of out-of-sample performance scores for the original and RBF-GFR model and the CART, RF and XGBoost models in combination with the GFR model using the second simulated dataset.

Models	R^2 (order or basis)	R^2 (CART)	R^2 (RFs)	R^2 (XGBoost)
Original GFR	-0.972 ± 1.77	0.235 ± 0.196	0.443 ± 0.396	0.491 ± 0.192
RBF-GFR	-2.35 ± 3.65	0.440 ± 0.225	0.593 ± 0.107	0.535 ± 0.10

500 XGBoost iterations were used because the cross-validation R^2 score increased as the iteration number increased, as shown in Fig. 5.6. No over-fitting problems occurred in the second simulated dataset.

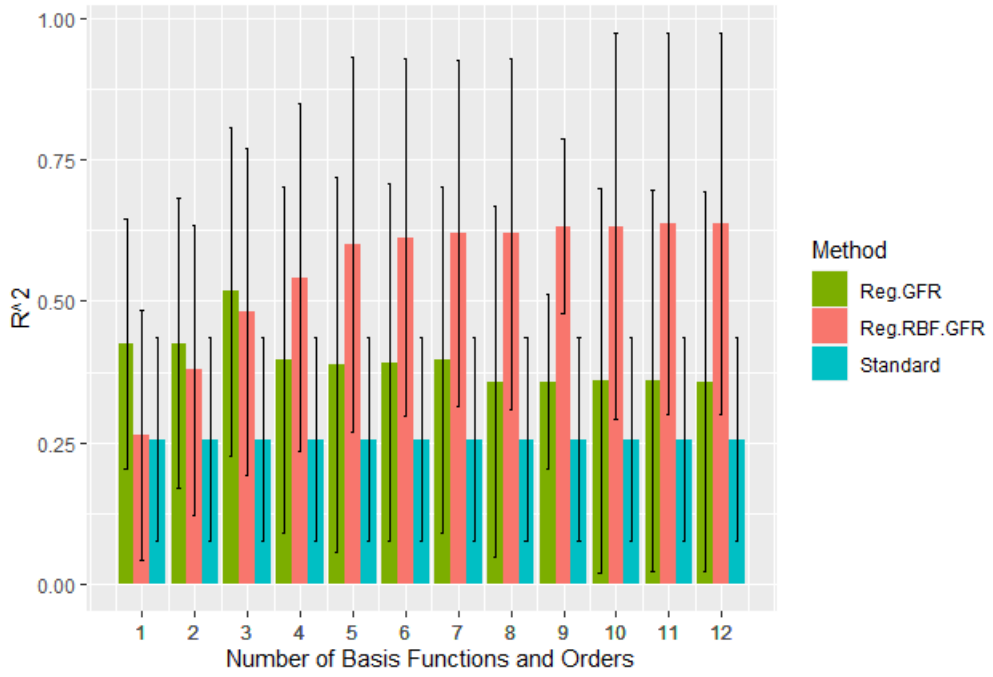


Figure 5.5: Comparison of performance score for the standard, regularized GFR and regularized RBF-GFR models applying to the second simulated dataset; bars are $\pm$ MAD.

5.2.3 Results of the Sparrow Dataset

For the sparrow dataset, the AIC and BIC scores are not consistent as the best number of basis functions in the RBF-GFR model based on the AIC score is three and the BIC score is one. However, the difference between the scores is small, as shown in Fig. 5.7, so the first and third order polynomial model and one and three radial basis functions were represented, first order and one basis functions model here and third order and three basis function models in Appendix A.5 Table A.3.

The best number of Gaussian components for the RBF-GFR model was 39. However, each colony consists of 40 data points, and the number of components is not allowed to be close or more than the data points. This is because each component will contain one point and cause singularity issues. Thus, the best number of components was set to 18 components as it is less than half of the number of data points in each colony. Table 5.7 presents the model selection scores for the original GFR and RBF-GFR models, using the

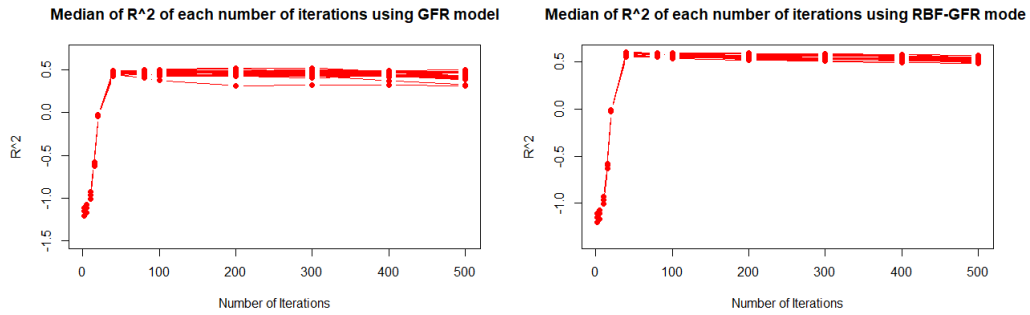


Figure 5.6: Optimization of iteration number using Algorithm 4 on a combination of the XGBoost with the original GFR (left panel) and XGBoost using RBF-GFR (right panel) as applied to the second simulated dataset.

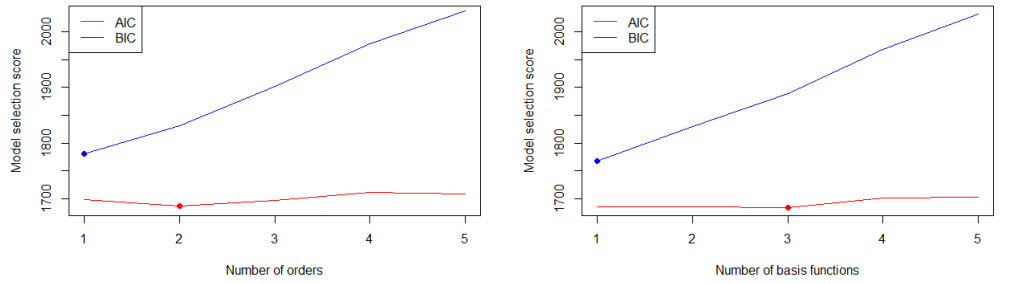


Figure 5.7: Optimization of the polynomial order number using model selection scores for the original GFR model (left panel) and the RBF-GFR model (right panel) as applied to the sparrow population dataset. The two points refer to the best number of orders and basis functions based on AIC and BIC. The AIC and BIC scores are not consistent as the best number of basis functions in the RBF-GFR model based on the AIC score is three and the BIC score is one.

percentage of *grass*, *bush*, and *roof* as covariates, in addition to the pseudo R^2 representing the variability in dependent variables as calculated using Eq. (2.62). The pseudo R^2 score was used to measure the proportion of the total variability explained by the model. Although the RBF-GFR model performed better than the original GFR model, neither model accounted for the variability in the space use data, based on their low pseudo R^2 scores of 6.1%.

Table 5.7: Comparison of the original GFR (first order) and the RBF-GFR method (one basis function) using three main variables for sparrow population data.

Method	AIC	BIC	Pseudo R^2
Original GFR	1698.76	1781.234	0.0606
RBF-GFR	1698.68	1781.16	0.0607

To better account for the variability in both models, the size of each colony was added as an additional explanatory variable representing values applying uniformly to each sampling instance in both models, along with the three main variables of *grass*, *bush*, and *roof*. The size of each colony was included in both models as large colonies behave differently from small colonies. The full RBF-GFR model is as follows:

$$\begin{aligned} use \;= \exp\{\; & \gamma_{0,0} + \gamma_{1,0} grass + \gamma_{2,0} bush + \gamma_{3,0} roof + \gamma_{4,0} size + \sum_{m=1}^{M} \delta_{0,1}^{(m)} I_{grass,m} + \\ & \sum_{m=1}^{M} \delta_{0,2}^{(m)} I_{bush,m} + \sum_{m=1}^{M} \delta_{0,2}^{(m)} I_{bush,m} + \sum_{m=1}^{M} \delta_{1,1}^{(m)} (grass \cdot I_{grass,m}) + \\ & \sum_{m=1}^{M} \delta_{1,2}^{(m)} (grass \cdot I_{bush,m}) + \sum_{m=1}^{M} \delta_{1,3}^{(m)} (grass \cdot I_{roof,m}) + \sum_{m=1}^{M} \delta_{2,1}^{(m)} (bush \cdot I_{grass,m}) + \\ & \sum_{m=1}^{M} \delta_{2,2}^{(m)} (bush \cdot I_{bush,m}) + \sum_{m=1}^{M} \delta_{2,3}^{(m)} (bush \cdot I_{roof,m}) + \sum_{m=1}^{M} \delta_{3,1}^{(m)} (roof \cdot I_{grass,m}) + \\ & \sum_{m=1}^{M} \delta_{3,2}^{(m)} (roof \cdot I_{bush,m}) + \sum_{m=1}^{M} \delta_{3,3}^{(m)} (roof \cdot I_{roof,m}) + \delta_{4,1} (grass \cdot size) + \\ & \delta_{4,2} (bush \cdot size) + \delta_{4,3} (roof \cdot size) \} \end{aligned} \tag{5.3}$$

where M is the best number of basis functions in the RBF-GFR model, $\gamma_{i,0}$ is the inter-

cept that does not depend on changes in availability for the ith covariate, and $\delta_{i,j}^{(m)}$ is the coefficient of $\gamma_i(\mathbf{x})$ for the m^{th} basis function of the j^{th} variable.

A noticeable improvement occurred in terms of the ability to account for the variability on this inclusion in both models: at that point, the RBF-GFR model explained 30% of the null deviance and the original GFR model explained 31%. A leave-one-out cross-validation scheme was then applied to calculate the out-of-sample performance score R^2: thus, each time the models were applied, all the colonies except one were used, for a total of 31 colonies, allowing tests to be performed on the colony not used in the training set. Table 5.8 shows the out-of-sample performance using the original GFR, regularized GFR, RBF-GFR and regularized RBF-GFR models.

Table 5.8: Comparison of the out-of-sample R^2 between the RBF-GFR model with the original GFR model on the sparrow population data with non-regularized and regularized approaches.

Method	R^2 (order or basis)	R^2 (regularized)
Standard model	0.265± 0.603	-
Original GFR	0.338 ± 1.01	0.241± 0.561
RBF-GFR	0.306 ± 0.673	0.252± 0.538

The median of out-of-sample R^2 scores over the 32 colonies is shown in Table 5.9, which highlights that the forecasting performance scores of the CART and the RFs in combination with the GFR and RBF-GFR models are better than those of the original GFR and RBF-GFR models.

Table 5.9: Median of out-of-sample performance scores for the GFR and RBF-GFR models and the CART, RF models and XGBoost in combination with the GFR model using the sparrow dataset.

Models	R^2 (order or basis)	R^2 (CART)	R^2 (RFs)	R^2 (XGBoost
Original GFR	0.338 ± 1.01	0.619 ± 0.674	0.730 ± 0.311	0.834 ± 0.594
RBF-GFR	0.306 ± 0.672	0.885 ± 0.171	0.861 ± 0.198	0.861 ± 0.205

Using XGBoost in combination with either the GFR or RBF-GFR models for the sparrow data causes no over-fitting problems, as shown in Fig. 5.8, on using Algorithm 4. A

total of 500 iterations was thus used to apply the XGBoost model in combination with the GFR models and the results are in Table 5.9.

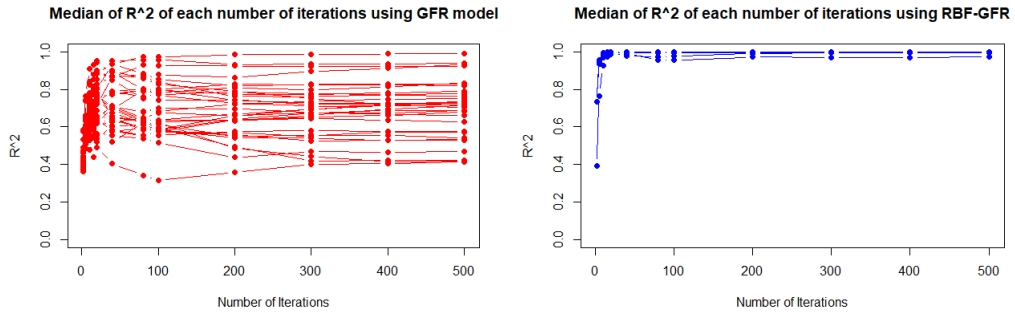


Figure 5.8: Optimization of iteration number using Algorithm 4 on a combination of the XGBoost with the original GFR (left panel) and XGBoost using RBF-GFR (right panel) as applied to the sparrow population dataset.

5.2.4 Results of the Wolf Dataset

The original GFR and the RBF-GFR models were applied using the first-order GFR model and the RBF-GFR model with one basis function for the wolf dataset. Higher polynomial orders or more basis functions result in a non-identifiable model as the 11 wolves observed belonged to just five packs, leading to a lack of high diversity between the packs. 17 Gaussian mixture components were used to approximate habitat availability based on the model selection score BIC, as described in Appendix A.3.

Table 5.10 shows the median out-of-sample performance scores from the 11 wolves in wolf dataset using an 11-fold cross-validation scheme.

Table 5.10: Median of out-of-sample performance scores for the standard, original GFR and RBF-GFR models using the wolf dataset.

Models	$R^2 \pm$ (MAD $\times$c)
Standard model	0.215 ± 0.603
Original GFR	0.156 ± 0.250
RBF-GFR	0.219 ± 0.157

The results, as seen in Table 5.11, suggest that the use of RF within the GFR models offers better predictions than the original and RBF-GFR models.

Table 5.11: Median of out-of-sample performance scores for the original GFR and RBF-GFR models and the CART, RF and XGBoost models in combination with the GFR model using the wolf dataset.

Models	R^2 (order or basis)	R^2 (CART)	R^2 (RFs)	R^2 (XGBoost)
Original GFR	0.157 ± 0.250	0.222 ± 0.086	0.769 ± 0.082	0.405 ± 0.200
RBF-GFR	0.219 ± 0.158	0.182 ± 0.075	0.760 ± 0.080	0.345 ± 0.173

Based on Algorithm 4, the best number of iterations within the wolf dataset was 100 for the original GFR model in combination with the XGBoost model, as shown in the left panel of Fig. 5.9. However, only 40 are required on combining the RBF-GFR model with the XGBoost model, as suggested by the right panel of Fig. 5.9. A potential explanation for this significant difference is that the wolf dataset may contain more outliers than the other datasets, as suggested by Fig. 5.9.

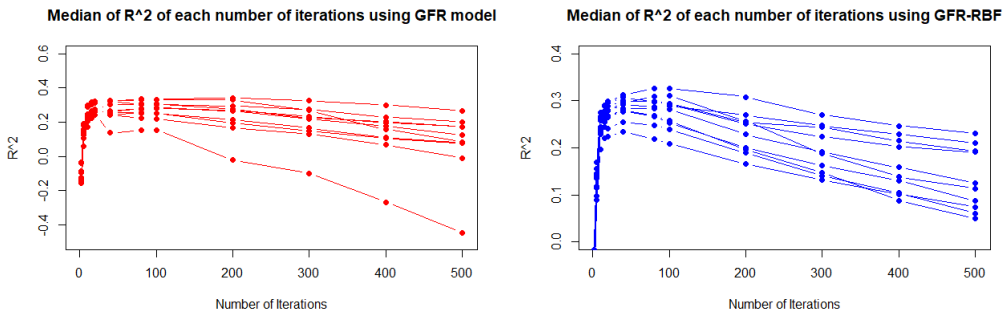


Figure 5.9: Optimization of iteration number using Algorithm 4 on a combination of the XGBoost with the original GFR (left panel) and XGBoost using RBF-GFR (right panel) as applied to the wolf dataset.

The results in Table 5.11 suggest that the use of XGBoost within the GFR models offers better predictions than the original and RBF-GFR models.

5.3 Relevance for CART, RF and XGBoost

The CART, RF, and XGBoost approaches were used in combination with the GFR and RBF-GFR models to increase the out-of-sample prediction accuracy. The difference between the out-of-sample scores using CART, RF and XGBoost and those resulting from using the GFR models was significant across most datasets, as shown in Fig. 5.10. However, the use of regularized GFR models in the second simulated dataset addresses the over-fitting problem in this dataset, suggesting that the regularization approach is relevant if the ensemble approach is not used. However, with an ensemble tool, the regularized method is less critical, as shown in Fig. 5.10. The CART, RFs and XGBoost approaches in the second simulated dataset were thus used in combination with non-regularized GFR models, as shown in Fig. 5.11.

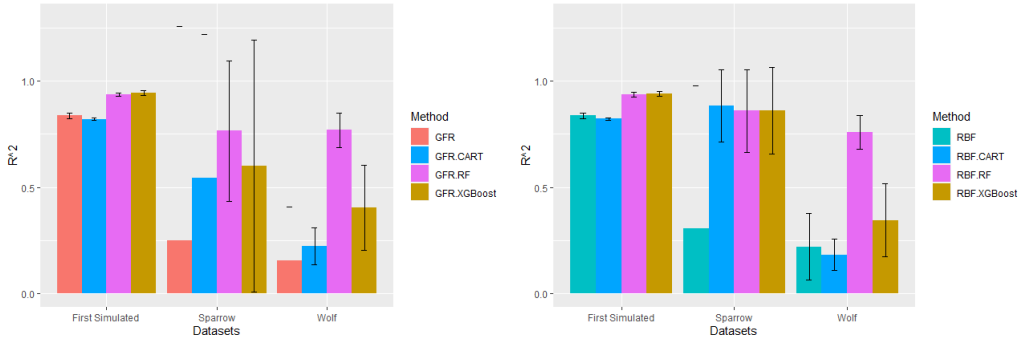


Figure 5.10: Comparison of performance scores for the original GFR, CART, RF and XGBoost using the original GFR (left panel) and the RBF-GFR, CART, RF and XGBoost using RBF-GFR (right panel), as applied to three different datasets.

5.4 Comparison Between RF and XGBoost Models

The RF and XGBoost models were combined with the original GFR and RBF-GFR models to increase the latter's predictive power. The out-of-sample performance scores of the RF and XGBoost models were thus compared. Table 5.12 and Fig. 5.12 show the out-of-sample scores of the original GFR and RBF-GFR models alongside those where RF and

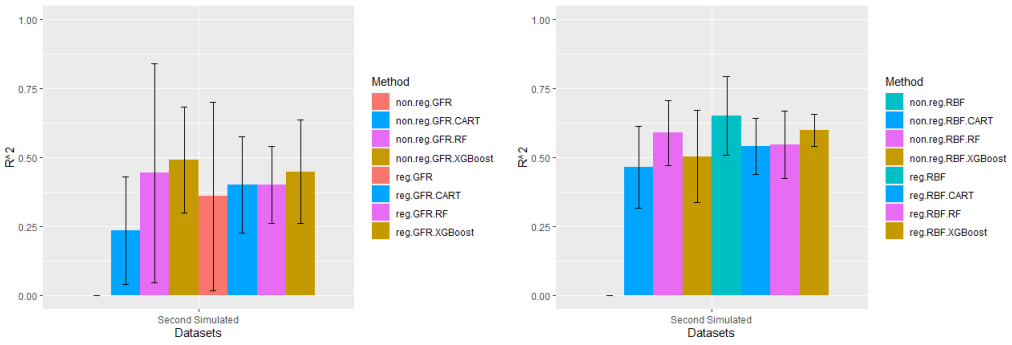


Figure 5.11: Comparison of performance scores for regularized GFR, non-regularized GFR , regularized GFR-CART, non-regularized GFR-CART, regularized GFR-RF, non-regularized GFR-RF, regularized GFR-XGBoost and non-regularized GFR-XGBoost (left panel) using the second simulated datasets. The right panel is a comparison of performance scores for regularized RBF-GFR, non-regularized RBF-GFR, regularized RBF-GFR-CART, non-regularized RBF-GFR-CART, regularized RBF-GFR-RF, non-regularized RBF-GFR-RF, regularized RBF-GFR-XGBoost and non-regularized RBF-GFR-XGBoost using the second simulated dataset.

XGBoost were applied to the four datasets. The predictive power of the GFR models in combination with the RF model exceeds that of the GFR models in combination with the XGBoost model based on the wolf dataset, while the differences between the models using the other datasets are insignificant.

Table 5.12: Median of out-of-sample R^2 scores for the RF and XGBoost models in combination with the GFR and RBF-GFR modelS applied to the four datasets.

Datasets	GFR-RF	GFR-XGBoost	RBF-GFR-RF	RBF-GFR-XGBoost
First simulated	0.936 ± 0.008	0.944 ± 0.012	0.937 ± 0.010	0.941 ± 0.011
Second simulated	0.443 ± 0.396	0.491 ± 0.192	0.571 ± 0.122	0.535 ± 0.102
Sparrow	0.730± 0.311	0.834± 0.594	0.861 ± 0.196	0.861 ± 0.205
Wolf	0.769 ± 0.082	0.405 ± 0.200	0.760 ± 0.080	0.345 ± 0.173

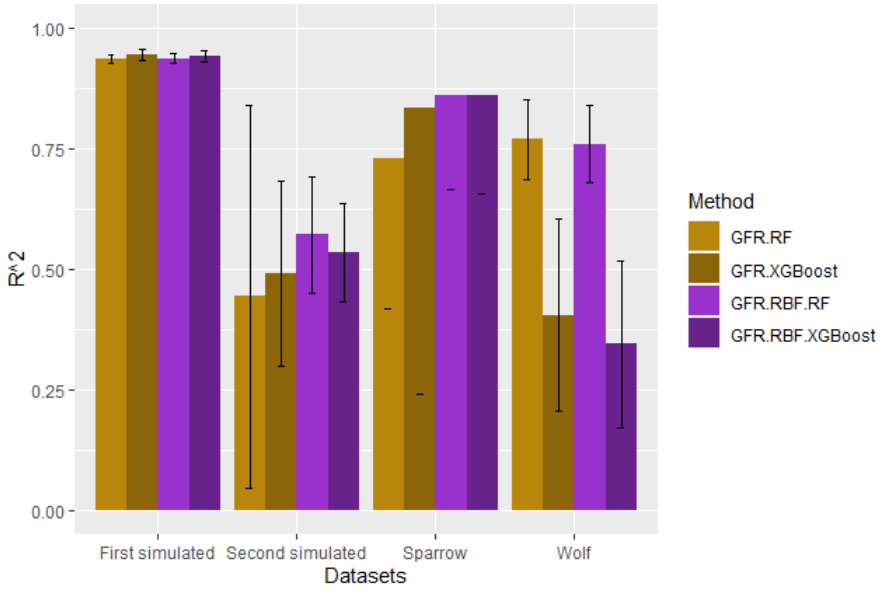


Figure 5.12: Comparison of out-of-sample R^2 scores for the RF and XGBoost using the original GFR and the RF and XGBoost using RBF-GFR applied to four different datasets.

5.5 Relevance of Regularization

The bar charts in Fig. 5.13 show the out-of-sample R^2 scores both with and without regularization for the original GFR and RBF-GFR models across the four datasets. In three out of four datasets, there was no substantial difference between regularized and non-regularized models, indicating that the unregularized models did not have an over-fitting problem. However, applying the models to the second simulated dataset without regularization caused an over-fitting issue with high variance and poor forecasting performance scores. Ridge regression was then applied to reduce variance to prevent this over-fitting problem. The regularized models outperformed the non-regularized model in the second simulated dataset.

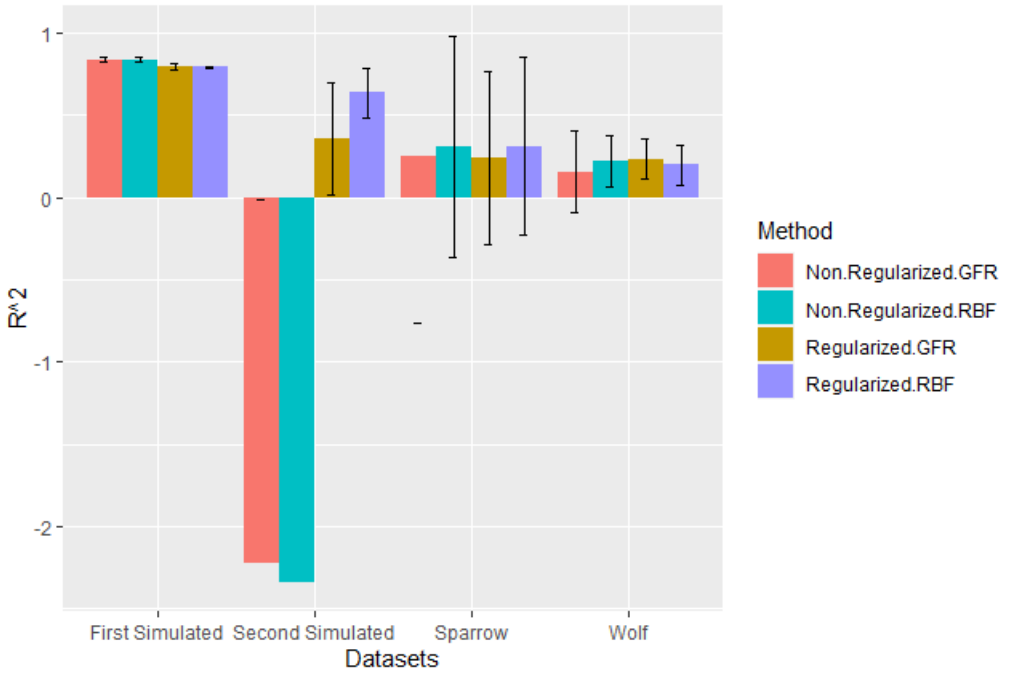


Figure 5.13: Comparison of performance scores for regularized and non-regularized GFR and RBF-GFR approaches applied to four different datasets.

5.6 Model Ranking

The models were then ranked by performance for each of the four data sets in turn, as shown in Fig. 5.14. This table thus offers the ranks and the detailed out-of-sample performance of all the models included in the study, as shown in Fig. 3.5. While none of the individual models consistently outperformed all other models across all data sets, a pattern did emerge whereby the ensemble methods, which use bagging or boosting for the creation of random forests, tend to outperform all other models as a class (namely the "class" of ensemble methods, as opposed to individual models). In particular, the combination of the proposed RBF-GFR and GFR models with bagging, as represented by the two models shown in the top rows of Fig. 5.14, consistently achieve ranks in the top 40% of the performance spectrum. This offers evidence of more stable performance than the non-ensemble models, while the latter show higher variability, as exemplified by

the regularized RBF-GFR model, which appears as the best model for the second simulated data set, but as the third-worst model for the first simulated data set. Regularization was not applied to the individual models included in the ensembles, with the results thus suggesting that, in terms of improving out-of-sample generalization performance, model averaging over ensembles offers an alternative to regularization, confirming similar findings in Machine Learning literature (Sollich and Krogh, 1996). The combination of the proposed RBF-GFR model with random forests (RBF-GFR-RF) produced the best model overall, consistently achieving a place in the top three performance rankings. An important additional finding was that almost all the methods proposed in this study outperform the original GFR model from Matthiopoulos et al., 2011, which was the initial aim motivating the present work. As shown in Fig. 5.14, the GFR model never achieves a rank better than 6. The computation time for the methods depends on the dataset. The random forest and XGBoost models take minutes to fit in the first simulated dataset but hours to fit in the second simulated dataset. The rest of the models take less than an hour to fit in all datasets. R^2_{DEV} in Eq. (2.61) is generally a better behaved measurement than R^2 in Eq. (2.60) for count data as described in Section 2.9. I used R^2_{DEV} to calculate the out-of-sample predictive performance in the two simulated datasets as shown in Fig. A.22 in Supplement A.7. However, the overall ranks using R^2_{DEV} are not different from the overall ranks using R^2 in Eq. (2.60) (comparing the average rank in Fig. 5.14 with ranks in Fig. A.22 in Supplement A.7).

5.7 Visualising Model Predictions

The predictions of animal habitat usage derived from the models used were presented in spatial maps and further visualisations of species abundance were generated using the second simulated dataset, which contained 20 sample instances, with each sample being formed of 2,500 observations (50 x 50 arena). One map was reserved from the cross-validation scheme and used to generate predictions from all models. Samples # 1 & # 17 were selected to represent, along with selected samples are shown in Supplement A.9.

Figure 5.15 shows a heat map of species abundance and geographical predictions of abundance in terms of latitude and longitude for the ground truth and the various models

	Sim1	Sim2	Sparrow	Wolf	Average
RBF-GFR-RF	0.937	0.593	0.86	0.76	0.782
GFR-RF	0.936	0.443	0.73	0.769	0.719
RBF-GFR-XGBoost	0.94	0.535	0.861	0.354	0.673
GFR-XGBoost	0.944	0.491	0.834	0.405	0.669
RBF-GFR-CART	0.822	0.44	0.884	0.182	0.582
GFR-CART	0.821	0.235	0.619	0.222	0.474
Reg RBF-GFR	0.796	0.635	0.252	0.199	0.471
Reg GFR	0.796	0.359	0.241	0.234	0.408
GLM	0.731	0.256	0.265	0.215	0.367
GFR	0.837	-0.97	0.338	0.156	0.09
RBF-GFR	0.837	-2.4	0.356	0.219	-0.103

Figure 5.14: Rank table of the out-of-sample R^2 scores of the models using the two simulated, sparrow, wolf datasets and the average score of out-of-sample R^2. The shading of colours indicates the ranks of the models. For each column, the colour shading ranges from yellow to dark red, with yellow indicating the lowest score in the respective column, and dark red indicating the maximum value.

shown in Table 3.1. Light colours indicate low abundance levels, so the abundance levels increase as the colour shading gets darker. The two panels differ in colour range, with the same output range used for all models in the upper panel, while in the lower panel, the colour range encompasses the whole range of model outputs, which may be different for different models, as the minimum and maximum values for which colours are plotted are limited by the minimum and maximum actual values. Model outputs larger than the maximum value of the truth are thus treated as missing values and are shown in white.

These results suggest that the RBF-GFR-RF model, which is overall the best model according to Fig. 5.14, also offers the best qualitative agreement with the ground truth

for predicted spatial abundance profiles. The RBF-GFR-RF predictions faithfully reproduce the high-intensity hotspots near the top right corner of the map, around coordinates (0.8,0.8), as well as those near the left margin, around coordinates (0.2,0.4). The alternative models also tend to capture the ground truth pattern qualitatively, but these display larger deviations. For instance, the GLM model shows reasonable agreement with the ground truth in the left panel, i.e., when plotted on the same scale as the ground truth; however, on its individually adjusted intensity scale (lower panel), the GLM predictions are systematically lower than the ground truth values, implying that the GLM model systematically underestimates extremes, while the GFR model shows an opposing trend, systematically overestimating extremes, as indicated by the white patches in the lower panel. Furthermore, the out-of-sample R^2 scores shown in Table 5.13 of sample instance # 1 for the various models used to predict the heat maps in Figs. 5.15 illustrate my finding from the heat maps where the RBF-GFR-RF score is higher than the scores of other models.

Table 5.13: Out-of-sample R^2 scores for the various models shown in Fig. 3.5 of sample instance # 1 from the second simulated dataset.

Models	R^2
GLM	0.612
GFR	-3.6
Reg GFR	0.695
GFR-CART	0.327
GFR-RF	0.673
GFR-XGBoost	0.764
RBF-GFR	-0.734
Reg RBF-GFR	0.742
RBF-GFR-CART	0.628
RBF-GFR-RF	0.769
RBF-GFR-XGBoost	0.713

Figure 5.16 shows the predicted abundance profiles for a different sample instance: as the regularized RBF-GFR model achieved the strongest performance in terms of out-of-sample R^2 scores, it is not surprising to see that the 2D intensity profiles suggest that the regularized RBF-GFR model shows very good agreement with the ground truth, faith-

fully reproducing its true intensity hotspots around coordinates (0.9,0.6) and (0.3,0.1). However, despite not being the absolute best models on this occasion, the two models' ensemble RBF-GFR-RF and RBF-GFR-XGBoost show 2D abundance profiles that are very similar, and differences are hardly discernible by the eye. The other models show stronger deviations from the ground truth, with the lower panel of the figure suggesting that the majority of the alternative models systematically underestimate extreme values.

5.8 Variable Ranking

Colony size, measured by the maximum number of males in each colony, is the most important feature of the sparrow population, which was determined using the best two models overall, the RF approach in combination with the GFR and RBF-GFR models. The percentage of *bush*, on the other hand, has the lowest importance score compared to the other main variables, as seen in Fig. 5.17. The importance scores were calculated using the mean decrease in accuracy from permuting out-of-bag data, as described in Section 2.7.1.1. Both features positively affect the habitat suitability of the sparrows based on the GFR and RBF-GFR models.

For the wolf dataset, using the RF approach in combination with the RBF-GFR model, *distance to high human use* is the most important covariate, positively affecting the wolves' habitat preference. The *slope* and *distance to high human use* strongly influence habitat preference; decreasing the *slope* or increasing *the distance to high human use* increases habitat preference. In contrast, *rocks* have the lowest impact on the wolves' habitat preference, as seen in the right panel of Fig. 5.18. However, the *slope* is more important than the *distance to high human use* based on the RF approach in combination with the GFR model, as seen in the left panel of Fig. 5.18. Since the *distance to high human use*, *slope* and *distance to high human use* strongly influence habitat preference based on the importance scores in Fig. 5.18, the low importance variables are excluded from the GFR-RF and RBF-GFR-RF models to achieve more parsimonious model to increase the predictive performance and avoid overfitting problems. However, the predictive power of the GFR-RF and RBF-GFR-RF models using the three variables are 0.758 and 0.75, respectively, while the predictive power of the same models using all variables are 0.769 and 0.76, indicating

that exclusion of the variables with low importance scores does not affect the predictive performance.

5.9 Conclusion

The predictive performance of the GFR, RBF-GFR and their extensions were found to be significantly better than the currently used SDM approaches and approximately consistent across the four datasets. Modelling 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 showed moderate improvements in the out-of-sample R^2 score. However, combining the original GFR and RBF GFR with the ensemble approach using bagging and boosting showed a substantial improvement in terms of the out-of-sample R^2 scores. This combination has appeared consistently in the top rows of the rank table in Fig. 5.14, whereas the performance of other models has been less consistent. From prediction visualisation, applying the GFR model increased the risk of under-predicting the ground truth. In contrast, the flexible RBF-GFR model provided a version that exaggerates the extreme abundance values. The results from the ensemble and regularization approaches confirmed the finding that these approaches reduce over-fitting. The regularization and ensemble approaches were also able to control over-fitting and significantly increase predictive performance (increases in R^2 from 0.25 to 0.85 in some cases). The improvement in predictive performance in the simulated datasets was better than the improvement using the two real-life datasets. The different amounts of predictive improvement can be attributed to the fact that the simulated datasets were simulated using assumptions and roles that could provide better adherence to the spatial stationarity of covariates. Furthermore, the significant difference in the improvement from the dataset can result from different data types; the sparrow and wolf datasets are use-availability datasets with poorer information compared with the abundance dataset (Yates et al., 2018). Using the best predictive model based on the rank table in Fig. 5.14, the most important variables of the model ensemble in combination with the GFR model had almost the same essential roles when the model ensemble was combined with the RBF-GFR model.

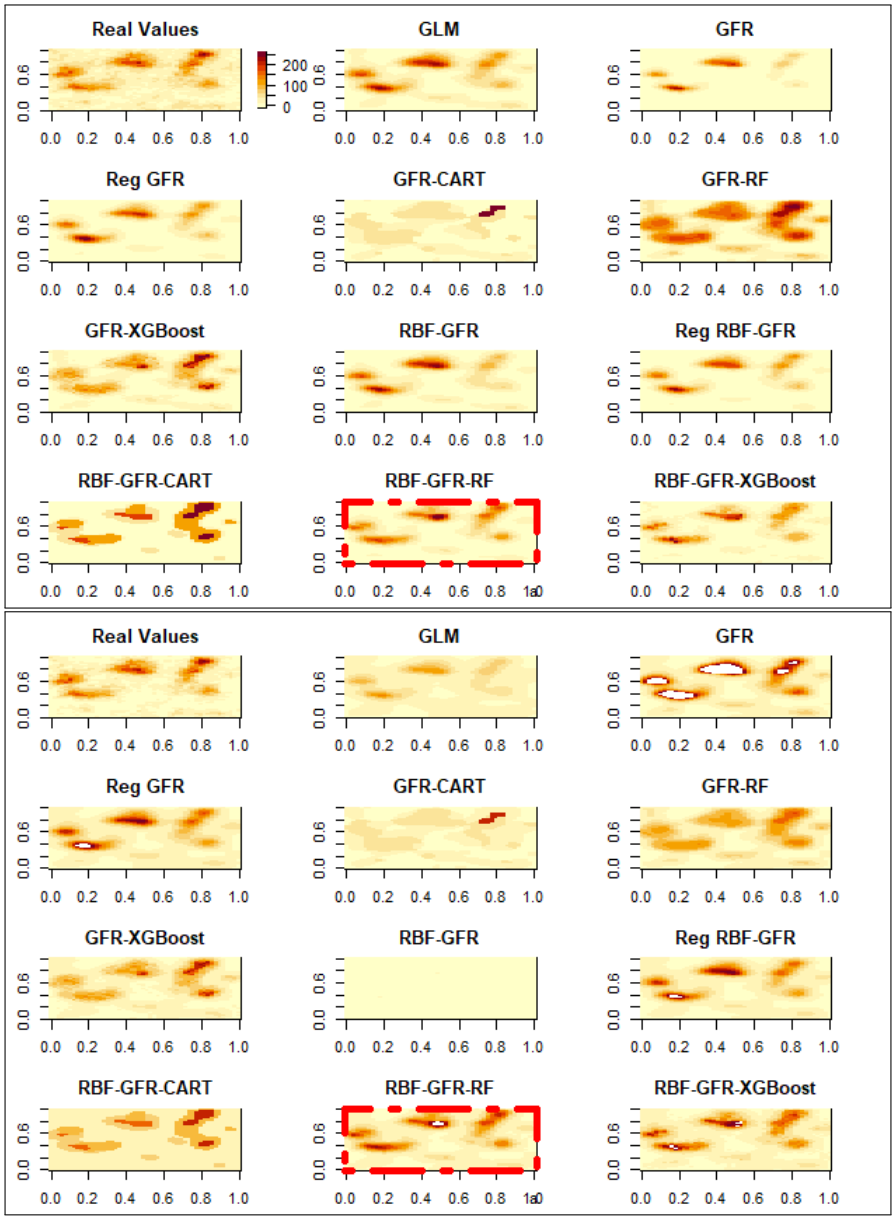


Figure 5.15: A heat map of abundance and geographical predictions of the abundance of sample instance # 1 from the second simulated dataset in terms of geographical dimensions: latitude and longitude for the ground truth and the various models shown in Table 3.1. The two panels differ in colour range. In the upper panel, I use the same output range for all models. In the lower panel, the colour range encompasses the whole range of model outputs and may be different for different models but the minimum and maximum values for which colours should be plotted are limited by the minimum and maximum numbers of the true values. Model outputs that are larger than the maximum value of the truth are treated as missing values and are shown in white. The map with red borders is the best predictive model based on out-of-sample R^2.

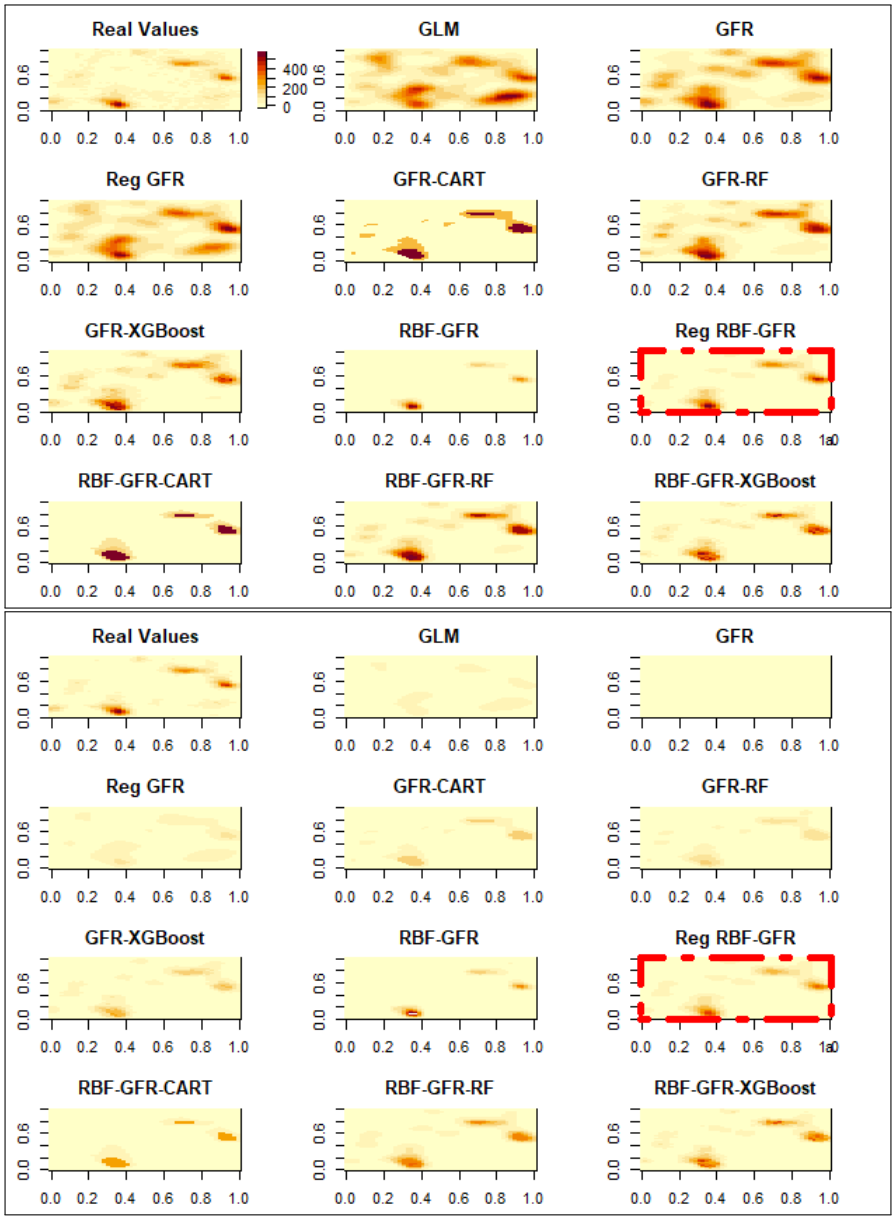


Figure 5.16: A heat map of abundance and geographical predictions of the abundance of sample instance # 17 from the second simulated dataset in terms of geographical dimensions: latitude and longitude for the ground truth and the various models shown in Table 3.1. The two panels differ in colour range. In the upper panel, I use the same output range for all models. In the lower panel, the colour range encompasses the whole range of model outputs and may be different for different models but the minimum and maximum values for which colours should be plotted are limited by the minimum and maximum numbers of the true values. Model outputs that are larger than the maximum value of the truth are treated as missing values and are shown in white. The map with red borders is the best predictive model based on out-of-sample R^2.

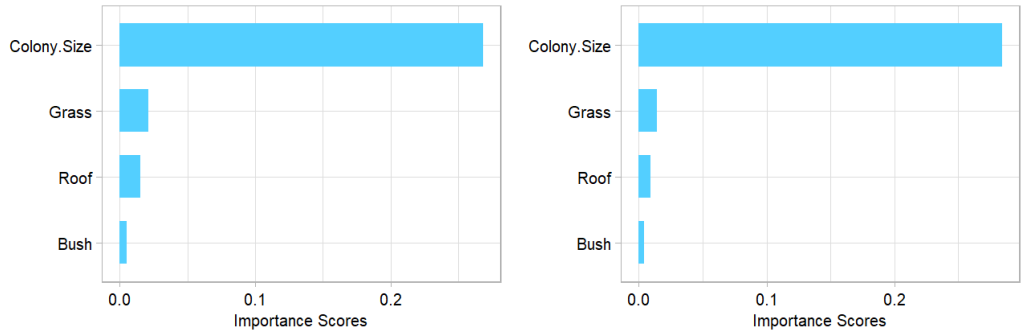


Figure 5.17: Importance scores for the main variables in the sparrow population dataset, using the GFR-RF model in the left panel and the BF-GFR-RF model in the right panel.

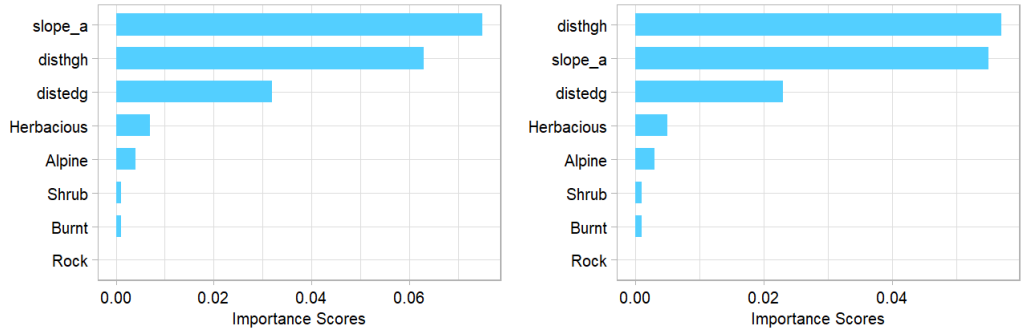


Figure 5.18: Importance scores for the main variables in the wolf dataset, using the GFR-RF model in the left panel and the RBF-GFR-RF model in the right panel.

Chapter 6

Quantifying and Interpreting the Variability of Selectivity Coefficients

6.1 Introduction

The main goal of the previous chapters was to increase the predictive power (i.e., the transferability) of SDMs, by improving on the performance of the GFR model. I have developed robust models that predict out-of-sample observations with high accuracy. To explore these models, we can look at the spatial predictions (i.e., look at scenario-specific plots of predicted usage over geographical space), which I have done in Chapter 5. A higher level of abstraction can be achieved by exploring the output of these models for regression selection coefficients $\beta_{i,b}$ (i.e., look at plots that are space-independent, but scenario-specific). At an even higher level, we can assess the models by looking at the GFR selectivity coefficients γ_i (i.e., look at plots that are both space- and scenario-independent). Looking at the selectivity coefficients γ_i offers a lower dimensional space of information, so it is easier to explore visually and also, it captures the essence of the behaviour of the model since the selectivity coefficients γ_i are context-independent. They are not affected by the environmental context at the location of a spatial point (e.g., exploring the spatial prediction in Chapter 5) or the broader availability of habitats in space (e.g., exploring using $\beta_{i,b}$). The selectivity coefficients can be thought of as characteristics of the species, whereas the

selection coefficients $\beta_{i,b}$ and the spatial predictions emerge from the interaction between the species and its environment at different scales.

It is also essential to know whether these models allow us to look beyond the predictions by offering us some explanatory power of the mechanisms mediating species distributions. In order to do this, it is necessary to explore patterns in the behaviour of fitted models, probe these, using statistical and visual summaries and interpret them post-hoc, in the light of the assumptions that underpin the data-generating process. Output patterns in the fitted models can be examined at the end-product (e.g., the predicted species distributions) or at some more informative intermediate stage (e.g., the selection coefficients generated for each environmental scenario). The aim of this chapter is to explore such patterns, and hence gain new insights into how fundamental biological mechanisms (such as the pursuit of resources) give rise to particular values of regression coefficients. In mapping, the "mechanistic to the empirical" several indirect (and maybe counter-intuitive) effects may become apparent, so it is useful to build some intuition in this respect. Explanatory models aim to find causative relationships (Sainani, 2014) by building hypotheses about invisible structures that help explain visible phenomena (Harré, 2002). In this chapter, images have been used to summarize the behaviour of different parts of the statistical SDMs under wide ranges of different environmental scenarios.

Transforming information into images helps explore the information, discover any patterns or behaviours, and analyse this information. Building images provides approaches that effectively help manage more information and rapidly analyse the data (Huber and Healey, 2005). Visualising a dataset using graphical approaches provides a better understanding of the dataset and discovers any data irregularities (Frankel and Reid, 2008). Visualising the predictions of the models used in the present study helped to understand certain patterns in the predictions, analyse the predictions and assess the transferability of each model, as discussed in Section 5.7. For deeper insights into varying coefficient models such as the GFR model, it is preferable to go beyond merely visualising spatial predictions of abundance by, instead, visualising the changes in the regression coefficients. This allows direct observation of the relationship between the availability of habitats and apparent preference for these by animals. Here, I focus on the simulated dataset from Matthiopoulos et al. (2011), which is described in Section 4.2.2. Knowing the mecha-

nisms generating these data is the key reason behind using this simulated dataset. I begin this chapter by discussing the specific biological rules of this simulation. The models are assessed by their ability to infer these mechanisms by looking at the patterns illustrated in images, where these patterns are purely the result of the statistical model being fitted to data. Furthermore, I use the simulated dataset from Matthiopoulos et al. (2011) because it is a simpler version of Matthiopoulos et al.'s (2015) dataset, which consists of 20 samples and looks at just two resources (e.g., food and cover). It is hard to carry out this assessment for real datasets because the mechanisms that control the real animals are unknown, and these data consist of spatial locations (e.g., colonies in the sparrow population dataset), that are not complete grids with full maps.

In this chapter, I focus on the coefficients of the regularized GFR and RBF-GFR models because they are the best models applied to the dataset based on Fig. 5.14.

6.2 Selectivity Coefficients Concept

The fixed selection coefficients β_i in conventional SDMs based on the GLM structure in Eq. (2.1) quantify the slopes of the response variable in relation to its environmental covariates. However, in varying-coefficient models such as the regularized GFR and RBF-GFR, the $\beta_{i,b}$'s in Eq. (2.2) change with the environmental composition. Specifically, they are represented by the integral of the weighted habitat availability $f_b(\mathbf{x})$, where the weights $\gamma_i(\mathbf{x})$ can be thought of as *selectivity coefficients*, which are space- and scenario-independent. The selectivity coefficients of the GFR models are formulated using a polynomial function, whereas the selectivity coefficients of the RBF-GFR models are formulated using a radial basis function, as seen in Eqs. (2.4) and (3.3). The $\gamma_i(\mathbf{x})$ describes how the selection coefficient $\beta_{i,b}$ adapts to changes in habitat availability $f_b(\mathbf{x})$. More formally, $\gamma_i(\mathbf{x})$ represents the increment or decrement applied to $\beta_{i,b}$ through the addition into the accessible environment of an extra unit of a particular habitat $\mathbf{x}$. The value of the gamma tells us whether an increase in the availability of that particular habitat is likely to increase the apparent preference or decrease it and by how much. In other words, the gamma values represent how much an extra unit of a particular habitat x affects the value of beta. Therefore, if I want to add one more square metre of a particular habitat, such as a particular

value of food and cover, the gamma values tell me what this is going to do to the betas and how this is going to affect the behaviour of the animal. As a result, this gives either an apparent increase in preference or a decrease. A positive $\gamma_i(\mathbf{x})$ value implies that adding one more unit of habitat $(\mathbf{x})$ will tend to create the impression of a stronger preference for the i^{th} environmental variable. In contrast, negative values tend to create the impression of lower preference for the i^{th} environmental variable when adding a unit of habitat $\mathbf{x}$. Values of $\gamma_i(\mathbf{x})$ close to zero indicate that the coefficient $\beta_{i,b}$ is unresponsive to changes in the availability of habitat $\mathbf{x}$.

6.3 Simulation Rules

In the second simulated data described in Section 4.2.2, I use different environmental scenarios of food and cover availabilities to simulate the usage, where each scenario is a spatial grid of 50x50 of these features used by animals. The simulation uses a biased random walk to attract the animal towards hotspots of food (when it is hungry) and hotspots of cover (when it is sated). According to the simulation rules in Section 4.2.2, I expect that if I introduce a poor habitat in one resource or both, I will increase the apparent preference of the resources because that makes the resources rarer. On the other hand, if I add a habitat type that is rich in food, I will reduce the apparent preference for food because the animal could use up to a certain amount of food (E_1), irrespective of food abundance. Upon satiation ($E > E_1$), the animal stopped feeding and climbed up the gradient of cover until reaching a local maximum. When E fell below a starvation threshold (E_2), the animal climbed up the food gradient until reaching a local maximum. Based on this rule, if I add a habitat type that is rich in food, I will reduce the apparent preference for food because the animal could use up to a certain amount of food (E_1), irrespective of food abundance, which mimics real-life scenarios because the animal could eat food until it is saturated in real life and then do another activity such as looking for cover to hide from other animals.

6.4 Visualising Selectivity Coefficients

The selectivity coefficient values for the regularized GFR are formalized as follows:

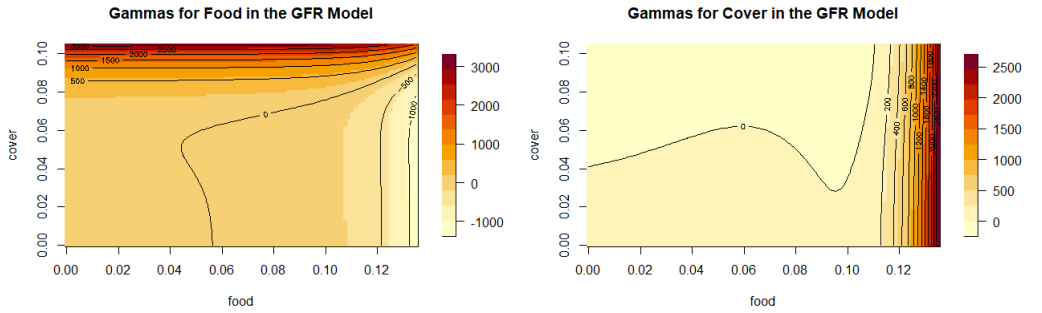
$$\gamma_i(\mathbf{x}) = \sum_{j=1}^{I} \sum_{m=0}^{M_j} \delta_{i,j}^{(m)} x_j^m \tag{6.1}$$

where the coefficient of $\gamma_i(\mathbf{x})$ for the m^{th} power of the j^{th} variable is $\delta_{i,j}^{(m)}$. The $\gamma_i(\mathbf{x})$ for the regularized RBF-GFR model is as follows:

$$\gamma_i(\mathbf{x}) = \sum_j \sum_m \delta_{i,j}^{(m)} \exp\left(-\frac{1}{2} \frac{(x_j - \xi_{j,m})^2}{\sigma_{j,m}^2} \right) \tag{6.2}$$

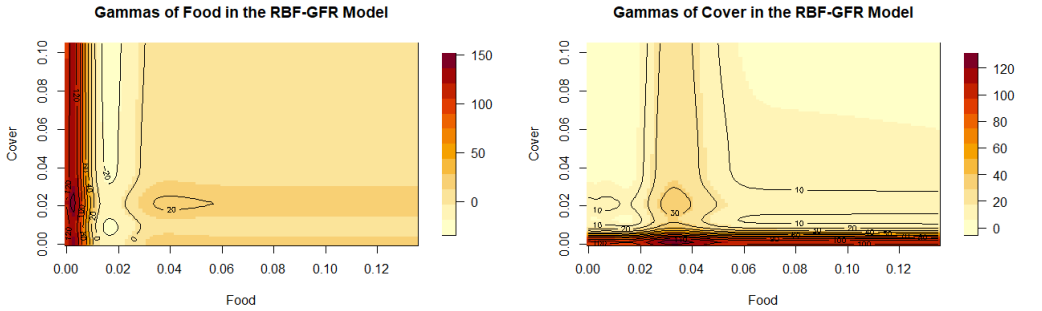
where $\xi_{j,m}$ is the centre of the mth basis function for the jth covariate and $\sigma_{j,m}$ is its bandwidth parameter. To obtain $\gamma_i(\mathbf{x})$ in both models, $\delta_{i,j}^{(m)}$ have to be estimated using Eqs. (2.6) and (3.18) for the regularized GFR and RBF-GFR models, respectively. The estimated $\delta_{i,j}^{(m)}$ is based on the selected value of λ in Eq. (3.21) chosen from an equidistant grid that minimizes the model selection score BIC, as described in Section 3.3. Using the simulated dataset from Matthiopoulos et al. (2011), Fig. 6.1 shows the selectivity coefficients plots for both habitat variables using the regularized models. The plots describe environmental space, so each point in this 2-D space represents a type of habitat with these characteristics, where x-axis and y-axis represent the two resources: food and cover. So, for example, a point in the bottom-left corner represents a habitat that is poor in both resources (food and cover) whereas a point in the upper-right corner represents habitats that are rich in both. I am interested in visualising how the selection coefficient for a particular environmental variable is affected by the addition of a single unit of each habitat across environmental space. Because the interpretation is made with a particular environmental variable in mind, each plot is specific to a named variable (this focal variable is usually reported in the title of each plot). The required effects in the selection coefficients $\beta_i(\mathbf{x})$ are indicated by the $\gamma_i(\mathbf{x})$ values that are plotted as colour gradients in each plot. The $\gamma_i(\mathbf{x})$ values that are large (compared to the baseline value of $\beta_i(\mathbf{x})$) indicate a steepening in the apparent response of the organism to the focal environmental variable (a change in $\beta_i(\mathbf{x})$),

while the negative values create an impression of reduction in the organism apparent response. The $\gamma_i(\mathbf{x})$ values close to zero tend to have no effect.



(a) Food selectivity coefficients (regularized GFR model)

(b) Cover selectivity coefficients (regularized GFR model)



(c) Food selectivity coefficients (regularized RBF-GFR model)

(d) Cover selectivity coefficients (regularized RBF-GFR model)

Figure 6.1: Selectivity coefficients for (a) food and cover (b) using the regularized GFR model and food (c) and (d) for the RBF-GFR model.

In addition, I changed the parameters, a and den, in the simulation to understand the behaviour of the $\gamma_i(\mathbf{x})$ values of the regularized GFR and RBF-GFR models by increasing the feeding rate, a, from low to high and to observe any gradual transition in the shape as I increase the metabolic cost. The plots in Figure 6.1 are the gamma values for food and cover in the regularized GFR and RBF-GFR models using 10 basis functions, den = 0.03, and the feeding parameter, a = 0.2. Figs. 6.2a and 6.3a show gamma plots using the regularized GFR model for food and cover, where a equals 0.5, 1, 1.5, and 5 and den

equals 0.001 and 0.003, respectively. Figs. 6.2b and 6.3b show the gamma plots using the regularized RBF-GFR model for food and cover, where a = 0.5, 1, 1.5, and 5 and den = 0.001 and 0.003 respectively.

6.5 Selectivity Coefficients Explication

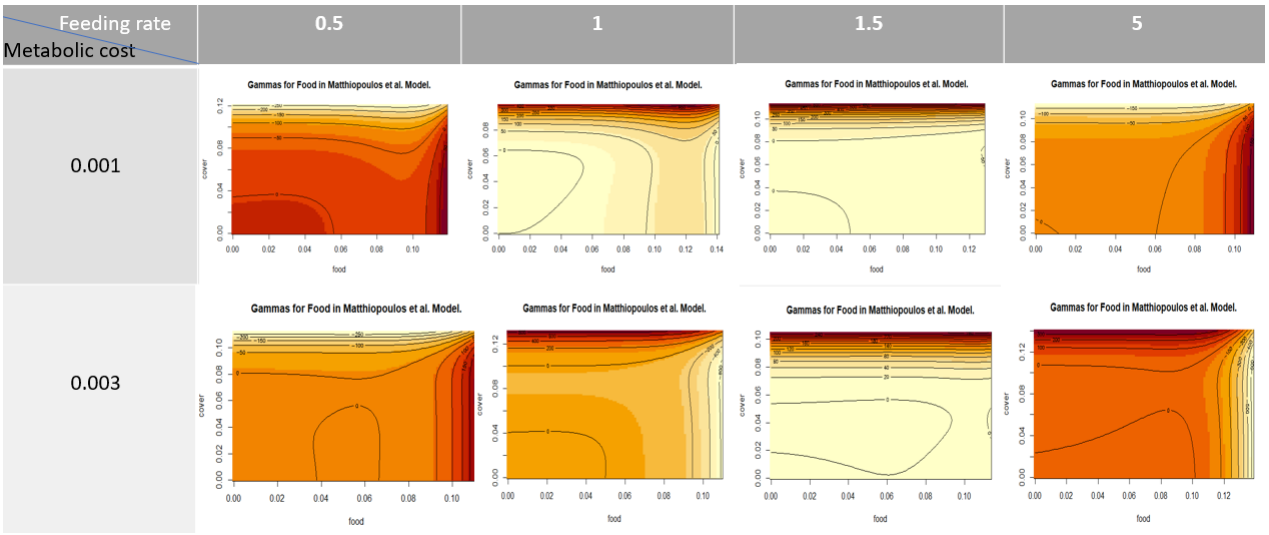
6.5.1 Models' Selectivity Coefficients

From the GFR model in Fig. 6.1a, adding one unit of food in a habitat rich in cover will increase the habitat preference of food, no matter how much food was in the habitat. However, adding one more unit of cover in a habitat rich in food increases the cover preference, regardless of how much cover there was, as seen in Fig. 6.1b.

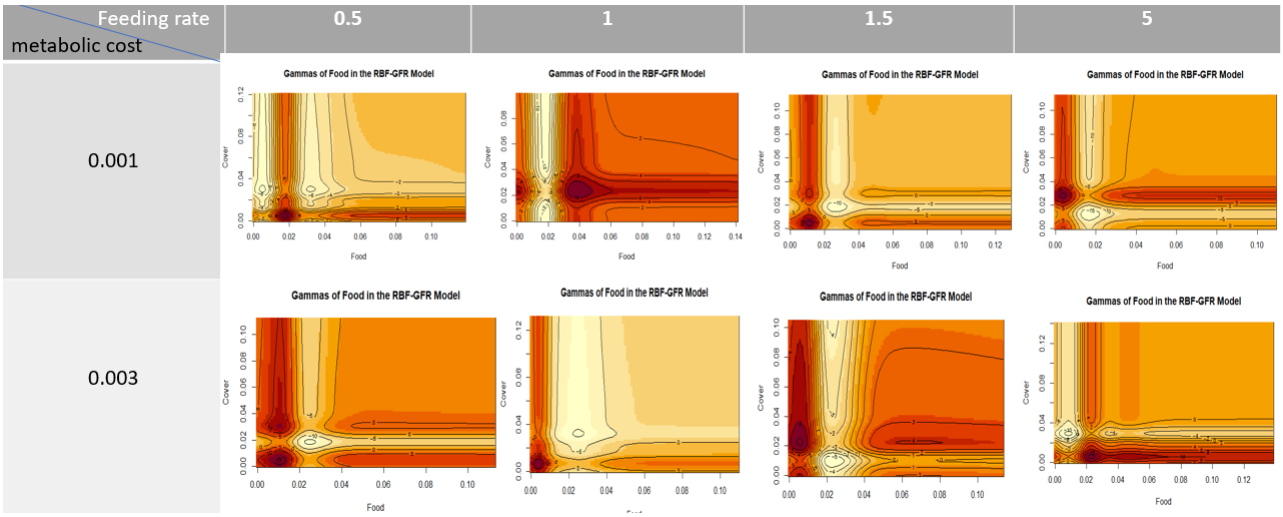
By using the RBF-GFR model in Fig. 6.1c, adding one more unit of food in a habitat that lacks food increases the food preference, regardless of the cover amount. In addition, if I increase the cover by one unit in a rare cover habitat, the cover preference increases, no matter how much food is in the habitat, when using the RBF-GFR model.

6.5.2 Varying the Simulation's Biological Parameters

The apparent preference or avoidance towards environmental variables, as expressed by the γ coefficients of the model, appear to be sensitive to changes in the maximum feeding rate a and basal metabolic cost den, as seen in Figs. 6.2 and 6.3. However, because the γ coefficient values of RBF-GFR model shows that there is no explicit causal link between the abundance of food and preference for cover, as seen in Fig. 6.1d, it might be expected that the gamma for cover would not be affected by the changes in a and den, which is illustrated in Fig. 6.3b. My results (Fig. 6.2b) indicate that there is an emergent sensitivity of a and den on the food preference, which I already observed in Fig. 6.1c, where the preference for food using this model depends on the amount of food in that habitat. The response of cover is sensitive to varying the feeding rate and metabolic cost using the regularized GFR model, as seen in Fig. 6.3a, and this behaviour is expected because the response of cover depends on the food amount, as observed in Fig. 6.1b. However,

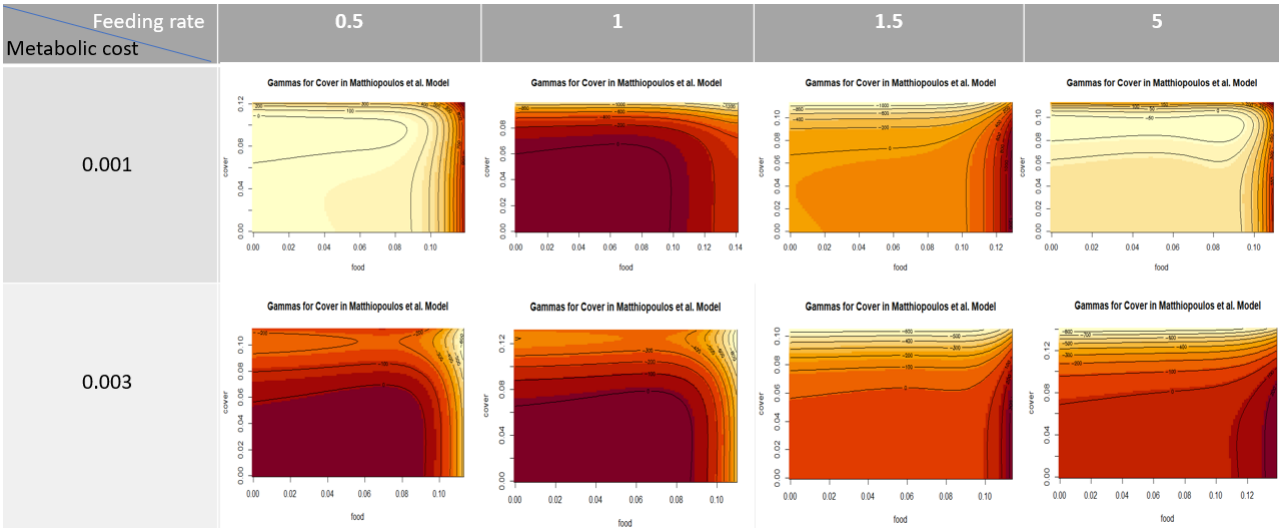


(a) Food selectivity coefficients (regularized GFR model)

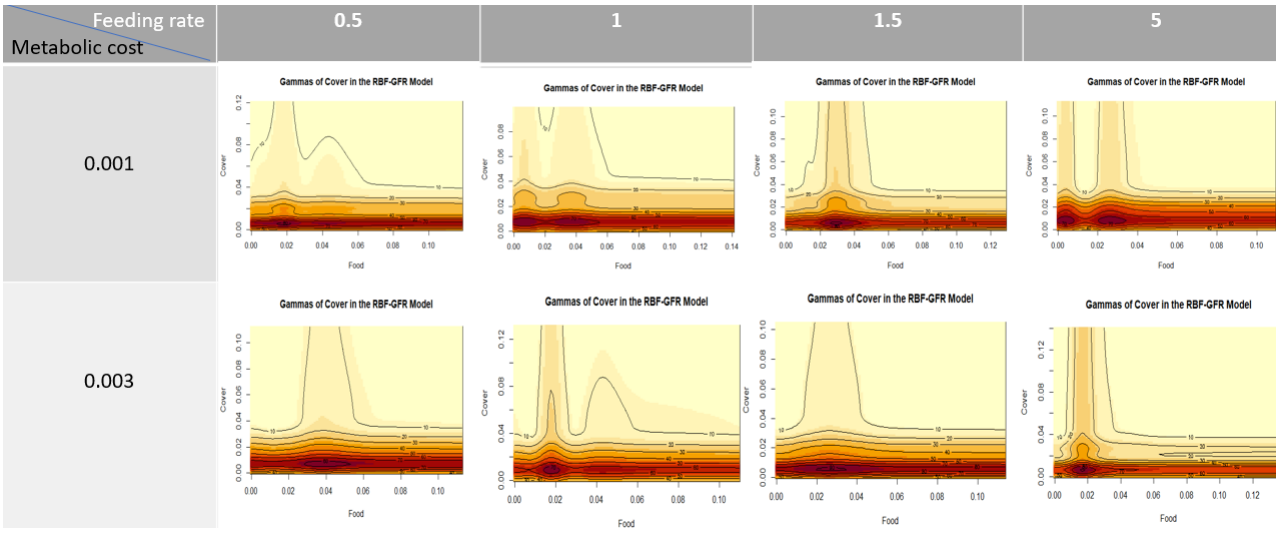


(b) Food selectivity coefficients (regularized RBF-GFR model)

Figure 6.2: (a) Selectivity coefficients for food from (a) the regularized GFR model and (b) the regularized RBF-GFR model and for different values of a in the columns and den in the rows, where a = 0.5, 1, 1.5, and 5 and den = 0.001 and 0.003.



(a) Cover selectivity coefficients (regularized GFR model)



(b) Cover selectivity coefficients (regularized RBF-GFR model)

Figure 6.3: (a) Selectivity coefficients for cover from (a) the regularized GFR model and (b) the regularized RBF-GFR model and for different values of a in the columns and den in the rows, where a = 0.5, 1, 1.5, and 5 and den = 0.001 and 0.003.

the response of food is sensitive to varying feeding rates, and metabolic costs using the regularized GFR model, as seen in Fig. 6.2a, even though the response of food does not depend on food abundance, as shown in Fig. 6.1a, which is not expected.

6.6 Investigating the Reasons for the Differences in the Selectivity Coefficients between Models

Given the stark differences in the gamma plots in Fig. 6.1 between the two models, the expected habitat preferences quantified by the regularized GFR and RBF-GFR models should be different, even though they were both found to be the best-performing models in terms of predictive power in Chapter 5. It is therefore interesting to explore whether these differences between gamma coefficients propagate to the beta coefficients (the regression coefficients that are ultimately responsible for the models' goodness-of-fit and predictive ability). Mathematically, the coefficient β is the integral of the coefficient γ times the probability density of $\mathbf{x}$, as follows:

$$\beta_{i,b} = \int \gamma_i(\mathbf{x}) f_b(\mathbf{x}) d\mathbf{x} \tag{6.3}$$

where the values of $\gamma_i(\mathbf{x})$ using the GFR model have different patterns than the values resulting from the RBF-GFR model and $f_b(\mathbf{x})$ is the probability density function for habitat availability in the b^{th} sampling instance. Here, I have used the probability density function for habitat availability derived from all environmental scenarios $f(\mathbf{x})$, which here are represented by the kernel-smoothed habitat availability. The kernel-smoothed habitat availability was used to create spatial autocorrelation between the cells like in the data simulation, where the kernel smooth has been used to create spatial autocorrelation between the seed layers (Matthiopoulos et al., 2011). The kernel bandwidth was selected using a normal reference bandwidth $\hat{h}$ (Venables and Ripley, 2013) as follows:

$$\hat{h}(\mathbf{x}) = 1.06 \times \min\left(\sigma, \frac{R(\mathbf{x})}{1.34}\right) \times n^{-0.5} \tag{6.4}$$

where σ is the standard deviation of $\mathbf{x}$ and $R(\mathbf{x})$ is the difference between the third and first quartiles of $\mathbf{x}$.

$f(\mathbf{x})$ is an attention function that allows the images to focus on the relevant regions of environmental space that, because of their high frequency of occurrence in the model-fitting data, are likely to have had a large influence on the model parameters, particularly the selectivity parameters $\gamma_i(\mathbf{x})$ that are habitat-independent, intrinsic features of the study species. Fig. 6.4 shows the effect of the smoothed habitat availability of the cells in the dataset. Most of the cells have a low amount of food and cover, as shown in the bottom left of Fig. 6.4, and as illustrated by the marginal distributions plots for food and cover availability in Fig. 6.5.

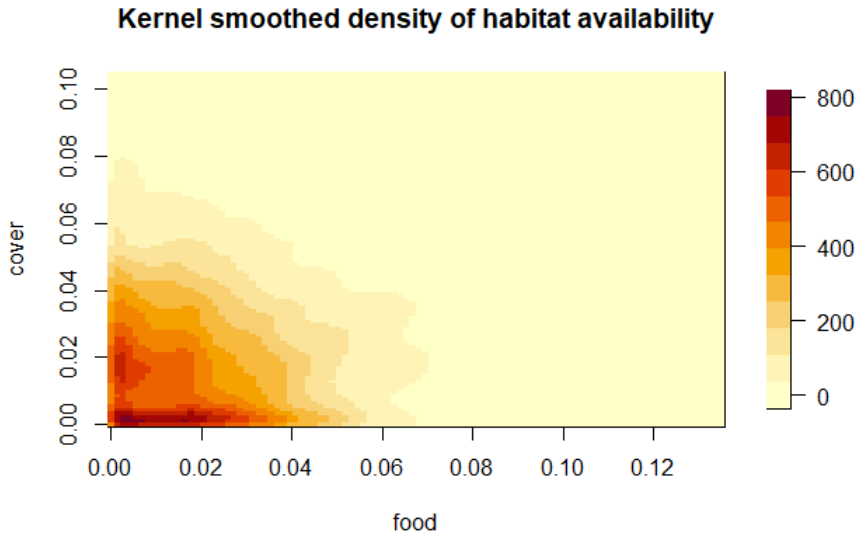


Figure 6.4: Kernel-smoothed density of habitat availability ($f(\mathbf{x})$ in Eq. (6.3)) in Matthiopoulos et al.'s (2011) dataset.

Fig. 6.6 is the multiplication of the kernel-smoothed habitat availability surface, $f(\mathbf{x})$ in Eq. (6.3), by all of the selectivity coefficients maps in Fig. 6.1 that have been derived from the regularized GFR and RBF-GFR models.

For a better view of the low-intensity regions in Fig. 6.6, I used the log-transformation

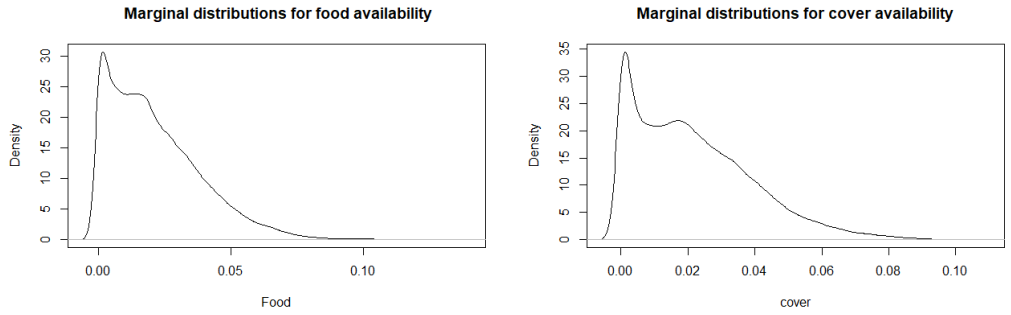
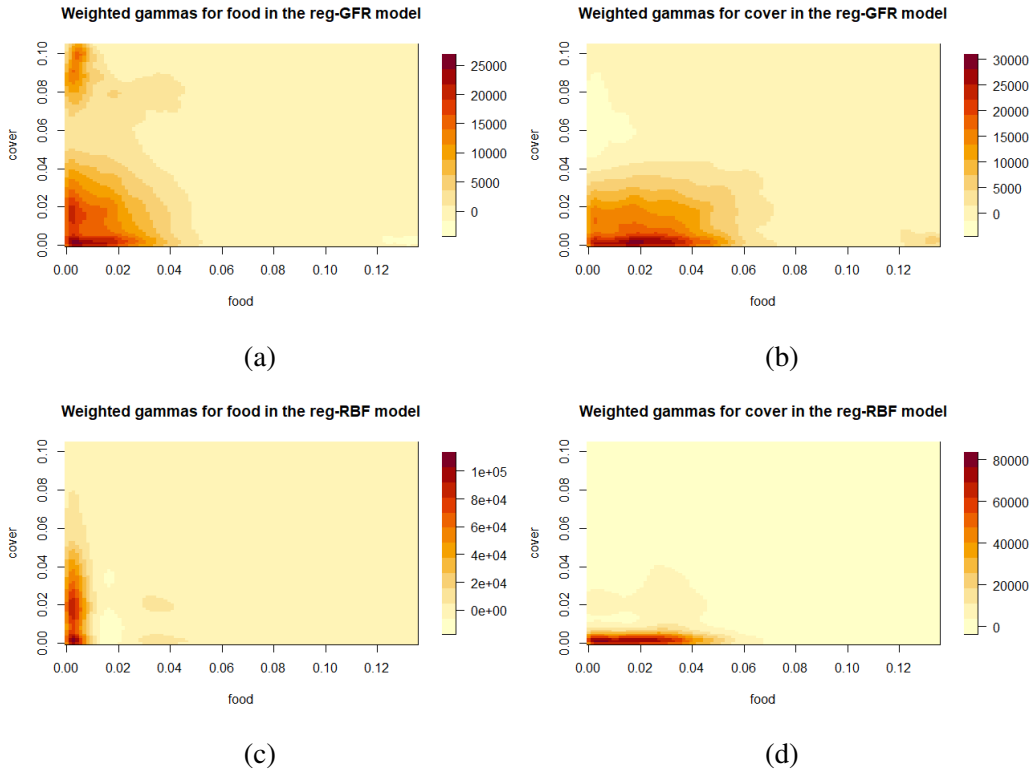


Figure 6.5: The marginal distributions for food availability in the left panel and cover availability in the right panel in Matthiopoulos et al.'s (2011) dataset.



(a) (b)

(c) (d)

Figure 6.6: Kernel-smoothed density of habitat availability ($f(\mathbf{x})$ in Eq. (6.3)) multiplied by the selectivity coefficients plots for food (a) and cover (b) using the regularized GFR model and food (c) and cover (d) using the regularized RBF-GFR model.

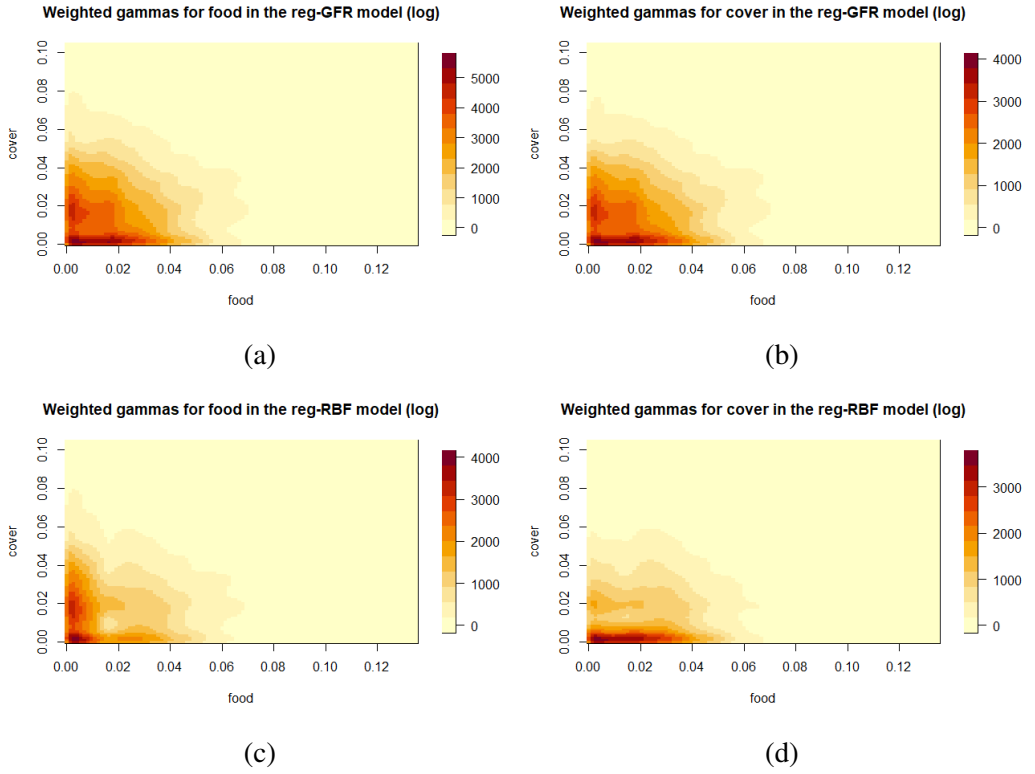
function, which changes the scale such that low-intensity pixels are shown at a high resolution and high-intensity pixels are compressed. It is used for image enhancement and to discern patterns by expanding the values of dark pixels while compressing the higher values (Manikpuri and Yadav, 2014). The log-transformation function used is:

$$t(z_i) = \log(z_i - min(\boldsymbol{z}) + \varepsilon) \tag{6.5}$$

where ε and $min(\boldsymbol{z})$ are used to ensure $z_i > 0$ (here, I used $\varepsilon = 1$). The transformed selectivity coefficients across environmental space were multiplied by the kernel-smoothed habitat availability as $t(\gamma_i(\mathbf{x}))f(\mathbf{x})$. Fig. 6.1 shows that the extreme values of the selectivity coefficients (high preference of the variable) are located in different areas, depending on the model and variable. However, the selectivity coefficients might be significant in a certain area, but the kernel-smoothed probability density of the availability of this area is small, leading to a downgrade of the high impact of this area by the smoothed availability, which will have very little influence on the βs, as seen in Fig. 6.7.

Fig. 6.7 shows a broad consistency between models as a result of applying the correction of the habitat availability $f(\mathbf{x})$, which focuses on frequently encountered habitats. The regions that have a maximum impact of selectivity coefficients (i.e., the area around cover = 0.10 in the top left panel and food > 0.12 in Fig. 6.1b) are effectively not observable in the data and, therefore, can be thought of as representing spurious features generated by the model. Still, this difference does not affect the transferability of the models because the β values are affected by the probability density of the habitat availability, and the extreme values of the selectivity coefficients are in very low-density areas. This conclusion is illustrated by plotting the log transformation for the different values of a and den for food and cover from the two regularized models in Figs 6.2a, 6.3a, 6.2b, and 6.3b. Figs 6.8a, 6.9a 6.8b, and 6.9b show the log-gamma for the different values of a and den from the two models for food and cover, respectively; both models show consistent behaviour of the selectivity coefficients, which are upgraded in the high-density region by the kernel-smoothed density of habitat availability. In Fig. 6.8b, the log transformation of food selectivity coefficients has some features that do not appear in the other plots (i.e., Figs. 6.8a, 6.9a, and 6.9b), which are the white patches in the high-density regions. These

small light regions (i.e., white patches in Fig. 6.8b) are small positive values resulting after transforming negative values of the selectivity coefficients using a log transformation in regions located in the high kernel-smoothed density of habitat availability.

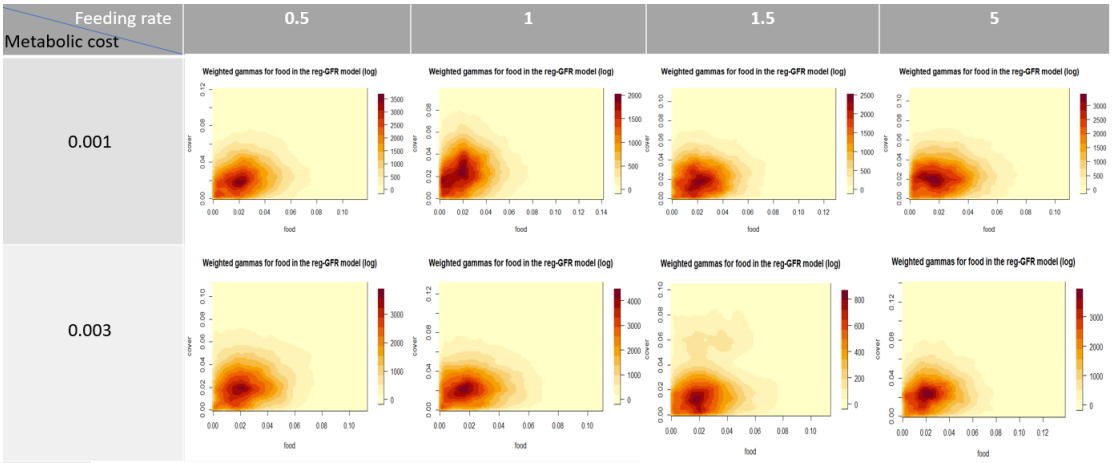


(a) (b)

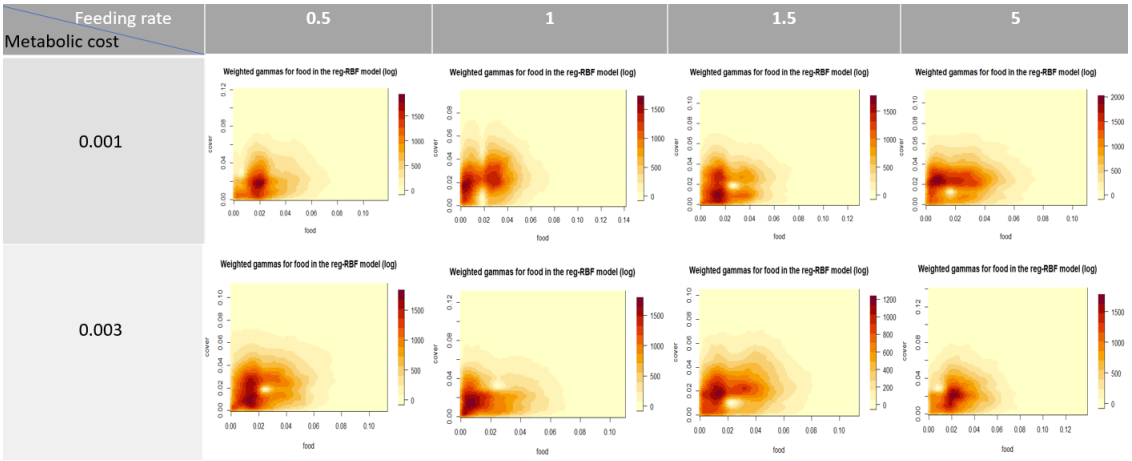
(c) (d)

Figure 6.7: Kernel-smoothed density of habitat availability ($f(\mathbf{x})$ in Eq. (6.3)) multiplied by log-gamma plots for food (a) and cover (b) for the GFR and food (c) and cover (d) for the RBF-GFR model.

Essentially, the features in the bottom left of each plot in Fig. 6.7 are a combination of how influential each particular habitat unit is and how often it has been encountered in the data. Interpreting these plots is complicated because they have the same features resulting from multiplying the log-transformation of selectivity coefficients by the kernel-smoothed density of the habitat availability function of the variables, which is used as an attention function. Instead of using the kernel-smoothed density of the availability function as an attention function, it can instead be used as a filter function to show the selectivity

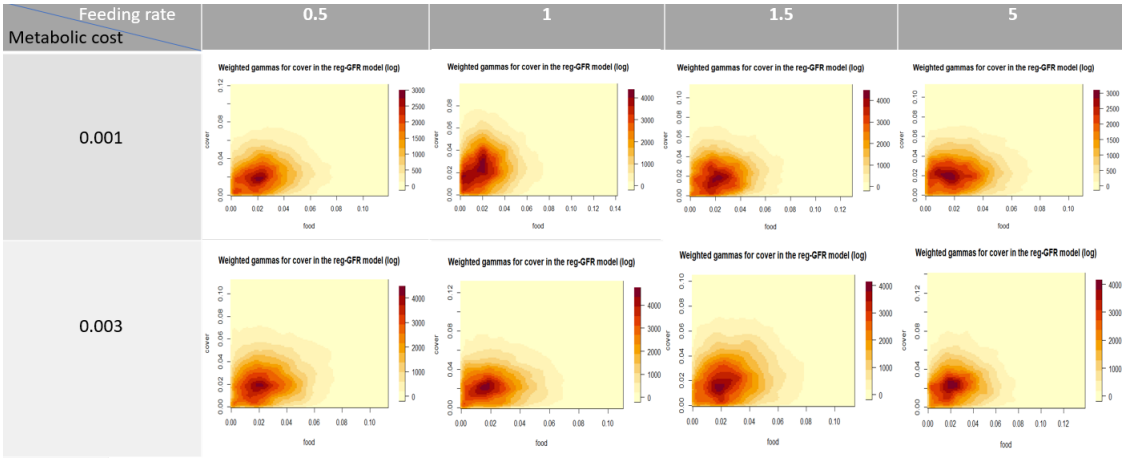


(a) Log-food selectivity coefficients (regularized GFR model)

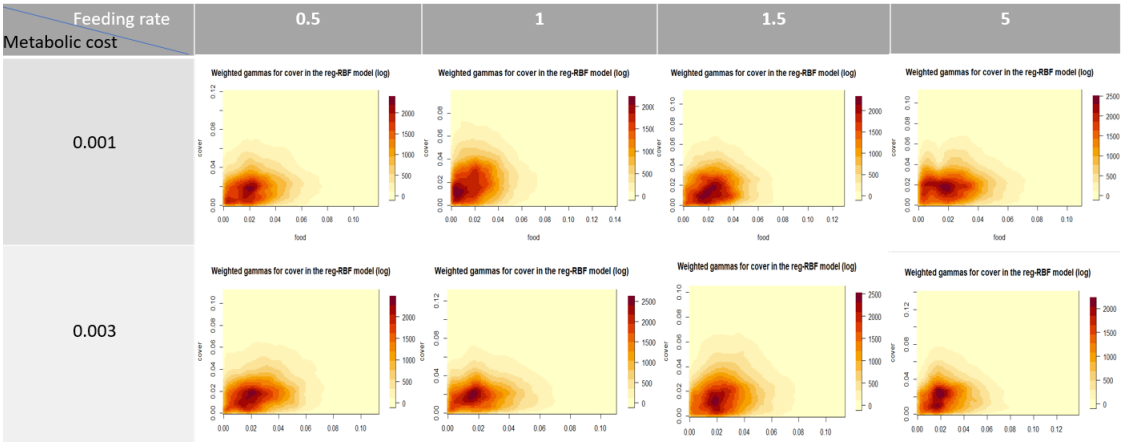


(b) Log-food selectivity coefficients (regularized RBF-GFR model)

Figure 6.8: (a) Log-transformation of selectivity coefficients for food from (a) the regularized GFR model and (b) the regularized RBF-GFR model for different values of a in the columns and den in the rows, where a = 0.5, 1, 1.5, 5 and den = 0.001 and 0.003.



(a) Log-cover selectivity coefficients (regularized GFR model)



(b) Log-cover selectivity coefficients (regularized RBF-GFR model)

Figure 6.9: (a) Log-transformation of selectivity coefficients cover from (a) the regularized GFR model and (b) the regularized RBF-GFR model for different values of a in the columns and den in the rows, where $a = 0.5, 1, 1.5, 5$ and $den = 0.001$ and 0.003.

coefficients only for those regions of plots that are sufficiently data-rich to be informative. The new plots in Fig. 6.10 allow us to compare the behaviour of the gammas in influential regions of environmental space, without the weighting previously applied by the function f. The focus now is on the configuration space, where the smoothed probability density is higher than 200 because there are extensive regions of environmental space below that threshold, as seen in Fig. 6.4. Based on the regularized GFR model in the top panels of Fig. 6.10, adding a habitat that is rare in food and has low amount of cover creates the strongest effect on the apparent preference of food, while adding a habitat that has intermediate richness in food and low amount of cover creates the strongest effect on the apparent preference of cover.

From the bottom panels of Fig. 6.10, adding one more unit of a habitat that contains rare food and rare or intermediate richness of cover increases the food preference in the habitat using the regularized RBF-GFR model. In addition, the strongest effect on apparent preference of cover occurs when add one unit of a habitat that has a rare or intermediate richness of food amount and low amount of cover using the same model, that is, the RBF-GFR model.

6.7 Qualitative Assessment of the Selectivity Coefficients

There is no direct correspondence between the statistical parameters $\gamma_i(\mathbf{x})$ and biological parameters used in the simulation such as the satiation threshold E_1, starvation threshold E_2, feeding rate a, and metabolic cost den, described in Section 4.2.2. However, referring to the mathematical model used to simulate this dataset described in Section 4.2.2, we can assess the selectivity coefficients by how plausible their values are based on the simulation rules. From the simulation rules, we know that the preference for both food and cover primarily depends on the food amount of a habitat because the behaviour of the simulated animal depends on the satiation threshold E_1 and starvation threshold E_2 when moving between feeding and hiding. The preference of food is consistent between the regularized GFR and RBF-GFR models, as seen from Figs. 6.10a and 6.10c, where adding a habitat that is poor in food makes food-rich patches rarer, causing the animals to show a stronger food preference mechanism, which is in the line with the simulation rule.

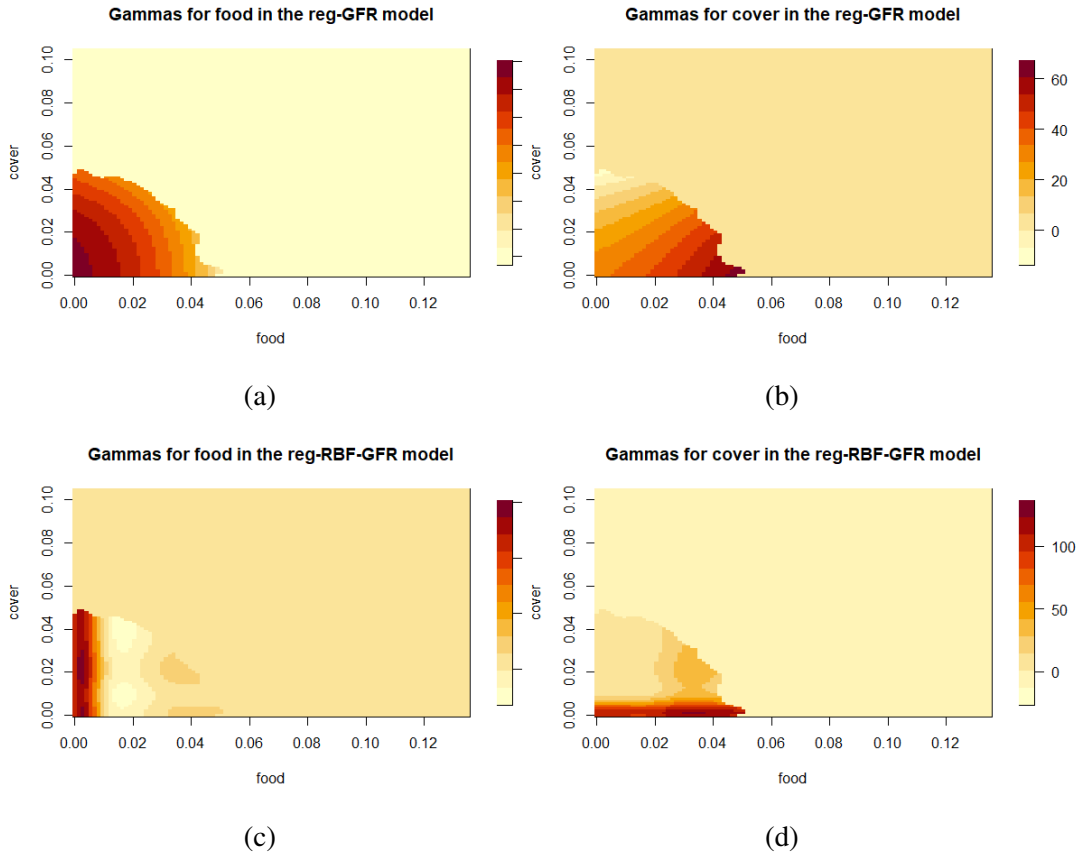


Figure 6.10: Selectivity coefficients for food (a) and cover (b) of the GFR and for food (c) and cover (d) of the RBF-GFR model using the kernel-smoothed density of the availability as a filter function (threshold is 200).

Because the response of cover depends on the food variable, not the cover variable, based on the simulation rule, the values of the selectivity coefficient of cover from both models are plausible. From Figs. 6.10b and 6.10c, if the food abundance is very low, the cover does not matter, which means the preference for cover decreases because the animal will starve and look for food. Both models conclude that the preference for cover increases as the food becomes more prevalent, which is consistent with the mathematical model of the simulation.

6.8 Conclusion

The behaviour of the selectivity coefficients of the regularized GFR and RBF-GFR models are different in both qualitative and quantitative terms, where the selectivity coefficient gradients are presented with different directions and steepnesses (Fig. 6.1). The different models (regularized GFR and RBF-GFR) are in agreement both qualitatively and quantitatively when looking at the availability-weighted behaviour of the log-selectivity coefficients (Fig. 6.7). This explains why the two models are able to explain the observed data in similar ways (i.e., produce similar values of the beta coefficients for each environmental scenario they are presented with). The different models are in qualitative agreement only by looking at the availability-filtered behaviour of the selectivity coefficients (Fig. 6.10); i.e., the gradients are positioned in the same direction, even if the steepnesses differ. The behaviour of the availability filtered of the selectivity coefficients using the regularized GFR and RBF-GFR models is consistent with the mechanisms generating this data. Furthermore, different selectivity coefficients can result in different betas and, hence, spatial predictions. However, the radically different selectivity coefficients of the regularized GFR and RBF-GFR models in regions of environmental space not often seen in the data result in similar betas, showing that the work here has moved a long way towards the aim of robust and transferable SDMs. This conclusion and interpretation of the selectivity coefficients is difficult to obtain without having the availability-filtered of the selectivity coefficients. The preference for both food and cover primarily depends on the food amount of habitat because the behaviour of the simulated animal depends on the satiation threshold E_1 and starvation threshold E_2 when moving between feeding and hiding. The behaviour

of the selectivity coefficient for food and cover is consistent between the regularized GFR and RBF-GFR models but the steepness is slightly different. The regularized RBF-GFR model is slightly better than the regularized GFR model because the selectivity coefficient behaviour of this model is more consistent with the mechanisms generating this data, especially when the food abundance is very low, the cover does not matter, which means the preference for cover decreases because the animal will starve and look for food and this is clearer with RBF-GFR model.

Chapter 7

Using GFRs to Predict Continental Patterns of Biodiversity

7.1 Introduction

The four validating datasets I have been using to develop the varying coefficient models are obtained from small-scale, single-species datasets. However, at larger scales, varying coefficient models are more likely to be useful because, across a large map, the prevailing conditions are likely to vary quite a lot. These expansive datasets would therefore play the role of multiple scenarios in the four validating datasets I have been using up until now. Furthermore, although larger-scale and multispecies datasets may be difficult to model or less interpretable, the analysis of these kinds of datasets better represents the reality of this species' world. Thus, the need to reflect this reality drives the demand for the mathematical interpretation of the world in which these animals live. In addition, there are emergent trends in how biodiversity increases or declines with ecological context. These patterns are not easily interpretable (certainly not in terms of the behaviour or energetics of the constituent species), but they may be no less predictable. Applying this modelling framework to multispecies biodiversity patterns is therefore a worthwhile phenomenological exercise, especially if it transpires that these GFR models can extend our predictive capability for biodiversity trends. Biodiversity is a univariate reduction of a multi-species community,

a statistical summary of ecosystem composition (i.e., the vector of species abundances). The statistical summary can be driven using information theory.

Information theory holds a central place in probability and statistics (Kullback, 1997). It aims to measure the amount of complexity needed to describe data or observed patterns (Brillouin, 2013); its first application in ecology was in 1955 (MacArthur, 1955; Ulanowicz, 2001) to measure biodiversity, which is a measure of variability in the species composition of ecological communities. The Shannon entropy score is the most frequently used measure of biodiversity in ecology (Sherwin and Prat i Fornells, 2019). It summarizes the information about species abundance within a spatiotemporal sampling unit or ecological community (Ricotta, 2002). In this chapter, I first model individual species distributions in the large scale North American Breeding Bird Survey (BBS) dataset by applying the generalized function response (GFR) model and various recent extensions using land cover types and the temperature of each segment as covariates. I assess the transferability achieved by using the GFR models by measuring the information content in the dataset under study. Second, I model the relationship between biodiversity and land cover types using the GFR models. I use the entropy score as my response variable in the biodiversity-habitat association models. Finally, I investigate the importance of legacy effects arising from extinction debts and colonisation credits (Haddou et al., 2022) in the GFR models of land cover types on biodiversity.

7.2 The Shannon Entropy Score

The Shannon entropy score is a popular diversity metric in ecology and can be defined as follows:

$$H(X) = -\sum_{i=1}^{m} p(x_i) log_2 p(x_i) \tag{7.1}$$

where m is the total number of classes and $p(x_i)$ is the proportion of individuals belonging to the i^{th} species in the dataset of interest.

The Shannon entropy score can be used to measure the uncertainty and randomness in the data (Higashi and Klir, 1982). The more randomness in a dataset, the higher will be the entropy and lower the information gain. If the entropy is low, information will be high.

To use entropy to express the information content in a community, I measured $p(x_i)$ for each species separately as the number of individuals in each sampling unit divided by the total number of individuals in the dataset, as follows:

$$H_s(X) = -\sum_{n=1}^{N} p_s(x_n) log_2 p_s(x_n) \tag{7.2}$$

where n is the sampling unit (segment of route in the BBS dataset), N is the number of sampling units, s refers to the species in the dataset and $p_s(x_n)$ is defined as:

$$p_s(x_n) = \frac{number\ of\ individuals\ that\ belong\ to\ species\ s\ in\ cell\ n}{total\ number\ of\ individuals\ that\ belong\ to\ species\ s\ in\ the\ dataset}$$

Here, $H_s(X)$ is the entropy score for each species being calculated over all sampling units in the dataset. In this case, a high entropy score is interpreted as low information content in the data because having similar abundance scores of a species in all sampling units is not informative. A variation of the abundance measures in the dataset leads to a lower entropy score, which leads to a higher information content in the dataset.

Furthermore, the Shannon entropy score is used here to observe the effect of land cover type on the biodiversity of species in each sample unit. In this situation, $p(x_i)$ is the number of individuals of each species in a sample unit divided by the total number of individuals in the same sample unit, as follows:

$$H_n(X) = -\sum_{s=1}^{S} p_n(x_s) log_2 p_n(x_s) \tag{7.3}$$

where S is the total number of species in the dataset and $p_s(x_n)$ is calculated as follows:

$$p_n(x_s) = \frac{number\ of\ individuals\ that\ belong\ to\ species\ s\ in\ cell\ n}{total\ number\ of\ individuals\ of\ all\ species\ \ in\ cell\ n}$$

Here, $H_n(X)$ is the entropy score for each sample unit being calculated over all species abundance scores in each sampling unit. The entropy score for each sampling unit is used as the response variable in the GFR models, with land cover types and temperature as the covariates to model the relationship between biodiversity and land cover types to predict

biodiversity in a new dataset using the GFR models. For example, the entropy score for each sample unit $H_n(\boldsymbol{z};\boldsymbol{\theta})$ using the GFR model can be expressed as a function of fixed effects of covariates and pairwise interactions between covariates and their moments:

$$H_n(\boldsymbol{z};\boldsymbol{\theta}) \;=\; \exp\left\{ \gamma_{0,0} + \sum_{i=1}^{I} \left(\sum_{m=0}^{M_i} \delta_{0,i}^{(m)} E[X_{n,i}^m]_b + \gamma_{i,0} x_{n,i} + x_{n,i} \sum_{j=1}^{I} \sum_{m=0}^{M_j} \delta_{i,j}^{(m)} E[X_{n,j}^m]_b \right) \right\} \tag{7.4}$$

where $\boldsymbol{\theta}$ is a parameter vector composed of the parameters γ_i and δ_i and $\boldsymbol{z}$ is a vector combining habitat variables $x_{n,i}$ in cell n and their expectation values $E[X_{n,i}^m]$, as well as their product terms.

Because the legacy effects in the relationship between habitat changes and species responses are common in nature (Haddou et al., 2022; Daskalova et al., 2020; Lira et al., 2019; Sala et al., 2000), the entropy score in Eq. (7.3) is used to investigate the effect of time lags of land cover covariates on the prediction of species biodiversity.

7.3 Increasing the Scale of the Neighbourhood Used to Characterise Environmental Context

In the present Chapter, I used the large-scale BBS dataset described in Section 4.3.3. The land cover covariates are the percentage of the forest, grass, urban, crop, wet, water, and barren within a 400 m buffer around the segment from which bird abundances were taken. These covariates have been used as habitat features and to represent the habitat availability $f_b(\mathbf{x})$ in Eq. (2.3) when applying the GFR, regularized GFR, GFR-CART, GFR-RF and GFR-XGBoost models. The models were applied to look at single species abundance patterns at the first stage. I used the out-of-sample scores to evaluate the models' transferability, as described in Section 2.9. After observing that one point is insufficient to describe the environment of each block p and that using just five description points of 400 m buffer to describe the 40 kilometres route (point per 8 kilometres) are insufficient to describe the environment of each route, I used 1 kilometre of radius instead of 400 m around each selected stop points (the 1^{st}, 11^{st}, 21^{st}, 31^{st} and 41^{st} stop points) in each route and a smaller

buffer (100 m) around 19 points selected randomly within the 1 kilometre buffer as the availability points, here considering the percentage of the land cover types within a 100 m buffer around the availability points, as seen in Fig. 7.1. Fig. 7.1 represents the process of increasing the scale of the neighbourhood used to characterise the environmental context in the model. Each route was represented using 5 stop points, the response variable for each stop point was the abundance of each species, and the covariates were the land cover types (habitat features) within a 400 m buffer of the stop points. After observing that 5 points are insufficient to describe the environment of the 40-kilometre route (point per 8 kilometres), I include more points around each stop point to use them as availability points in the model. The response variable does not change (the abundance of each species for each stop point), but the covariates are the land cover types within a 1-kilometre buffer of the stop points, which include 19 additional points with a 100 buffer each to describe the environment in the 1 kilometre. These availability points are used to increase the habitat description around each selected stop point by calculating the moments $E[X_j^m]_b$ in Eq. (2.5). Thus, the availability points increase the scale of the environmental variables.

7.4 Performance Evaluation

In the first stage, I assessed the transferability of the GFR models using the out-of-sample performance of different models to measure the abundance prediction of each bird separately using the land cover covariates and temperature of each segment, as described in Section 4.3.3. I split the dataset into two parts: the training set was for 2001 to 2016, and the testing dataset was for 2019 and used years as the sample instance (blocks). To increase the ability of the model to predict out-of-sample data, I used the state route partitions in each year as the blocks and extended the dataset using the availability points, as described in Section 7.3. The entropy score for each species being calculated over all sampling units in the dataset in Eq. (7.2) is used here to measure the information gain of each species' abundance in the BBS dataset. Furthermore, the entropy score for each sample unit being calculated over all species abundance scores is used as a response variable to find the effect of land cover types on the biodiversity of the ten birds in the BBS dataset, as described in Eq. (7.3). The same entropy score in Eq. (7.3) is used to investigate the importance of

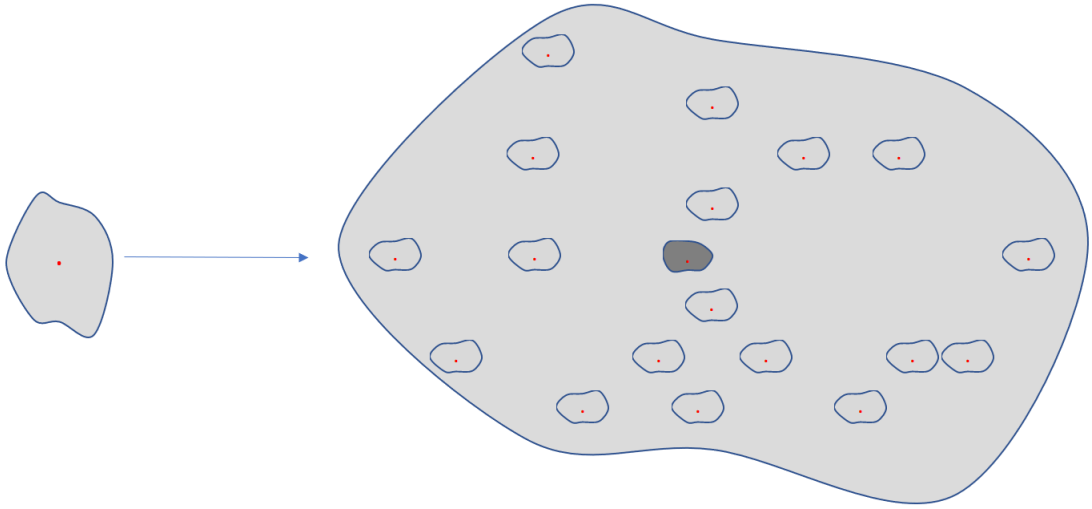


Figure 7.1: A diagram explaining the availability points used to increase the description around each selected point. The left polygon is the 400 m buffer around each stop point I used, and the right polygon is the new 1 kilometre buffer around each stop point. The small polygons are the 100 m buffers around the randomly selected points within the 1 kilometre buffer whereas the dark grey polygon contains the selected stop point.

legacy effects in the GFR models of land cover types on biodiversity.

7.5 Results

7.5.1 The GFR Models

To measure the transferability of the GFR models, the out-of-sample performance of different models was used to measure the abundance prediction of each bird separately using the land cover covariates and temperature of each segment, as described in Section 4.3.3. I split the dataset into two parts: the training set was for 2001 to 2016, and the testing dataset was for 2019. The out-of-sample performance of the GFR models was poor, as seen with the Mourning Dove in Table 7.1. The poor performance of the transferability of a model could be a result of several reasons, such as using an insufficiently flexible model, insufficient geographic separations, missing important predictors, or a lack of eco-

logical information (Yates et al., 2018). The extended variants of the GFR model provided non-negligible improvements in predictive performance when applied to four different datasets, as seen in Chapter 5, so I did not consider insufficient model flexibility as a plausible explanation. Using the years as sample instances (blocks) was insufficient in terms of transferability because the data was a large-scale dataset of more than 700 routes in the United States with very different geographical features across the states.

Table 7.1: The out-of-sample R^2 scores of the original GFR with its extended models using years as blocks in Mourning Dove.

Method	GFR	Reg-GFR	GFR-CART	GFR-RF	GFR-XGBoost
R^2	-18044.39	-0.740	0.069	0.183	0.115

To increase the ability of the model to predict out-of-sample data different from those used for model fitting, I used the state route partitions in each year as the sample instance (blocks). I have aimed to improve model performance by increasing the description of the landscape around each observation, as described in Section 7.3. A 40 kilometres route with just five segments is not sufficiently homogeneous; using just five description points of 400 m buffer to describe a route of 40 kilometres (point per 8 kilometres) is insufficient to describe the environment of each route. To describe the environment around each stop point, I used the stop point in addition to the 19 additional availability points (a total of 20 points to describe the environment of each stop point). Including more points is better to increase the predictive power, but is computationally expensive. A collection of 19 observations around each point of birds to describe the environmental features around each survey point shows a significant improvement of the model's transferability based on the results of the out-of-sample R^2 scores, specifically for the GFR model's transferability using state route partitions in each year as blocks and 19 availability points around each survey point in Table 7.2 compared with Table 7.1.

Although the GFR model's transferability increased using the availability points, the predictive ability of the GFR models is still considered poor based on the out-of-sample R^2 scores. This poor transferability of the GFR models was investigated numerically by checking for multicollinearity between the covariates and regularizing the model parameters. The multicollinearity was checked by the variance inflation factor (VIF), which is a

Table 7.2: Out-of-sample R^2 of the standard and GFR models for all species in the BBS dataset using state route partitions in each year as blocks and 19 availability points around each survey point.

Species	GLM	GFR	Reg-GFR	GFR-CART	GFR-RF	GFR-XGBoost
Mourning Dove	0.023	0.051	0.050	0.020	0.064	0.058
Red-winged Blackbird	0.073	0.184	0.195	0.187	0.462	0.327
American Crow	-0.026	-0.032	-0.025	-0.032	-0.107	-0.007
Blue Jay	0.002	0.009	0.008	-0.010	-0.081	-0.044
Common Yellowthroat	0.047	0.102	0.10	0.018	0.150	0.069
Barn Swallow	0.003	-4.79	-0.051	-0.016	0.025	0.079
Brown-headed Cowbird	0.05	0.073	0.094	0.075	0.184	0.130
Chipping Sparrow	-0.011	-0.137	-0.038	-0.081	0.077	0.055
European Starling	-0.01	-0.021	-0.014	-0.267	-0.064	0.020
American Robin	0.039	0.105	0.092	0.046	0.316	0.280

measure of the amount of multicollinearity in the multiple regression variables. I removed the variables that have VIF scores larger than five to reach an acceptable value for VIF (Gareth et al., 2013; Menard, 2002) and then applied the model to the other variables. Conceptually, multicollinearity might not affect the predictive performance, but when estimating the model's parameters, a rank-deficient matrix must be inverted. Therefore, in practice, if multicollinearity exists, $\mathbf{X}$ in Eq. (2.52) is not full rank, and its columns are linearly dependent. Therefore, the matrix $\mathbf{X}^T\mathbf{X}$ becomes a singular matrix and has no inverse. I avoided the problem by regularizing the model (ridge regression). While the expression $\mathbf{X}^T\mathbf{X}$ can be singular, the expression $\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}$ in Eq. (2.58) is mathematically non-singular for any $\lambda > 0$, and the inversion is numerically stable if lambda is sufficiently large. However, that basically leads to a biased model. For that reason, I also tried the alternative approach by removing parameters to avoid having rank deficient matrix and removing multicollinearity. Still, the reduced and regularized models did not improve the prediction performance. Both the reduced model after removing variables with high VIF scores and the regularization approach illustrate that poor transferability is not a multicollinearity consequence.

7.5.2 Entropy Score Uses

7.5.2.1 Information Gain

The poor predictive ability of a model does not always indicate a problem with model mis-specification. It could also be a result of errors in the data or missing covariates. If the information in a dataset is insufficient, using a complex model will not improve the transferability of the model. A high entropy score indicates that a noninformative dataset results from having similar density everywhere; the distribution is spread more equally across a lot of segments (Bishop, 2006). In contrast, if the density is very low for some segments but high for others, the entropy score will be low, which indicates that the information content is high in the dataset. The Shannon entropy score was found for each species in the data to measure the amount of information in each species dataset separately. For the BBS dataset, I calculated the entropy score to measure the information gain of each species abundance using Eq. (7.2) when the number of individuals in each cell is divided by the total number of individuals of each species, as shown in Table 7.3.

Table 7.3: The Shannon entropy scores for all species using Eq. (7.2) in the BBS dataset.

Species	Entropy score
Mourning Dove	8.655
Red-winged Blackbird	8.089
American Crow	8.386
Blue Jay	7.647
Common Yellowthroat	7.980
Barn Swallow	7.246
Brown-headed Cowbird	7.552
Chipping Sparrow	8.215
European Starling	7.212
American Robin	8.788

For comparison, I provided the entropy scores for the two real datasets (wolf and sparrow) in Table 7.4. The entropy scores for all species in the BBS datasets in Table 7.3 are higher than the entropy scores of the more precise datasets in Table 7.4, indicating that the BBS dataset has low information content. I experimented with additional information

Table 7.4: The Shannon entropy scores for sparrow and wolf datasets using Eq. (7.2).

Dataset	Entropy score
Wolf	2.15
Sparrow	5

to increase the models' transferability. For example, using the abundance of other species as additional covariates aimed to capture the effects of species interactions (competition, mutualism, and predation) on the distribution of any one focal species. Table 7.6 and Fig. 7.2 show the results of the out-of-sample R^2 scores using the standard model with and without the effect of other species abundances, indicating that the out-of-sample performance of most of the species that consider other species abundance is better than the result without. To test whether the impact of other species abundances is the missing information that can significantly increase the transferability of the models, I used the parametric t-test and non-parametric Wilcoxon test to compare the out-of-sample R^2 scores before and after adding other species abundance to the model (Stevens, 2013). Both tests were used because the parametric t-test is more powerful than the non-parametric test if the distributional assumption is valid, which is hard to check for with such a small sample size, as seen in Fig. A.31 in Appendix A.10, and the t-test is sensitive to outliers. Because the scores were calculated for the same species before and after and the two sets are paired, the following t-test was used for this case:

$$H_0 : \mu_D = 0 \qquad vs \qquad H_1 : \mu_D > 0$$

where μ_D is the mean of the difference between the scores before and after adding other species abundance. The following hypothesis was used when the Wilcoxon test was applied: $H_0 : m_1 = m_2$ vs $H_1 : m_1 \neq m_2$

where m_1 and m_2 are the medians of the scores before and after adding other species abundance, respectively. Both tests were used to test the following hypotheses:

H_0: The difference between the out-of-sample R^2 before and after adding other species abundance is zero

versus

H_1: There is a non-zero difference between the out-of-sample R^2 before and after adding other species abundance

Table 7.5: P-values of the t-test and Wilcoxon test for comparing the out-of-sample R^2 scores for the GFR models before and after adding other species' abundance measures to the models.

Test	P-value
T-test	0.669
Wilcoxon test	0.084

Table 7.5 shows the p-values of the t-test and Wilcoxon test for comparing the out-of-sample R^2 scores for the GFR models before and after adding other species' abundance measures. Based on the p-values, at the 5% significance level, I do not reject the null hypothesis, as illustrated in Fig. 7.3, when the two boxes of the scores for both cases seem to overlap. There is no sufficient evidence in the data to reject the hypothesis that the two means and medians in the populations are similar, so the impact of other species' abundances is not the missing information that can increase the transferability of the model.

Table 7.6: Out-of-sample R^2 for the standard (GLM) for all species in the BBS dataset before and after including other species' abundance as additional covariates.

Species	Without other species abundance	With other species abundance
Mourning Dove	0.023	0.033
Red-winged Blackbird	0.073	0.128
American Crow	-0.026	0.036
Blue Jay	0.002	0.007
Common Yellowthroat	0.047	0.052
Barn Swallow	0.003	-0.511
Brown-headed Cowbird	0.05	0.08
Chipping Sparrow	-0.011	0.021
European Starling	-0.009	0.002
American Robin	0.039	0.096

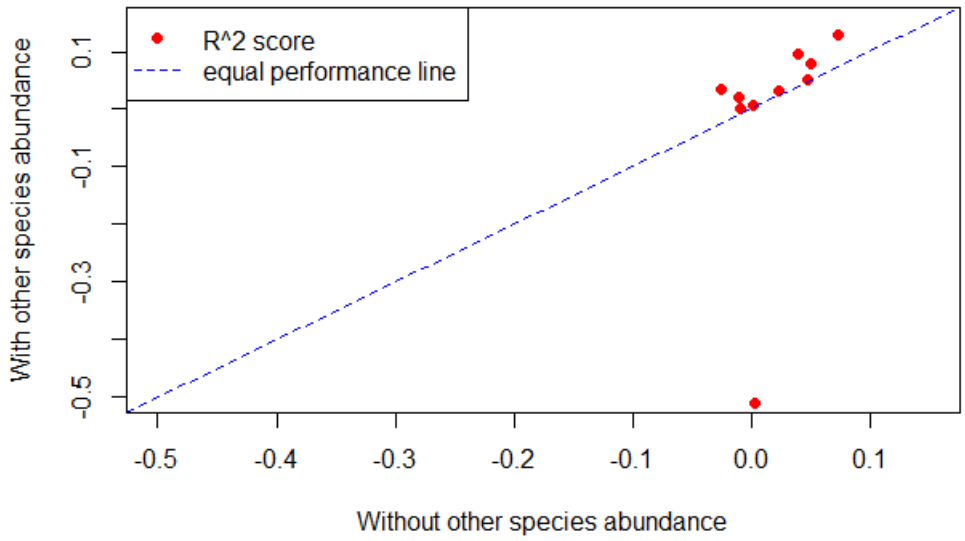


Figure 7.2: Scatter plot comparing the out-of-sample R^2 scores for the standard (GLM) model for all species in the BBS dataset without including other species' abundance measures as additional covariates in the horizontal axis and with including other species' abundance measures in the vertical axis, where each red dot refers to a species and the blue dashed line is the line of equal performance. The out-of-sample R^2 scores are better than the scores without including other species abundance, but the difference is not significant based on the p-values of the t-test (p-value = 0.669) and Wilcoxon test (p-value =0.084) at the 5% significance level. I concluded that there were no significant differences between the out-of-sample R^2 scores before and after including other species' abundance scores.

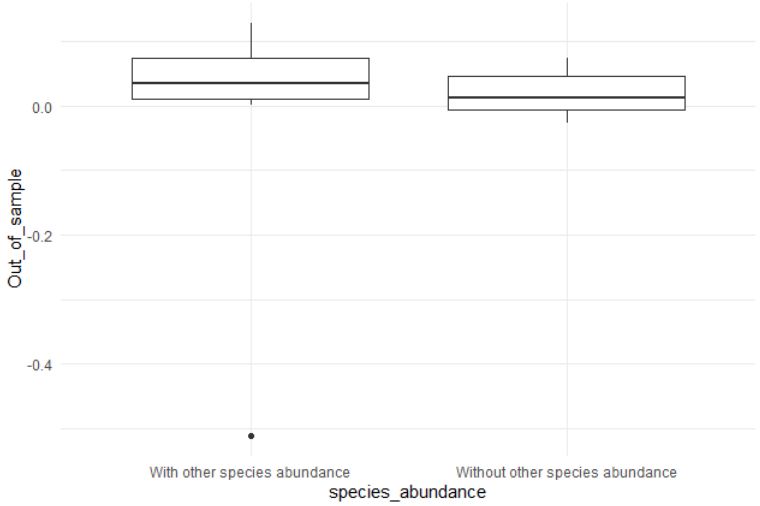


Figure 7.3: Box plots of the out-of-sample R^2 scores for the standard (GLM) model for all species in the BBS dataset after including other species' abundance measures as additional covariates in the left and before including other species' abundance measures in the right panel. The partial overlap of the boxes illustrates my finding from the p-values of the t-test (p-value = 0.669) and Wilcoxon test (p-value = 0.084) at the 5% significance level, showing that there are no significant differences between the out-of-sample R^2 scores before and after including other species' abundance scores.

7.5.2.2 Biodiversity of Species

A previous study has found evidence for the effect of land cover types on the biodiversity of bird species in the BBS dataset (Haddou et al., 2022). I used the GFR models to explore whether the dependence of the coefficients of the model on different land cover compositions could improve the transferability of these models. The response variable was the Shannon entropy scores of all species for each segment, calculated using Eq. (7.3), regressed against land cover types and temperate. I found that biodiversity patterns can be better predicted using the GFR model than the standard GLM model, as seen in Table 7.7. The efficient in-sample performance of GFR-XGBoost and GFR-RF models indicates that the model has the flexibility to learn the process. The transferability of the GFR-RF model is better than the other models, as shown in Table 7.7, because the RF approach guards to some extent against over-fitting. The increase in both the in-sample and out-of-sample

scores as the model flexibility increase indicates no over-fitting issue. However, there is a discrepancy between the in-sample and out-of-sample R^2 scores in the ensemble models. Any significant difference between the land cover types in the training and testing sets could cause discrepancies between the in-sample and out-of-sample scores. To make this comparison, I used the 2016 year dataset as a training set, which is the closest year to the test set in the data, that is, the 2019 set. I wanted to compare the outcomes from the same model using the training and testing sets to investigate the significant difference between the in-sample and out-of-sample R^2 scores. Thus, I used the same model applied to these two datasets without removing any terms. Fig. 7.4 shows the histograms of the entropy scores of the training and testing sets, where the height of each bar indicates the number of locations that has an entropy score within the corresponding bin. The two histograms show the agreement of the entropy scores in the training and testing sets, which is illustrated by the scatter plots in Fig. A.32 in Appendix A.11.

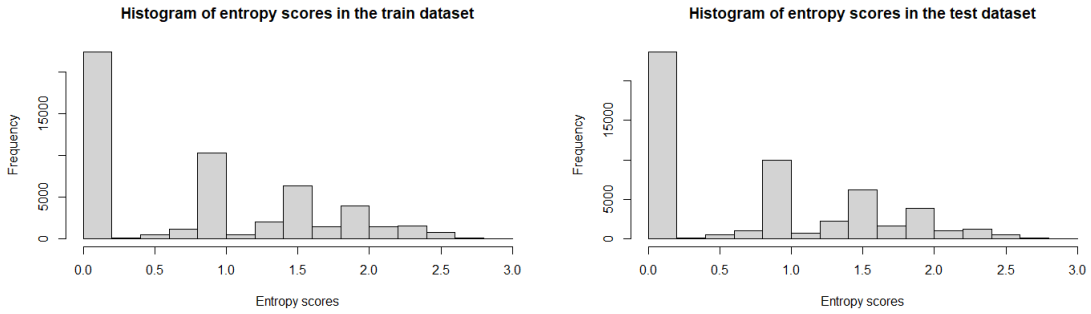


Figure 7.4: Histogram of the entropy scores of the training set in the left panel and the testing set in the right panel, where the height of each bar indicates the number of locations that has entropy scores within the corresponding bin.

The t-test and Wilcoxon test results in Tables 7.8 and 7.9 illustrate that there are discrepancies between the land cover covariates in the training and testing datasets, which are supported by Figs. A.33 and A.34 in Appendix A.11. I found the mean decrease in accuracy and SHAP feature importance scores, as described in Sections 2.7.1.1 and 2.7.2.1. For the random forest method, I extracted the importance measures for all the variables used in the random forest by applying the models to the training data, and retraining the model

using the test dataset to check if the variables would have a different level of importance in the test and training datasets. The importance scores were calculated using the mean decrease in accuracy from permuting out-of-bag data; the details for this calculation can be found in Section 2.7.1.1. Fig. 7.5 shows that the most important five variables using the GFR-RF model in the training set differ from the most important variable after retraining the same model using the test set. For the extreme gradient boosting model, I found SHAP feature importance scores measured as the mean absolute Shapley values, as described in Section 2.7.2.1. The most important five variables using SHAP feature importance scores of the GFR-XGBoost model in the training set differ from the most important variable by retraining the same model using the test set, as shown in Fig. 7.6. The scatter plots, parametric test, non-parametric test, importance score from the RF approach, and importance scores for the XGBoost illustrate that there is a discrepancy between the training set and test set that caused the difference between the in-sample and out-of-sample scores in Table 7.7.

Table 7.7: In-sample and out-of-sample scores for the GLM, GFR, REG-GFR, GFR-CART, GFR-XGBoost, and GFR-RF models when the entropy score is the response variable.

Models	In-sample R^2	Out-of-sample R^2
GLM	0.230	0.163
GFR	0.294	0.260
Reg-GFR	0.288	0.256
GFR-CART	0.263	0.240
GFR-XGBoost	0.948	0.293
GFR-RF	0.905	0.385

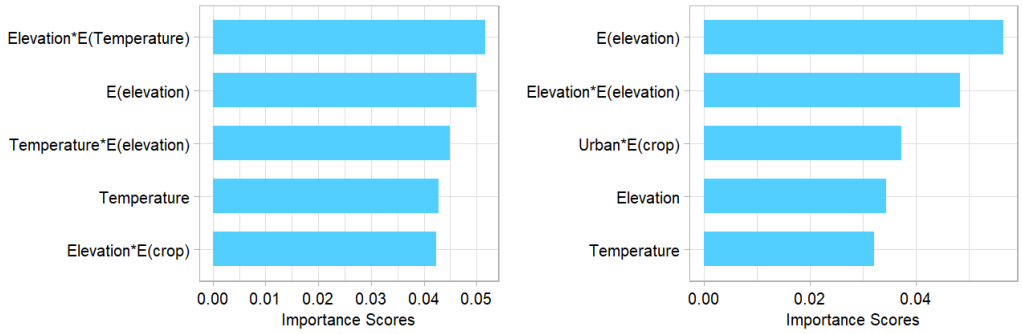


Figure 7.5: Importance scores using the mean decrease in accuracy for the most important five variables in the GFR-RF model using the training dataset (2016) in the left panel and test dataset (2019) in the right panel.

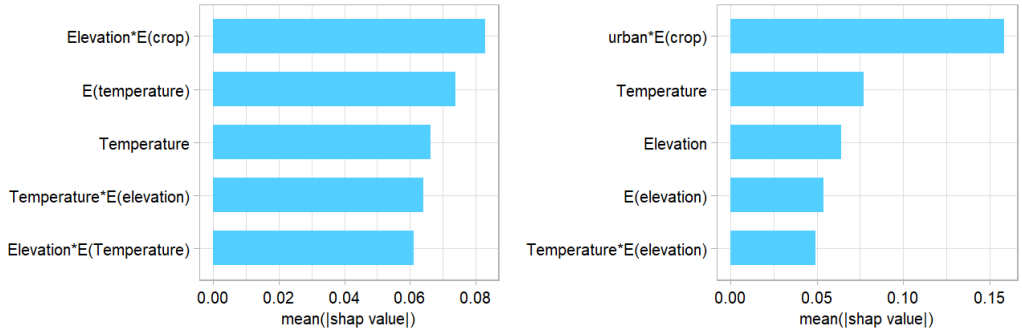


Figure 7.6: SHAP feature importance scores using the mean absolute Shapley values for the highest five variable scores in the GFR-XGBoost model using the training dataset (2016) in the left panel and test dataset (2019) in the right panel..

Table 7.8: T-test of the land cover variables in the training set (2016) vs. the test set.

variable	p-value (T-test)	conclusion at 5%
urban	<2.2e-16	significant difference
forest	7.474e-05	significant difference
grass	4.243e-07	significant difference
crop	0.05394	no significant difference
wet	0.0004768	significant difference
water	0.4053	no significant difference
barren	5.88e-06	significant difference
temperature	<2.2e-16	significant difference
elevation	0.765	no significant difference

Table 7.9: Wilcoxon test of the land cover variables in the training set (2016) vs. the test set.

variable	p-value (Wilcoxon test)	conclusion at 5%
urban	<2.2e-16	significant difference
forest	3.415e-05	significant difference
grass	2.51e-07	significant difference
crop	0.01662	significant difference
wet	8.771e-06	significant difference
water	0.2224	no significant difference
barren	0.0001715	significant difference
temperature	<2.2e-16	significant difference
elevation	0.575	no significant difference

7.5.2.3 Legacy Effect

Biodiversity often occurs with time lags in response to land cover types (Haddou et al., 2022; Daskalova et al., 2020; Lira et al., 2019; Sala et al., 2000). The BBS dataset I used in the present thesis covers the species abundance of the years 2001, 2004, 2006, 2008, 2011, 2013, 2016, and 2019 with gaps of 3, 2, 2, 3, 2, 3, and 3 years, respectively. To investigate the effect of time lags of the BBS dataset using the GFR models, I used the biodiversity scores in 2016 with the land cover type and temperature from 2013, the biodiversity scores in 2013 with the land cover type and temperature from 2011, the biodiversity scores in

2011 with the land cover type and temperature from 2008, the biodiversity scores in 2008 with the land cover type and temperature from 2006, the biodiversity scores in 2006 with the land cover type and temperature from 2004, and the biodiversity scores in 2004 with the land cover type and temperature from 2001 as the training set to predict the biodiversity scores in 2019, as shown in Table 7.10. Fig. 7.7 shows the out-of-sample and in-sample R^2 scores for the GFR models with time lags, as presented in Table 7.10 versus without time lags shown in Table 7.7.

Table 7.10: In-sample and out-of-sample scores for the GLM, GFR, reg-GFR, GFR-CART, GFR-XGBoost, and GFR-RF models when the entropy score is the response variable with a delay effect.

Models	In-sample	Out-of-sample
GLM	0.204	0.224
GFR	0.292	0.260
Reg-GFR	0.268	0.257
GFR-CART	0.263	0.240
GFR-XGBoost	0.961	0.303
GFR-RF	0.904	0.385

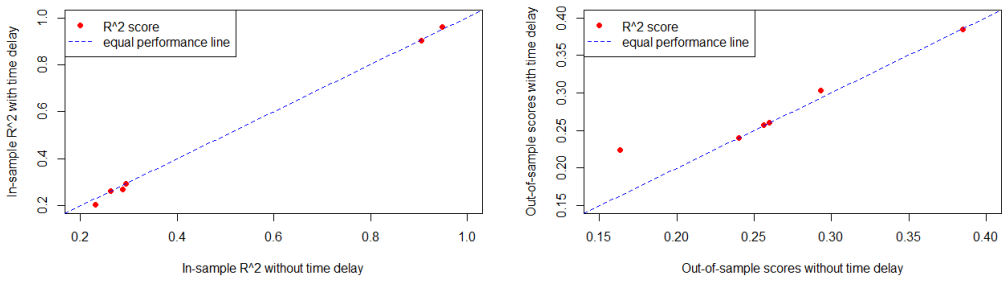


Figure 7.7: In-sample R^2 for the GFR models with versus without time lags in the left panel. The right panel is the out-of-sample R^2 scores for the GFR models with versus without time lags .

To test whether biodiversity in the BBS dataset occurs with time lags in response to land cover covariates, I used the paired t-test and Wilcoxon test to compare the out-of-sample R^2 scores, with and without including the delay effect using the GFR models.

Because the scores were calculated for the same models before and after including time lags, the two sets are paired, and the paired tests were used. Both tests were used because six scores are inadequate for the normality test, as seen in Figs. 7.8 and 7.9.

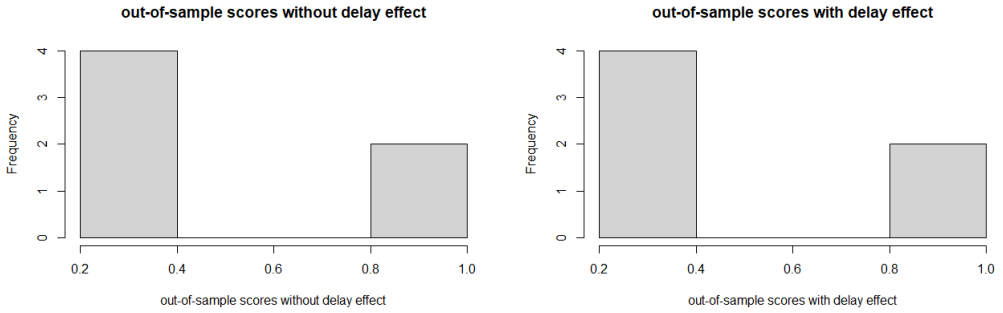


Figure 7.8: Histogram to check if the normality assumption is valid of the in-sample R^2 scores for the GFR models without time delay in the left panel and with time delay in the right panel.

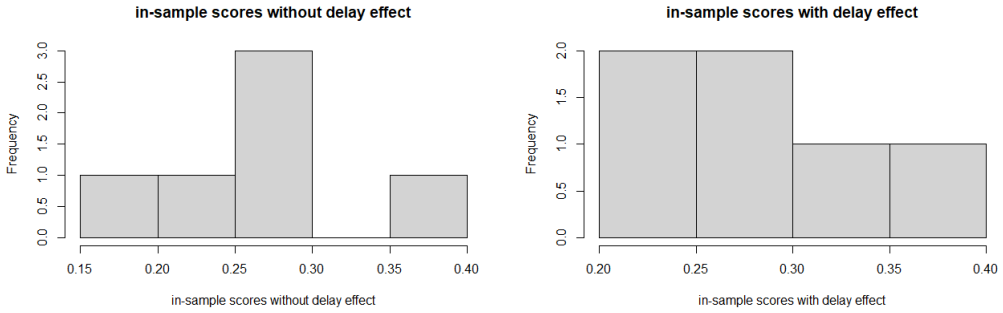


Figure 7.9: Histogram to check if the normality assumption is valid of the out-of-sample R^2 scores for the GFR models without time delay in the left panel and with time delay in the right panel.

Based on the p-values in Table 7.11 for the out-of-sample R^2 scores and in-sample R^2 scores, at the 5% significance level, I do not reject the null hypothesis that the two means and medians in the populations are the same, so biodiversity in the BBS dataset does not occur with time lags in response to land cover covariates when using the GFR models, as

Table 7.11: P-values of the t-test and Wilcoxon test for comparing the in-sample and out-of-sample R^2 scores with and without including the delay effect using the GFR models.

Test	in-sample scores p-value	out-of-sample scores p-value
T-test	0.823	0.141
Wilcoxon test	0.281	0.182

illustrated in Fig. 7.10, where the two boxes in the left panel for the in-sample scores and the two boxes in the right panel for the out-of-sample scores seem to overlap.

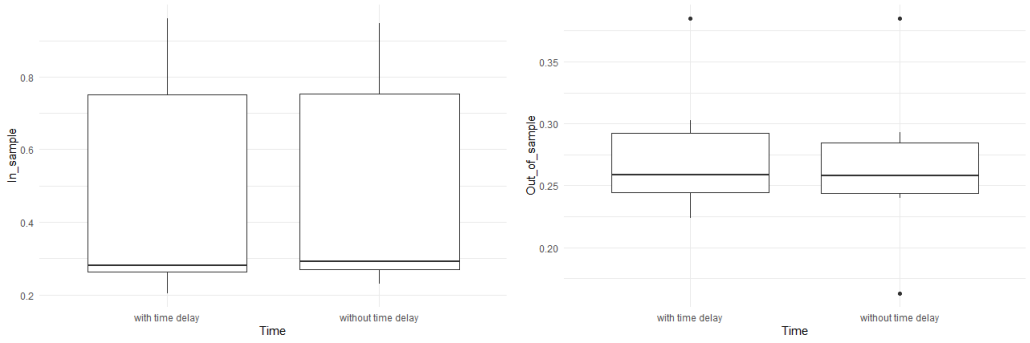


Figure 7.10: Box plots of the in-sample R^2 scores for the GFR models with and without time delay in the left panel. The right panel is the out-of-sample R^2 scores for the GFR models with and without a time delay. The partial overlap of the boxes illustrates my finding from the p-values of the t-test (p-value = 0.823) and Wilcoxon test (p-value = 0.281) for the in-sample scores and the p-values of the t-test (p-value = 0.141) and Wilcoxon test (p-value = 0.182) for the out-of-sample scores at the 5% significance level, indicating that there are no significant differences between the in-sample and out-of-sample R^2 scores for the GFR models with and without a time delay.

7.6 Conclusion

The challenging large-scale North American Breeding Bird Survey BBS dataset was used for several purposes in this chapter. The GFR model and its various extensions were used to model individual species distributions, using the species abundance of each segment as the response variable and the land cover (urban, forest, grass, crop, wet, water, elevation

and temperature) as the covariates. The Shannon entropy score was used to investigate three different aspects of these data. First, the entropy score for each species, calculated over all sampling units in the dataset, was used to investigate the poor transferability of the original GFR and its various recent extensions by measuring the information content in the dataset under study. I found that the information in the dataset was insufficient after measuring the entropy score of each species' abundance. This could be the reason for the poor predictive ability of each species' abundance using the GFR models. Second, the Shannon entropy score for each sample unit was calculated over all species abundance scores and used as a response variable to observe spatial patterns of biodiversity and to explore the ability of GFR tools to increase the predictive ability of these models. The GFR-RF model doubled the predictive power of the biodiversity model compared with the generalized linear model. Finally, using GFR models, I investigated the possibility of legacy effects on biodiversity in response to land cover change. Biodiversity in the BBS dataset did not occur with time lags in response to land cover covariates when using the GFR models. This conclusion conflicts with a recent study by Haddou et al. (2022) which, by analysing the same dataset, found that the past landscape composition predominantly affects the current effective number of species. This discrepancy could be for a number of reasons. For example, Haddou et al. (2022) model was not limited to landscape composition but also took into account some other variables not included in my model such as observer effects, the effective number of land cover types, time of day effects, between-route variation, and a quadratic fixed effect for temperature. However, most importantly, although my model was designed to quantify neighbourhood effects, Haddou et al. (2022) model was designed to quantify the exact contribution of past and present landscapes as a weighted function. Furthermore, the model in Haddou et al. (2022) was used to train the current biodiversity using past and current landscapes, but my model was designed to use the past landscape and current biodiversity to predict future biodiversity. However, the future debts and credits were used to test the direction predicted by Haddou et al. (2022) model and quantified by a Pearson correlation, R=0.28, which can be considered a low score. The time gap used to model the legacy effect is another difference: a 15-year window in Haddou et al. (2022) model and a 3-year window in my model.

The BBS dataset suffers from various inherent and well-documented limitations that

reduce its suitability for detecting species distribution patterns. The collection of the dataset was based on a single time window (breeding season) for each route, ignoring the abundance during the non-breeding season (Rosenberg et al., 2017). The BBS dataset suffers from a lack of remote geographic regions and coverage of hard-to-detect species (Rosenberg et al., 2017). Heterogeneity in observer experience is ignored in the BBS dataset resulting in very different efficiency of species identification (Peterjohn, 2001; Sauer et al., 1994). Dataset collection is from a random collection of roadside routes across North America, and these routes are not representative of the landscape in the regions (Peterjohn, 2001; Hanowski and Niemi, 1995). The observer records every seen or heard bird for three minutes, which is a technique shown to have deficiencies for sampling bird species (Peterjohn, 2001; Verner, 1985). Counts in stop points in an area do not provide complete counts of all birds present in that area (Peterjohn, 2001; Barker and Sauer, 1995). Another reason why the data are imperfect is that the collection of information does not occur every year (Currie and Venne, 2017). Even with all of these limitations that the very large scale, multispecies, error-prone, imperfect BBS dataset has, all of which reduce the data suitability for detecting the distribution patterns, the extended GFR models offer improvement in transferability of biodiversity, and the GFR-RF model doubled the predictive power of biodiversity compared with the GLM and original GFR models.

Chapter 8

Overall Conclusions

Given their extensive applications to questions of anthropogenic change (Iturbide et al., 2018) it is imperative that the predictive ability of SDMs be assessed by transferring models built in one region or time to another spatiotemporal frame where the prevailing environmental conditions are different, and possibly outside the range of covariate values previously measured (Duque-Lazo et al., 2016; Elith and Leathwick, 2009). The discrepancy between phenomenological models and the complex biological mechanisms they try to capture leads to the existence of highly non-linear functional responses in species-habitat associations (Mysterud and Ims, 1998). These are especially complicated in the cases of animals with higher mobility and cognition. Despite the increasingly recognised challenges of model transferability (Petitpierre et al., 2017; Yates et al., 2018; Wenger and Olden, 2012; Peterson et al., 2003; Randin et al., 2006; Townsend Peterson et al., 2007; Barbosa et al., 2009; Sundblad et al., 2009; Wenger et al., 2011), particularly for purely statistical models such as SDMs, there is a dearth of methods for functional responses in SDMs and a lack of comparative validation of such methods with synthetic and real data.

Here, I have built on the suggestions of Boyce and McDonald, 1999 and the early implementation of the GFR by Matthiopoulos et al., 2011. My work has addressed the lack of flexibility and control in the original GFR model and investigated how alternative models might be implemented within the broader GFR framework. Replacing the global polynomial functions of the original GFR with radial basis functions as well as using the Gaussian mixtures approximation to approximate habitat availability allowed the RBF-

GFR model to be more flexible than the original GFR model. This flexibility poses the risk of over-fitting, the opposite behaviour to the rigidity of classic SDMs (Paton and Matthiopoulos, 2016). Over-fitting is a fundamental issue for achieving transferable GFR models (Wenger and Olden, 2012). In this respect, I found regularization approaches to be effective in controlling over-fitting in GFRs.

I have also explored the suggestion made by recent publications that modern machine learning methods achieve better results than traditional statistical methods (Heikkinen et al., 2012; Elith* et al., 2006; Lawler et al., 2006; Prasad et al., 2006). I have achieved this by combining the RBF-GFR and GFR models with CART methods and, as a further extension, I have used ensemble approaches, random forests (RF) and extreme gradient boosting (XGBoost).

Ignoring structure dependence in data increases the susceptibility to over-fitting and causes autocorrelations and non-independence of model residuals (Roberts et al., 2017). The block cross-validation approach addresses the autocorrelation of dataset structures (Roberts et al., 2017). I have implemented the block cross-validation approach to account for autocorrelation of dataset structures. The simulated dataset in Matthiopoulos et al. (2015) was derived from multiple instances, each comprising 500 observations representing a sub-population in a different landscape (Matthiopoulos et al., 2015). The dataset has a spatial structure based on these scenarios. I have used these scenarios as dataset blocks when I applied the models. I have used a cross-validation approach based on these blocks (10-cross-validation where each fold contained 40 blocks) to measure the out-of-sample predicted performance. The simulated dataset in Matthiopoulos et al. (2011) is a simpler version of the Matthiopoulos et al. (2015) dataset consisting of 20 blocks (sample instances). I set the blocks to be the folds, meaning that I have used a 20-block cross-validation approach. In the sparrow population dataset, I have used the 32 colonies as the blocks and folds, resulting in 32-block cross-validation. The wolf dataset has a grouping structure based on the five packs that the wolves belong to. I have used five-blocks cross-validation based on these five packs when applying the models.

The resulting performance of the GFR, RBF-GFR and their extensions on four datasets showed that considerable gains in predictive performance could be achieved and that these were approximately consistent across data sets. Going from a global median regression

GFR model to the radial basis function model, offered local flexibility in the functional response curves but generated only moderate improvements in out-of-sample R^2 score. However, combining the ensemble approach using bagging and boosting with the GFR and RBF-GFR models substantially improved the out-of-sample R^2 scores. In general, ensemble methods, such as bagging and boosting were consistently among the top-scoring models, with no evidence of over-fitting, while for other models, performance varied more drastically. In essence, the original GLM model provides a much flatter version of the ground truth (i.e., under-predict the actual values), while applying overly flexible extensions of the GFR model can increase the risk of exaggerating extremes in species distribution (i.e., over-predicting abundance hot-spots/peaks and under-predicting cold-spots/troughs). I also replicated the finding that model ensembles can perform the same role as regularization, buffering the models' predictions from such variances of exaggeration.

It has been clear that simple SDMs homogenise predictions (Paton and Matthiopoulos, 2016) and that polynomial GFRs can be overly volatile. The key message from my work is that using measures against over-fitting (i.e., either regularization or ensemble modelling) can give consistent and impressive improvements in out-of-sample predictions, in some cases raising the R^2 from 0.25 to 0.85 (typical gains were from 0.35 to 0.80). This comes at a cost of implementation. The libraries required for fitting these models are not as user-friendly as the base GLM approaches, so some work will be needed in the future to develop automation in software workflows for functional responses. A key advantage of such approaches is that regularization is an efficient way of achieving a parsimonious models so, in addition to GFR flexibility, it would simultaneously facilitate issues of covariate selection.

The differences in performance between different data sets are as interesting as the consistent features of Table 5.14, but considerably harder to explain. Improvements in predictive performance were most dramatic in the two real data sets (the sparrows and wolves), despite the fact that simulated data sets were designed to offer better adherence to the spatial stationarity of covariates and distributional assumptions made by the models fitted to those data. Several reasons have been mentioned for poor transferability in the literature. For example, the poorer information content of occupancy compared to abun-

dance data (Yates et al., 2018), the definition of the scale of habitat availability (Barbosa et al., 2009; Paton and Matthiopoulos, 2016; Beyer et al., 2010), or the ranging behaviour of the study species (Vanreusel et al., 2007; Yates et al., 2018, Wogan, 2016). The suggestion of stratifying the data by the type of behavioural activity (Yates et al., 2018) as a solution to the problem of varying conditions might improve the model's transferability under the auspices of the GFR family of models.

In addition to the marked improvements in the predictive performance of the RBF-GFR and the extensions of GFR and RBF-GFR compared to the original GFR and GLM models, it is also essential to know how plausible these models are by offering some explanatory power of the mechanisms mediating species distributions. Understanding these mechanisms and rules helps assess the models' transferability by their ability to infer these mechanisms.

The assessment was made by visualizing the selectivity coefficients $\gamma_i(\mathbf{x})$ of the best two models, regularized GFR and regularized RBF-GFR, for the simulated dataset in Matthiopoulos et al. (2011). The availability-weighted behaviour of the log-selectivity coefficients $\gamma_i(\mathbf{x})$ of the two models is similar. The two models produce similar values of the selection coefficients $\beta_{i,b}$ for each environmental scenario they are presented with because the availability-weighted selectivity coefficients have similar behaviour. Most importantly, although the behaviour of the selectivity coefficients ($\gamma_i(\mathbf{x})$) is significantly different, the availability-filtered behaviour of the selectivity coefficients ($\gamma_i(\mathbf{x})$) is in qualitative agreement, and these coefficients are consistent with the mechanisms generating these data. This illustrates that the two models have moved towards the goal of robust and transferable SDMs.

The implementation of the improved versions of the GFR model was tested on small-scale and single-species datasets in the first part of my thesis. However, it was important to test the transferability of models for a large-scale (continent-wide), multi-species dataset. At larger scales, varying coefficient models are more likely to be useful, because, across a large map, such as the continental USA, prevailing conditions are likely to be varying a lot. The insufficient information content in the dataset, which was measured here by the Shannon entropy score, could be a reason for the poor predictive power of species distribution of the challenging large-scale North American Breeding Bird Survey BBS dataset, which

is known for its inherent limitations mentioned in the literature (i.e., Rosenberg et al., 2017; Peterjohn, 2001; Sauer et al., 1994; Peterjohn, 2001; Hanowski and Niemi, 1995; Verner, 1985; Currie and Venne, 2017). Furthermore, the Shannon entropy score was used to quantify the appropriateness of the GFR models to observe biodiversity patterns in the imperfect BBS dataset using the land cover covariates. Although the BBS data suffers from inherent limitations, the GFR-RF model was characterised by twice the predictive power of fixed-coefficient models of biodiversity and the original GFR model. The GFR models did not detect any effects of time lags in biodiversity in response to land cover covariates, which conflicts with a recent study by Haddou et al. (2022). Adding more covariates can improve model performance not only in detecting the effect of time lags in biodiversity but also can improve the predictive power of species abundances using the GFR models of the land cover covariates using the BBS dataset. Including historical bird data at each stop point location as an additional covariate in the BBS dataset would further improve the BBS population trend estimates (Hudson et al., 2017). Other covariates, such as wind turbines, can significantly affect the abundance of the species in the BBS dataset (Miao et al., 2019). Furthermore, Li et al. (2020) found that an increase in neonicotinoid use led to statistically significant reductions in bird biodiversity in the BBS dataset. Including more descriptions of the landcover type can improve the prediction of species abundance (Cazalis et al., 2019), where the abundance of forest species in protected forest sites is significantly higher than in unprotected forests.

For the RBF model and its extension models, the RBF approach is used to model $\gamma_i(\mathbf{x})$ in Eq. (3.3). All the explanatory variables in the two simulated, sparrow and BBS datasets are continuous variables, so extracting the parameters of the basis functions was possible. However, some datasets such as the wolf dataset, contained binary variables or factors, thus making the application of the basis function approach invalid because these binary variables can only take one of two values, which is impossible to set the centre and bandwidth parameters, $\xi_{j,m}$ and $\sigma_{j,m}$, for m basis functions from these two values. Therefore, the RBF-GFR models could be improved to deal with explanatory binary variables. For instance, I may try using the decision tree by modifying the model, where the binary variables can be used for the decision nodes and the continuous variables for the input to the model associated with the leaf nodes.

The parameters of the basis function are currently fixed. I could try making the basis function adjustable, and then the model becomes a neural network, which can improve the approximation in addition to its several advantages such as its ability to provide good generalization (Yu et al., 2011), and eliminate the effect of outliers (Chen and Jain, 1994). The backpropagation algorithm can be used to adjust the radial basis functions neural network parameters (Bishop, 1995). In addition, another flexible model, such as the spline approach rather than the radial basis function or polynomial function, can be used to describe $\gamma_i(\mathbf{x})$. The spline approach is a local function and more flexible than the polynomial function, which addresses the limitation of the global polynomial functions by dividing the input domain into regions and then fitting the polynomial function in each region. The basis spline function requires fewer parameters compared to the radial basis function, and hence it is computationally easier in order to derive β_i. Furthermore, the ability to predict in out-of-sample datasets increases when I apply the CART and ensemble models, which are non-linear models. As a result of this improvement, applying more flexible non-linear models, such as generalized additive models (GAMs) or spline regression in combination with the GFR and RBF-GFR, might increase the models' transferability.

The work done here instills computational robustness into a method that has been previously shown to work and opens the avenue for further comparative studies and biological interpretation. Better visualisation methods for how regression coefficients in species-habitat association models adapt to changes in overall habitat composition provide a link between these de-facto phenomenological models with some quintessentially mechanistic fields of environmental sciences, particularly behavioural and landscape ecology. For example, the analysis of the functions $\gamma_i(\mathbf{x})$ in Eq. (3.1), as derived from my GFR models, has clear parallels with models of consumer choice developed in the areas of ethology (Sih and Christensen, 2001) and the humanities (Raghavarao et al., 2010).

Similarly, by extending the SDMs to account for regional environmental context, the models might provide clues about more holistic processes at the level of landscape ecology. Therefore, the GFR, an approach that begun with the sole aim of trying to improve predictive performance may, through the generation of new hypotheses for habitat selection, lead to new insights about fundamental biology at the level of the individual and the landscape.

Appendix A

Additional Appendices

A.1 Derivation Details of the RBF-GFR Model

The simplified ζ in Eq. (3.12) is obtained using the following simplification steps of Eq. (3.11), as follows:

$$\zeta = \exp\left[\left(\ln\frac{1}{\sqrt{[\sigma^2_{j,j}]_b+\sigma^2_{j,m}}}\right)-\frac{1}{2}\left(\frac{(x_j-\xi_{j,m})^2\cdot\frac{[\sigma^2_{j,j}]_b}{[\sigma^2_{j,j}]_b+\sigma^2_{j,m}}+(x_j-[\mu_{j,k}]_b)^2\cdot\frac{\sigma^2_{j,m}}{[\sigma^2_{j,j}]_b+\sigma^2_{j,m}}}{\frac{[\sigma^2_{j,j}]_b\cdot\sigma^2_{j,m}}{[\sigma^2_{j,j}]_b+\sigma^2_{j,m}}}\right)\right]$$

$$= \exp\left[\left(-\frac{1}{2}\ln\left([\sigma^2_{j,j}]_b+\sigma^2_{j,m}\right)\right)-\frac{1}{2}\left(\frac{(x_j-\xi_{j,m})^2\cdot\frac{[\sigma^2_{j,j}]_b}{[\sigma^2_{j,j}]_b+\sigma^2_{j,m}}+(x_j-[\mu_{j,k}]_b)^2\cdot\frac{\sigma^2_{j,m}}{[\sigma^2_{j,j}]_b+\sigma^2_{j,m}}}{\frac{[\sigma^2_{j,j}]_b\cdot\sigma^2_{j,m}}{[\sigma^2_{j,j}]_b+\sigma^2_{j,m}}}\right)\right]$$

$$= \exp\left[-\frac{1}{2}\left[\ln\left([\sigma^2_{j,j}]_b+\sigma^2_{j,m}\right)+\left(\frac{(x_j-\xi_{j,m})^2\cdot\frac{[\sigma^2_{j,j}]_b}{[\sigma^2_{j,j}]_b+\sigma^2_{j,m}}+(x_j-[\mu_{j,k}]_b)^2\cdot\frac{\sigma^2_{j,m}}{[\sigma^2_{j,j}]_b+\sigma^2_{j,m}}}{\frac{[\sigma^2_{j,j}]_b\cdot\sigma^2_{j,m}}{[\sigma^2_{j,j}]_b+\sigma^2_{j,m}}}\right)\right]\right]$$

$$= \exp\left[-\frac{1}{2}\left[\frac{\ln\left([\sigma^2_{j,j}]_b+\sigma^2_{j,m}\right)\cdot\frac{[\sigma^2_{j,j}]_b\cdot\sigma^2_{j,m}}{[\sigma^2_{j,j}]_b+\sigma^2_{j,m}}+(x_j-\xi_{j,m})^2\cdot\frac{[\sigma^2_{j,j}]_b}{[\sigma^2_{j,j}]_b+\sigma^2_{j,m}}+(x_j-[\mu_{j,k}]_b)^2\cdot\frac{\sigma^2_{j,m}}{[\sigma^2_{j,j}]_b+\sigma^2_{j,m}}}{\frac{[\sigma^2_{j,j}]_b\cdot\sigma^2_{j,m}}{[\sigma^2_{j,j}]_b+\sigma^2_{j,m}}}\right]\right]$$

$$= \exp[-\frac{1}{2}[\frac{\ln\left([\sigma^2_{j,j}]_b+\sigma^2_{j,m}\right)\cdot\frac{[\sigma^2_{j,j}]_b\cdot\sigma^2_{j,m}}{[\sigma^2_{j,j}]_b+\sigma^2_{j,m}}+(x_j^2+\xi^2_{j,m}-2x_j\xi_{j,m})\cdot\frac{[\sigma^2_{j,j}]_b}{[\sigma^2_{j,j}]_b+\sigma^2_{j,m}}+}{\frac{[\sigma^2_{j,j}]_b\cdot\sigma^2_{j,m}}{[\sigma^2_{j,j}]_b+\sigma^2_{j,m}}}$$
$$\frac{(x_j^2+[\mu_{j,k}]_b^2-2x_j[\mu_{j,k}]_b)\cdot\frac{\sigma^2_{j,m}}{[\sigma^2_{j,j}]_b+\sigma^2_{j,m}}}{\frac{[\sigma^2_{j,j}]_b\cdot\sigma^2_{j,m}}{[\sigma^2_{j,j}]_b+\sigma^2_{j,m}}}]] \tag{A.1}$$

$$\zeta = \exp\left[-\frac{1}{2}\left[\frac{\kappa}{\frac{[\sigma^2_{j,j}]_b\cdot\sigma^2_{j,m}}{[\sigma^2_{j,j}]_b+\sigma^2_{j,m}}}\right]\right] \tag{A.2}$$

where

$$\kappa = \ln\left([\sigma^2_{j,j}]_b+\sigma^2_{j,m}\right)\cdot\frac{[\sigma^2_{j,j}]_b\cdot\sigma^2_{j,m}}{[\sigma^2_{j,j}]_b+\sigma^2_{j,m}}+(x_j^2+\xi^2_{j,m}-2x_j\xi_{j,m})\cdot\frac{[\sigma^2_{j,j}]_b}{[\sigma^2_{j,j}]_b+\sigma^2_{j,m}}+$$
$$(x_j^2+[\mu_{j,k}]_b^2-2x_j[\mu_{j,k}]_b)\cdot\frac{\sigma^2_{j,m}}{[\sigma^2_{j,j}]_b+\sigma^2_{j,m}} \tag{A.3}$$

$$\kappa = x_j^2-2x_j\left(\frac{\xi_{j,m}[\sigma^2_{j,j}]_b+\sigma^2_{j,m}[\mu_{j,k}]_b}{\sigma^2_{j,m}+[\sigma^2_{j,j}]_b}\right)+\left(\frac{\sigma^2_{j,m}[\mu_{j,k}]_b^2+\xi^2_{j,m}[\sigma^2_{j,j}]_b+\ln\left([\sigma^2_{j,j}]_b+\sigma^2_{j,m}\right)\cdot\left([\sigma^2_{j,j}]_b\cdot\sigma^2_{j,m}\right)}{[\sigma^2_{j,j}]_b+\sigma^2_{j,m}}\right)$$

$$\kappa = x_j^2-2x_jA+\left(A^2+C\right) \tag{A.4}$$

where $A = \left(\frac{\xi_{j,m}[\sigma^2_{j,j}]_b+\sigma^2_{j,m}[\mu_{j,k}]_b}{\sigma^2_{j,m}+[\sigma^2_{j,j}]_b}\right)$, and

$$A^2+C = \frac{\sigma^2_{j,m}[\mu_{j,k}]_b^2+\xi^2_{j,m}[\sigma^2_{j,j}]_b+\ln\left([\sigma^2_{j,j}]_b+\sigma^2_{j,m}\right)\cdot\left([\sigma^2_{j,j}]_b\cdot\sigma^2_{j,m}\right)}{[\sigma^2_{j,j}]_b+\sigma^2_{j,m}}$$

By solving for C:

$$C = \frac{\sigma_{j,m}^2[\mu_{j,k}]_b^2 + \xi_{j,m}^2[\sigma_{j,j}^2]_b + \ln\left([\sigma_{j,j}^2]_b + \sigma_{j,m}^2\right) \cdot \left([\sigma_{j,j}^2]_b \cdot \sigma_{j,m}^2\right)}{[\sigma_{j,j}^2]_b + \sigma_{j,m}^2} - \left(\frac{\xi_{j,m}[\sigma_{j,j}^2]_b + \sigma_{j,m}^2[\mu_{j,k}]_b}{\sigma_{j,m}^2 + [\sigma_{j,j}^2]_b}\right)^2$$

$$C = \frac{\left[\sigma_{j,m}^2[\mu_{j,k}]_b^2 + \xi_{j,m}^2[\sigma_{j,j}^2]_b + \ln\left([\sigma_{j,j}^2]_b + \sigma_{j,m}^2\right) \cdot \left([\sigma_{j,j}^2]_b \cdot \sigma_{j,m}^2\right)\right] \cdot}{\left([\sigma_{j,j}^2]_b + \sigma_{j,m}^2\right)^2} \qquad \text{(A.5)}$$
$$\left([\sigma_{j,j}^2]_b + \sigma_{j,m}^2\right) - \left(\xi_{j,m}[\sigma_{j,j}^2]_b + \sigma_{j,m}^2[\mu_{j,k}]_b\right)^2$$

Let $\ln\left([\sigma_{j,j}^2]_b + \sigma_{j,m}^2\right)$ equal α

$$C = \frac{\left[\sigma_{j,m}^2[\mu_{j,k}]_b^2 + \xi_{j,m}^2[\sigma_{j,j}^2]_b + \alpha \cdot \left([\sigma_{j,j}^2]_b \cdot \sigma_{j,m}^2\right)\right] \cdot \left([\sigma_{j,j}^2]_b + \sigma_{j,m}^2\right) - \left(\xi_{j,m}[\sigma_{j,j}^2]_b + \sigma_{j,m}^2[\mu_{j,k}]_b\right)^2}{\left([\sigma_{j,j}^2]_b + \sigma_{j,m}^2\right)^2}$$

$$C = \frac{\sigma_{j,m}^2[\mu_{j,k}]_b^2[\sigma_{j,j}^2]_b + \xi_{j,m}^2\sigma_{j,j}^4 + \alpha\sigma_{j,j}^4\sigma_{j,m}^2 + \sigma_{j,m}^4[\mu_{j,k}]_b^2 + \sigma_{j,m}^2\xi_{j,m}^2[\sigma_{j,j}^2]_b + \alpha[\sigma_{j,j}^2]_b\sigma_{j,m}^4}{\left([\sigma_{j,j}^2]_b + \sigma_{j,m}^2\right)^2}$$
$$+ \frac{-\xi_{j,m}^2\sigma_{j,j}^4 - \sigma_{j,m}^4[\mu_{j,k}]_b^2 - 2[\sigma_{j,j}^2]_b\sigma_{j,m}^2[\mu_{j,k}]_b\xi_{j,m}}{\left([\sigma_{j,j}^2]_b + \sigma_{j,m}^2\right)^2}$$

$$C = \frac{\sigma_{j,m}^2[\mu_{j,k}]_b^2[\sigma_{j,j}^2]_b + \alpha\sigma_{j,j}^4\sigma_{j,m}^2 + \sigma_{j,m}^2\xi_{j,m}^2[\sigma_{j,j}^2]_b + \alpha[\sigma_{j,j}^2]_b\sigma_{j,m}^4 - 2[\sigma_{j,j}^2]_b\sigma_{j,m}^2[\mu_{j,k}]_b\xi_{j,m}}{\left([\sigma_{j,j}^2]_b + \sigma_{j,m}^2\right)^2}$$

$$C = \frac{\sigma_{j,m}^2[\sigma_{j,j}^2]_b\left[[\mu_{j,k}]_b^2 + \alpha[\sigma_{j,j}^2]_b + \xi_{j,m}^2 + \alpha\sigma_{j,m}^2 - 2[\mu_{j,k}]_b\xi_{j,m}\right]}{\left([\sigma_{j,j}^2]_b + \sigma_{j,m}^2\right)^2}$$

$$C = \frac{\sigma_{j,m}^2[\sigma_{j,j}^2]_b\left[\left([\mu_{j,k}]_b - \xi_{j,m}\right)^2 + \alpha\left([\sigma_{j,j}^2]_b\sigma_{j,m}^2\right)\right]}{\left([\sigma_{j,j}^2]_b + \sigma_{j,m}^2\right)^2}$$

Finally,

$$C = \frac{[\sigma_{j,j}^2]_b \sigma_{j,m}^2}{\left([\sigma_{j,j}^2]_b + \sigma_{j,m}^2\right)^2} \cdot \left([\mu_{j,k}]_b - \xi_{j,m}\right)^2 + \frac{[\sigma_{j,j}^2]_b \sigma_{j,m}^2}{\left([\sigma_{j,j}^2]_b + \sigma_{j,m}^2\right)} \cdot \left(\ln\left([\sigma_{j,j}^2]_b + \sigma_{j,m}^2\right)\right)$$

By inserting Eq. (A.4) into Eq. (A.2), I get:

$$\zeta = \exp\left[-\frac{1}{2}\left[\frac{x_j^2 - 2x_j A + (A^2 + C)}{\frac{[\sigma_{j,j}^2]_b \cdot \sigma_{j,m}^2}{[\sigma_{j,j}^2]_b + \sigma_{j,m}^2}}\right]\right]$$

$$\zeta = \exp\left[-\frac{1}{2}\left[\frac{(x_j - A)^2 + C}{\frac{[\sigma_{j,j}^2]_b \cdot \sigma_{j,m}^2}{[\sigma_{j,j}^2]_b + \sigma_{j,m}^2}}\right]\right]$$

A.2 Comparison of the RBF-GFR model's parameter methods

The parameters of the RBFs, $\xi_{j,m}$ and $\sigma_{j,m}$, need to be determined in advance to find $[I_{j,m}]_b$ in Eq. (3.17). I used the histogram approximation and quantile approaches, discussed in Section 3.2.2, to select these parameters and the best method was chosen based on AIC and BIC. Table A.1 shows the result of the simulated dataset in Matthiopoulos et al. (2015), which was the first dataset used in this thesis. The number of basis functions (bins using histogram approximation and quantiles using the quantile approach) was varied from 1 to 13 and then the RBF-GFR model was applied to calculate AIC and BIC for comparison. The quantiles approach was chosen to select the RBF parameters since the model selection scores (AIC and BIC) using the quantile approach for most numbers of basis functions are less than the scores from using the histogram approach, as seen in Table A.1. Furthermore, Table A.2 provides a comparison of the AIC and BIC scores resulting

from using the histogram and quantile approaches to select the RBF parameters applying to the second simulated, sparrow, wolf datasets. The scores suggest that the quantiles approach is better based on AIC and BIC in the second simulated dataset. However, there is no difference between the score in the wolf dataset and just a slight difference between the score in the sparrow dataset; the scores from the histogram approach are slightly lower than the quantile approach scores. Based on these results, the quantile approach was used to determine the RBF parameters when applying the RBF-GFR model and its extensions.

Table A.1: The AIC and BIC scores of the RBF-GFR model in the first simulated dataset using the histogram and quantiles approachs to determine the basis function parameters for 1 to 13 basis functions.

Basis functions	AIC (histogram)	AIC (quantiles)	BIC (histogram)	BIC (quantiles)
1	929481.9	922463.9	929573.8	922555.8
2	919408.2	918543.3	919561.3	918696.4
3	915465.7	914724.1	915680	914938.4
4	912934.3	913538	913209.9	913813.6
5	910790.4	913027.4	911127.2	913364.2
6	910483.8	911026.5	910881.8	911424.5
7	910402.5	909267.2	910861.8	909726.5
8	908237.6	908344.1	908758.1	908864.7
9	907465.2	907256.1	908047	907837.8
10	907205.8	907205.8	907818.2	907818.2
11	907221.3	907205.8	907854.1	907818.2
12	907230.2	907205.8	907883.4	907818.2
13	907269.7	907205.8	907933.1	907818.2

Table A.2: The AIC and BIC scores of the RBF-GFR model for the second simulated, sparrow and wolf datasets using the histogram and quantiles approachs to determine the basis function parameters.

Dataset	AIC (histogram)	AIC (quantiles)	BIC (histogram)	BIC (quantiles)
Second simulated	595025.5	595015.3	595572.4	595544.5
Sparrow	1698.067	1698.687	1780.54	1781.161
Wolf	15546.34	15546.34	15881.76	15881.76

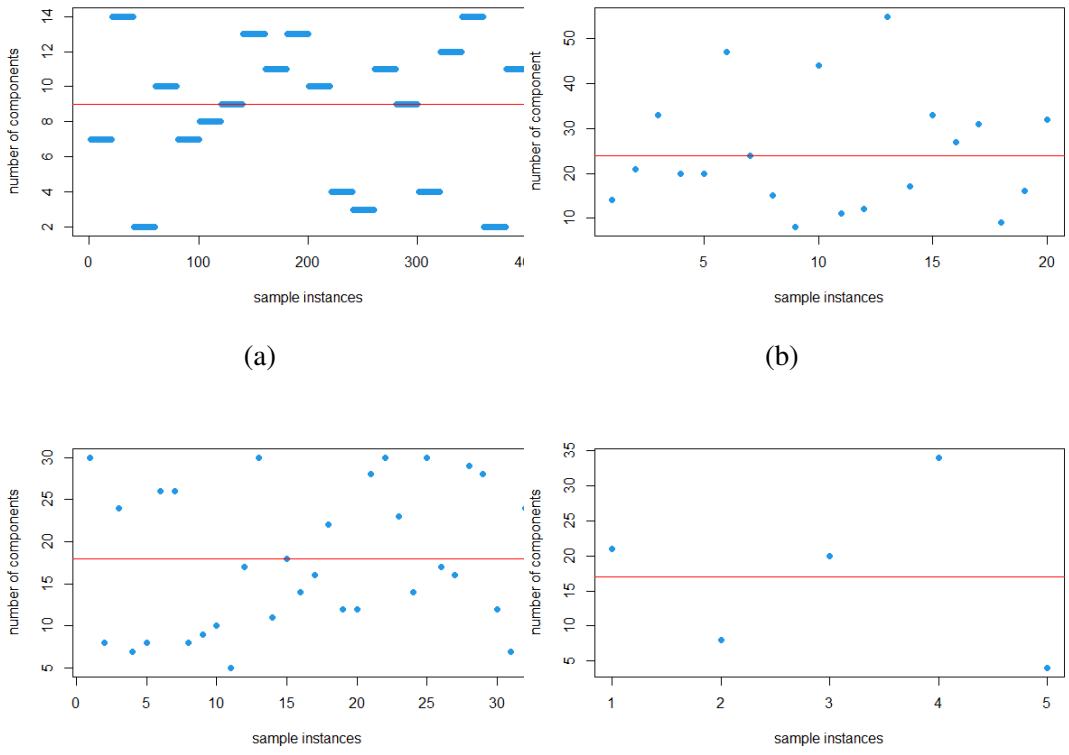
A.3 Optimal Number of Gaussian Mixture Components

The RBF-GFR model depends on different complexity parameters, such as the number of Gaussian mixture components, see Eq. (3.4). To optimize the number of Gaussian mixture components K, I found the number of components that minimize the BIC score for each block (scenario), then used the average of the number of components of all blocks as the optimal number of Gaussian mixture components for the RBF-GFR model and its extensions. The number of components is not allowed to be close or more than the data points. This is because each component will contain one point and cause singularity issues. Thus, the best number of components is set to less than half of the number of data points in a block. The optimal number of components using this method is 9, 24 and 17 for the first simulated, second simulated and wolf datasets as seen in Fig A.1. For the sparrow population dataset, the best number of Gaussian components for the RBF-GFR model was 39. However, each colony consists of 40 data points, and since the number of components is not allowed to be close or more than the data points, the best number of components was set to 18 components as it is less than half of the number of data points in each colony.

A.4 Model Diagnostics

I have used the scatter plot of residuals against the fitted values (Fig. A.2 - Fig. A.11) to check the first assumption, which is that the residuals are independent and identically distributed, and I have used the Quantile-Quantile plots (Fig. A.12 - Fig. A.21) to check if residuals are normally distributed. In the residual plots, the residuals spread around a horizontal line with distinct patterns. the bands in these plots correspond to different counts of the response variable. Furthermore, the Quantile-Quantile plots are not ideal; the deviations are in the tails in most cases. However, the focus of this thesis is on the out-of-sample predictive performance and the out-of-sample results are sufficiently encouraging even with this mismatch.

In 1976, the British statistician George Box wrote the famous line that “all models are wrong, but some are useful.” Usefulness, in the context of the research carried out for my thesis, is quantified by out-of-sample predictive performance, as the focus of my



(a) (b)

(c) (d)

Figure A.1: The best number of Gaussian mixture components that minimizes the BIC score for each block (blue points). The red line refers to the average of the number of components of all blocks; the optimal number of Gaussian mixture components for the RBF-GFR model and its extensions using (a) the first simulated dataset, $K = 9$ (b) the second simulated dataset, $K = 24$ (c) the sparrow population dataset, $K = 18$ (d) the Wolf dataset, $K = 17$.

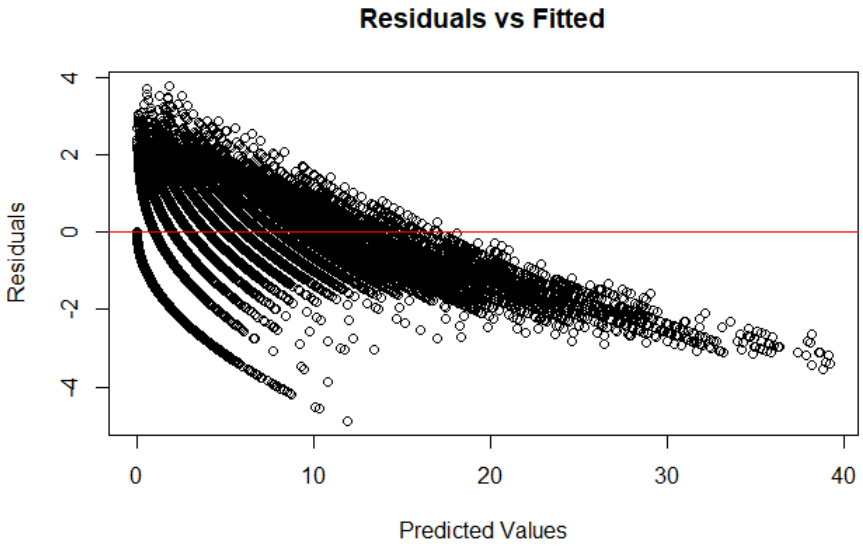


Figure A.2: Predicted values vs residuals of the GFR model in the first simulated dataset

research has been more on predictive rather than explanatory modelling. For that reason, I conclude that while the residual diagnostics clearly point to room for improvement in the modelling, they are no reason for undue concern as long as the out-of-sample predictive performance indicates a clear improvement over state-of-the-art models, which I have been able to successfully demonstrate.

A.5 Out-of-sample R^2 for the Sparrow Dataset

Table A.3 shows the out-of-sample R^2 scores for the sparrow dataset for all models using one basis function or order and 3 basis functions or orders

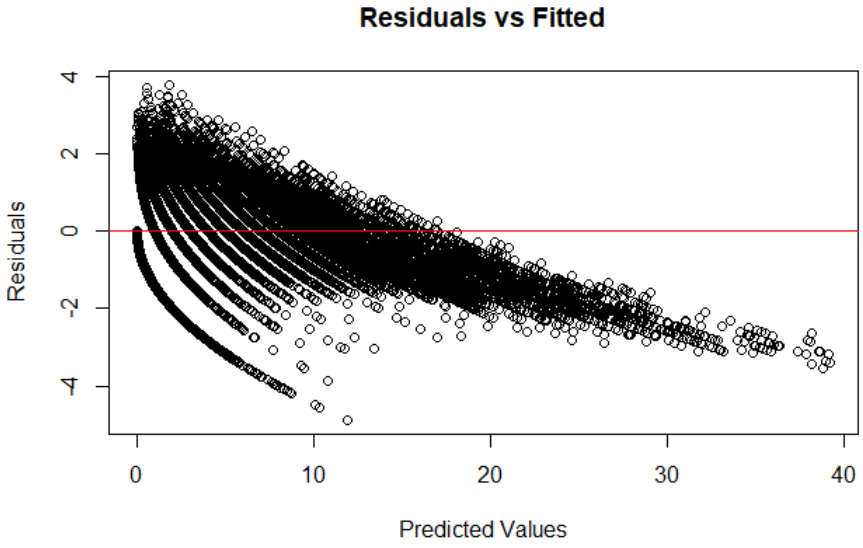


Figure A.3: Predicted values vs residuals of the RBF-GFR model in the first simulated dataset

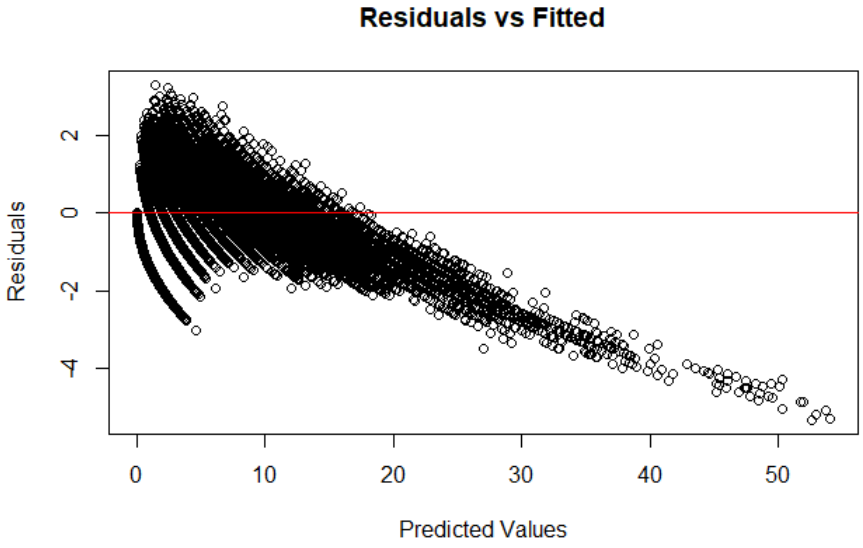


Figure A.4: Predicted values vs residuals of the regularized GFR model in the first simulated dataset

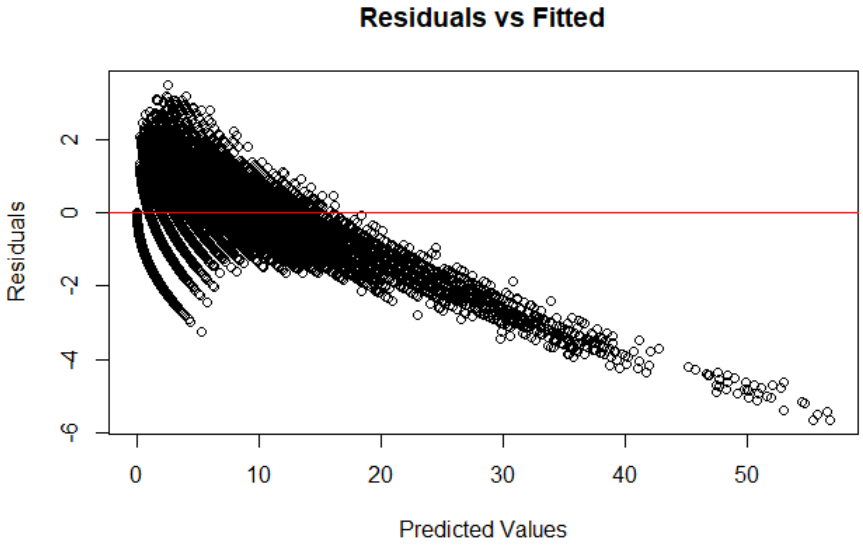


Figure A.5: Predicted values vs residuals of the regularized RBF-GFR model in the first simulated dataset

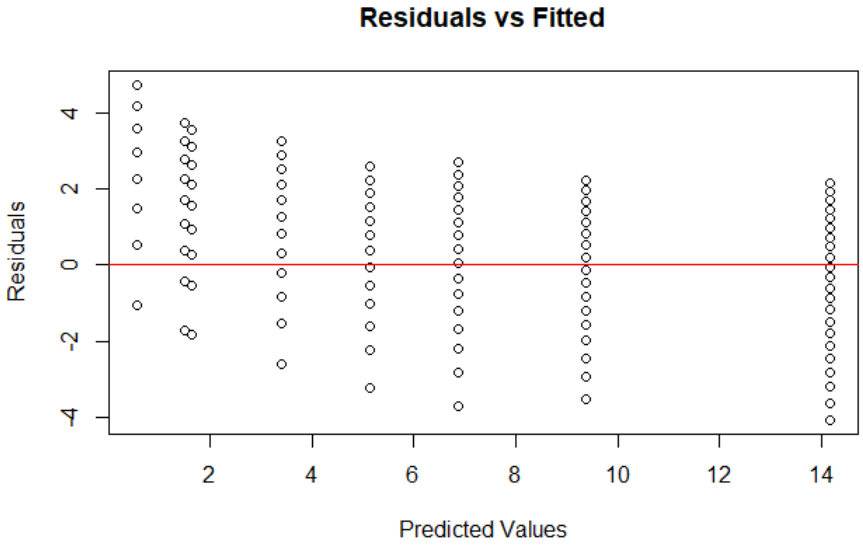


Figure A.6: Predicted values vs residuals of the GFR-CART model in the first simulated dataset

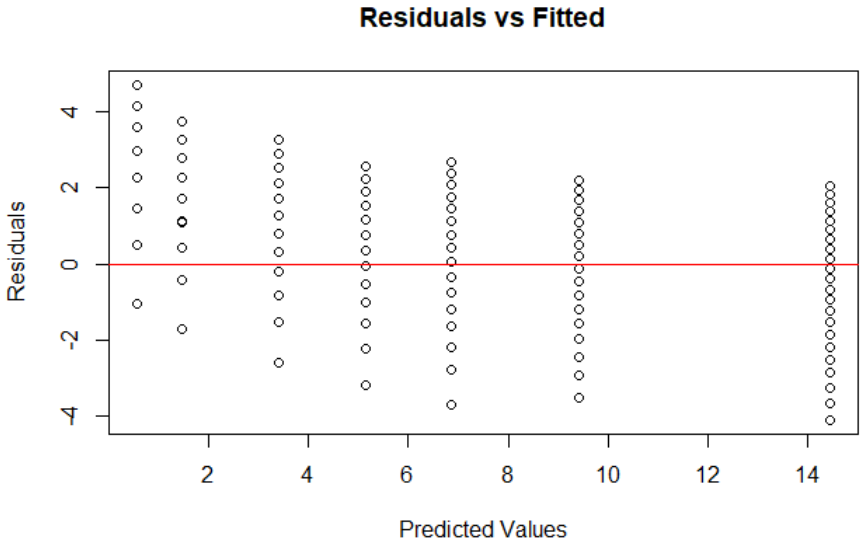


Figure A.7: Predicted values vs residuals of the RBF-GFR-CART model in the first simulated dataset

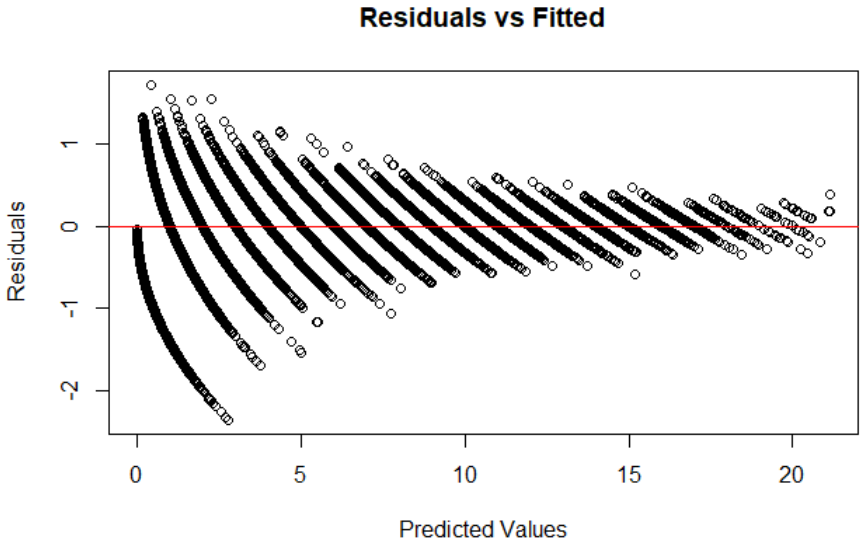


Figure A.8: Predicted values vs residuals of the GFR-RF model in the first simulated dataset

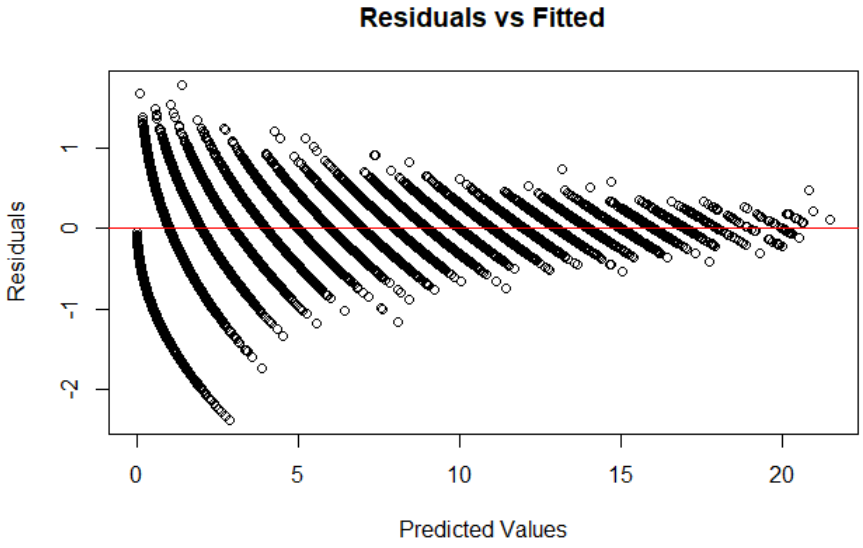


Figure A.9: Predicted values vs residuals of the RBF-GFR-RF model in the first simulated dataset

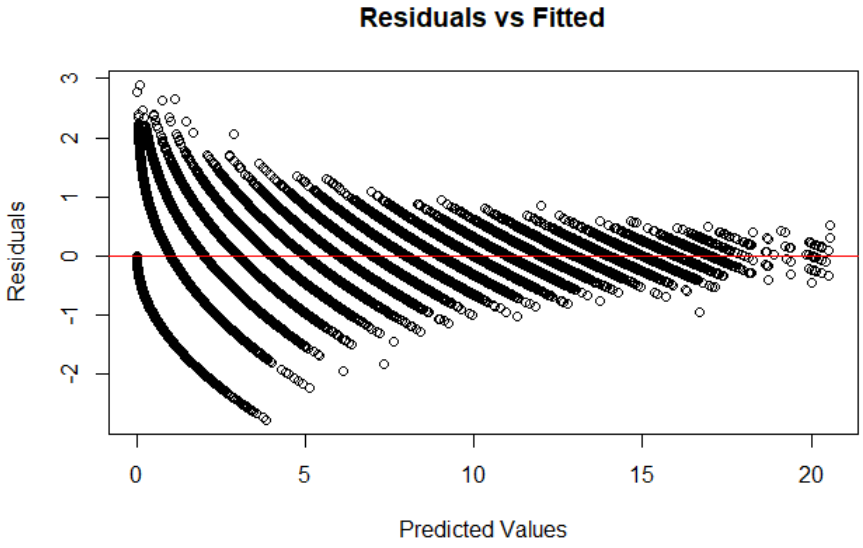


Figure A.10: Predicted values vs residuals of the GFR-XGboost model in the first simulated dataset

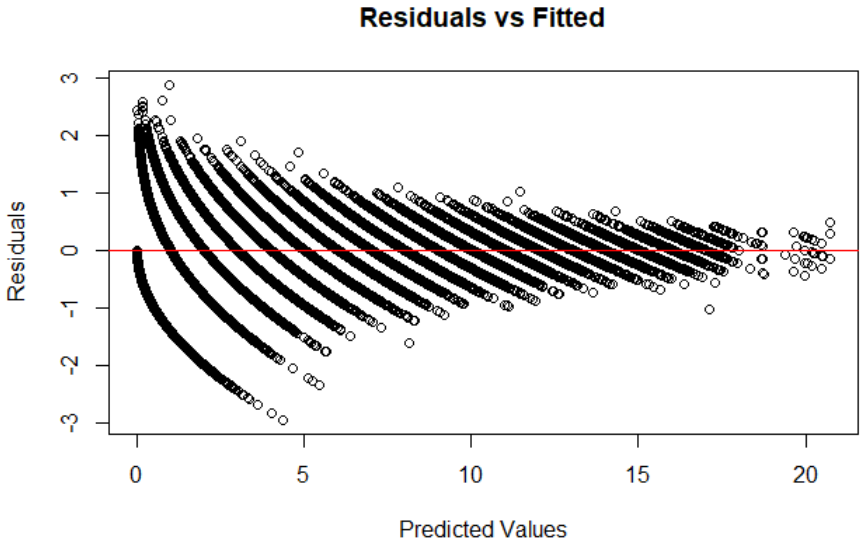


Figure A.11: Predicted values vs residuals of the RBF-GFR-XGboost model in the first simulated dataset

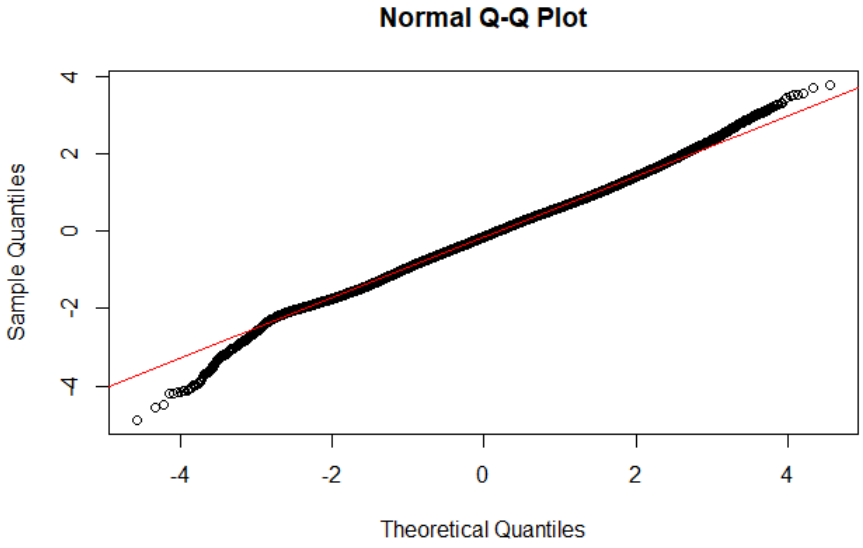


Figure A.12: Quantile-Quantile plot of the GFR model's residuals in the first simulated dataset

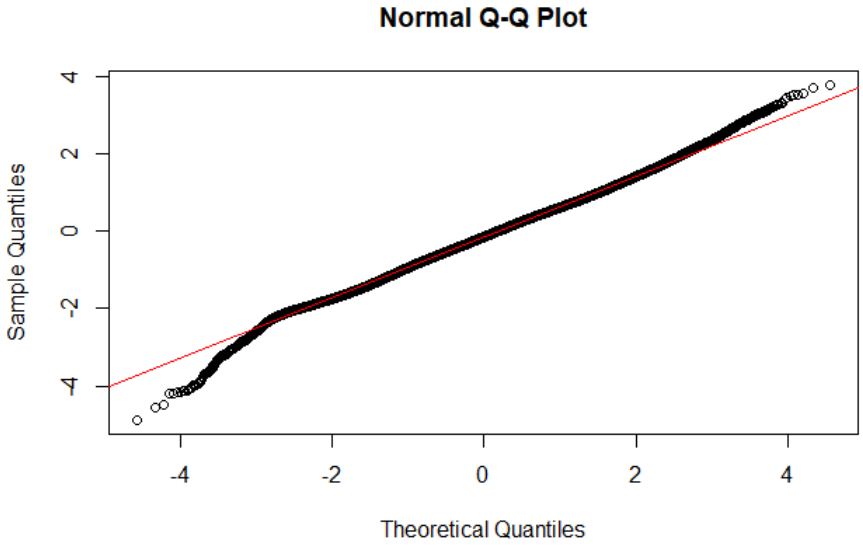


Figure A.13: Quantile-Quantile plot of the RBF-GFR model's residuals in the first simulated dataset

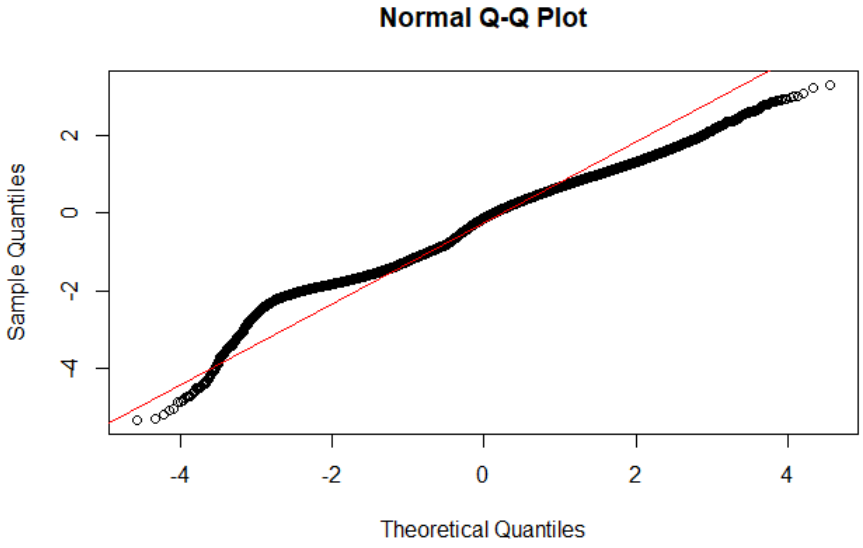


Figure A.14: Quantile-Quantile plot of the regularized GFR model's residuals in the first simulated dataset

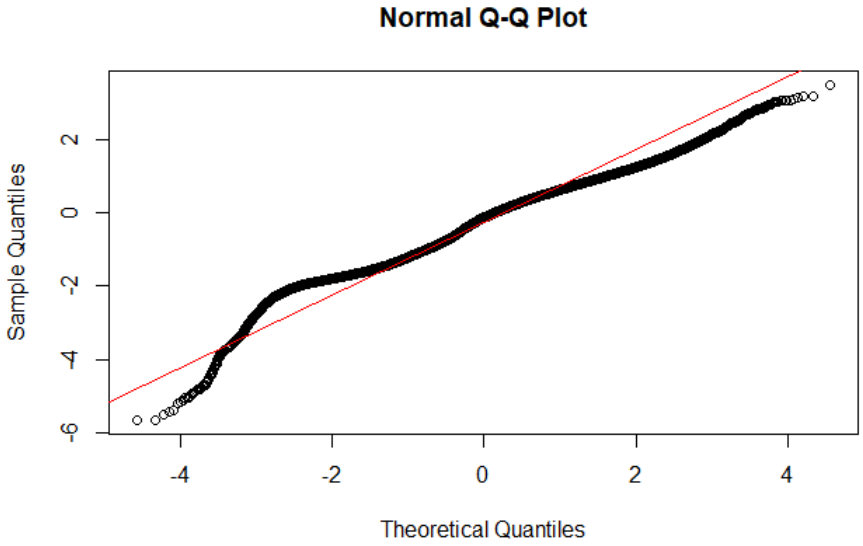


Figure A.15: Quantile-Quantile plot of the regularized RBF-GFR model's residuals in the first simulated dataset

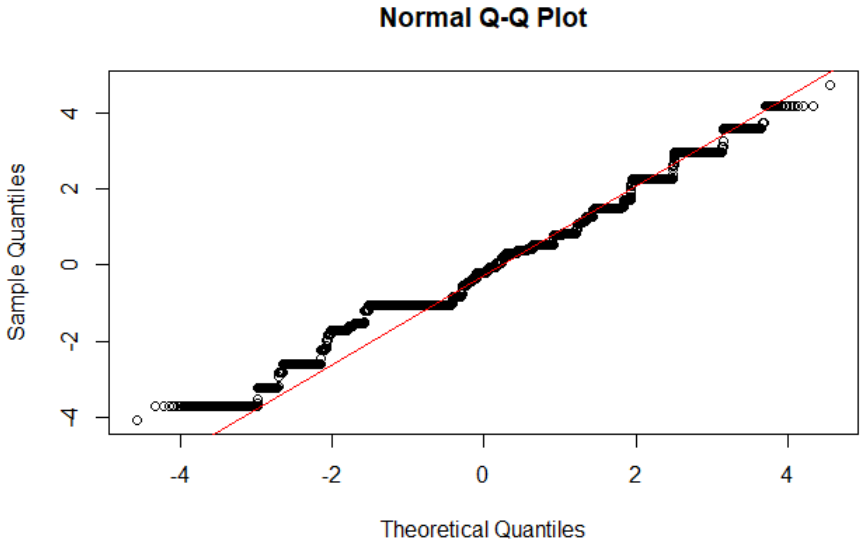


Figure A.16: Quantile-Quantile plot of the GFR-CART model's residuals in the first simulated dataset

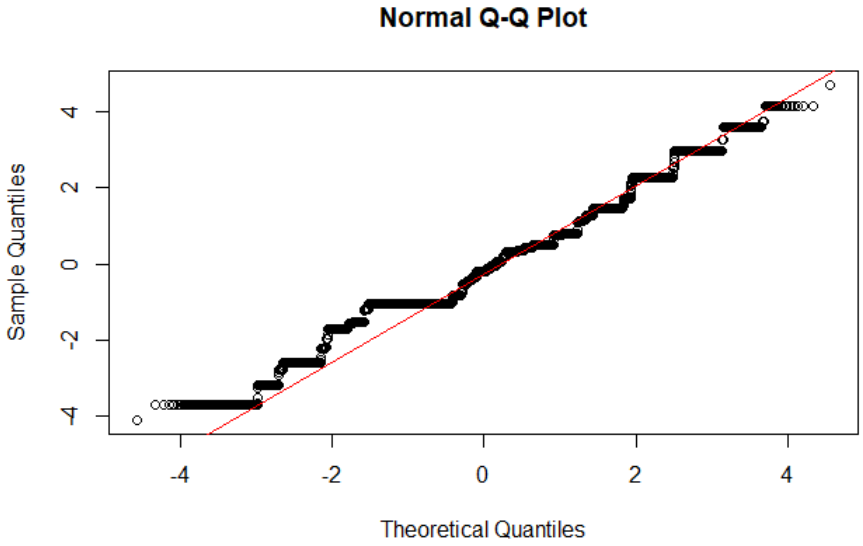


Figure A.17: Quantile-Quantile plot of the RBF-GFR-CART model's residuals in the first simulated dataset

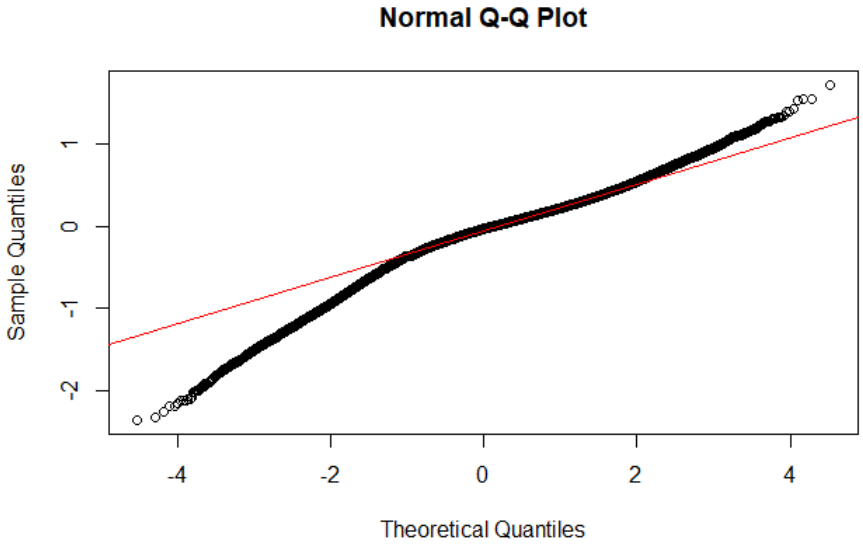


Figure A.18: Quantile-Quantile plot of the GFR-RF model's residuals in the first simulated dataset

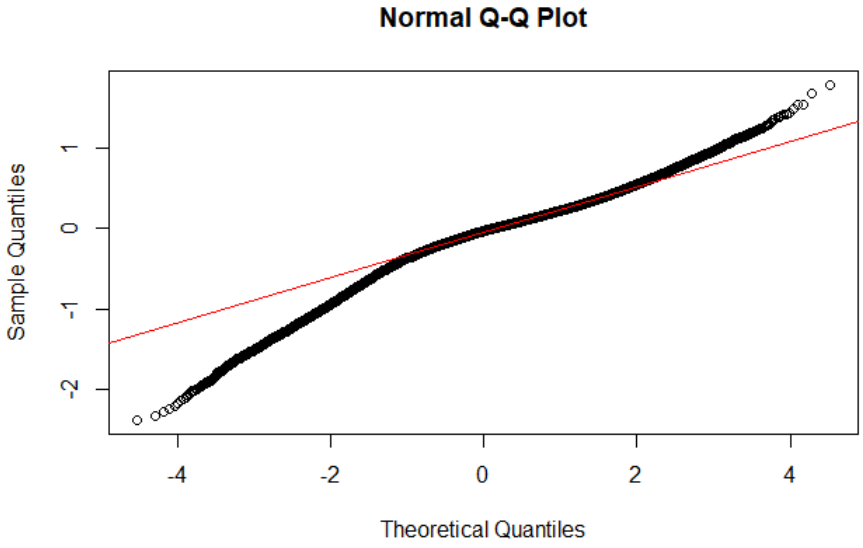


Figure A.19: Quantile-Quantile plot of the RBF-GFR-RF model's residuals in the first simulated dataset

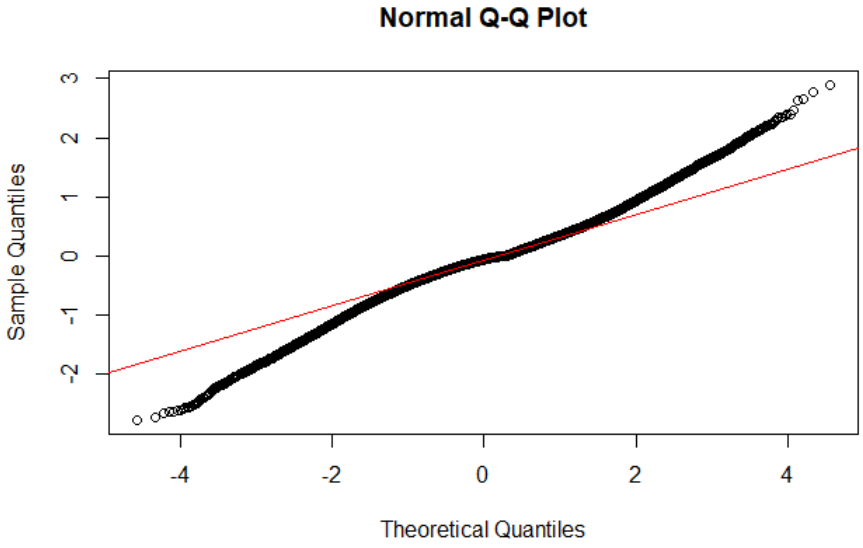


Figure A.20: Quantile-Quantile plot of the GFR-XGboost model's residuals in the first simulated dataset

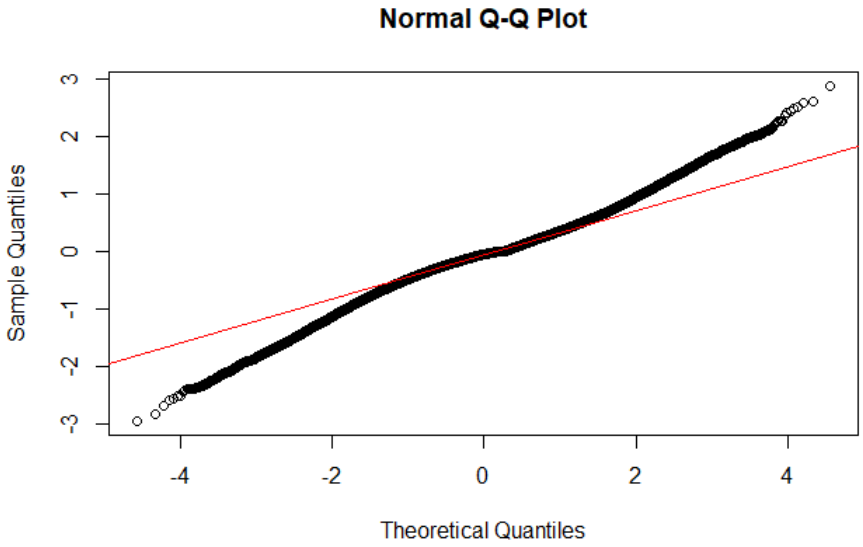


Figure A.21: Quantile-Quantile plot of the RBF-GFR-XGboost model's residuals in the first simulated dataset

Table A.3: Out-of-sample R^2 for sparrow data using one basis function or order and 3 basis functions or orders.

Models	One (basis or order)	Three (basis or orders)
GFR	0.338	0.250
GFR-CART	0.619	0.545
GFR-RF	0.730	0.765
GFR-XGBoost	0.834	0.599
RBF-GFR	0.306	0.351
RBF-CART	0.884	0.689
RBF-RF	0.861	0.920
RBF-XGBoost	0.861	0.935

A.6 Model Selection Scores for the GFR and RBF-GFR Models

The primary goal of the thesis was to increase the transferability of SDMs and compare the models' performance by the out-of-sample performance score. In addition, I reached

the model selection scores of the GFR and RBF-GFR models. The AIC and BIC scores for the RBF-GFR model are lower than the AIC and BIC scores for the GFR model for most of the datasets except for the second simulated data, where the scores are equal, as seen in Table A.4. The RBF-GFR model often outperforms the GFR model in terms of the out-of-sample R^2 and model selection scores.

Table A.4: The AIC and BIC scores of the GFR and RBF-GFR model were applied to the first simulated, second simulated, sparrow and wolf datasets.

Dataset	AIC (GFR)	AIC (RBF-GFR)	BIC (GFR)	BIC (RBF-GFR)
First simulated	907217.4	907205.8	907860.4	907818.2
Second simulated	595015.3	595015.3	595544.5	595544.5
Sparrow	1698.76	1698.687	1781.234	1781.161
Wolf	15625.83	15546.34	15961.25	15881.76

A.7 R^2_{DEV} for Count Dataset

R^2_{DEV} in Eq. (2.61) is generally better behave measurement based on deviance residuals than R^2 in Eq. (2.60) for count data regression models as described in Section 2.9. Since the two simulated datasets that I used are species abundance datasets, I used R^2_{DEV} to calculate the out-of-sample predictive performance in these datasets as shown in the rank table in Fig. A.22. However, the overall ranks using R^2_{DEV} are not different from the overall ranks using R^2 in Eq. (2.60) by comparing the average rank in Fig. 5.14 with Fig. A.22.

A.8 The Code and implementation

The following is a hyperlink to my GitHub repository that includes the code and implementation: https://github.com/shaykhah/rcodes/edit/main/README.md

	Sim1	Sim2	Sparrow	Wolf	Average
RBF-GFR-RF	0.891	0.646	0.86	0.76	0.789
GFR-RF	0.89	0.46	0.73	0.769	0.712
RBF-GFR-XGBoost	0.903	0.622	0.861	0.354	0.685
GFR-XGBoost	0.906	0.549	0.834	0.405	0.673
RBF-GFR-CART	0.732	0.494	0.884	0.182	0.573
GFR-CART	0.732	0.337	0.619	0.222	0.478
Reg RBF-GFR	0.787	0.663	0.252	0.199	0.475
Reg GFR	0.776	0.469	0.241	0.234	0.43
GLM	0.761	0.358	0.265	0.215	0.4
GFR	0.825	-0.453	0.338	0.156	0.216
RBF-GFR	0.825	-12.4	0.356	0.219	-2.75

Figure A.22: Rank table of the out-of-sample R^2_{DEV} scores of the models using the two simulated, sparrow, wolf datasets and the average score of out-of-sample R^2. Light colours indicate low ranks; the rank of the models increases as the colour shading gets darker.

A.9 Visualising Model Predictions for Some Samples

Images of predictions maps for the ground truth and the various models shown in Table 3.1 of some samples in the second simulated dataset such as samples # 2, 3, 5, 6, 10, 12, 15 and 20 in Figs A.23, A.24, A.26, A.27, A.28, A.29 and A.30 respectively. The predictions from the new extended models are much better than the predictions from the GFR model in the literature and the standard model in Eq. (2.1) based on comparing the maps to the truth. Bagging and boosting regression trees are the best performing models or close to the best models in each case. Overall, the GLM model provides a much flatter version of the truth, but if the wrong extension of the GFR model is chosen, the opposite risk of over predicting the extremes occurs. The GFR extension models are very different in terms of the out-of-sample R^2 scores from each other. The GFR model in sample instance # 5, from Fig. A.25 is homogeneous after scaling in the left panel; however, without scaling, as in the right panel, the GFR model seems to be extreme, without guarantee that, under extrapolation, the baseline will remain correct. The baseline for the average values, across the map is predicted by the average values of the covariances; thus, the coefficients obtained for the way the system responds to the average values of the coefficients do not work under extrapolation using the GFR model. In addition, Fig. A.25 shows one of the rare occasions where the GLM model actually over predicts some areas. In Fig. A.27, the RBF-GFR model maps in the scaling and non-scaling scenarios are white maps without numbers. The reason for this is seen in the left panel: it is not a problem of truncation but rather of singularity: the Gaussian kernel function has effectively converged to the degenerate case of extreme scenarios. This singularity problem was addressed by applying a regularization approach, as shown in the regularized RBF-GFR map, which also offers the best model in terms of out-of-sample prediction performance in this scenario. Sample instance # 20 is the only sample from the selected samples where the positions of the hotspots are not correctly predicted by most methods as shown in Fig. A.30.

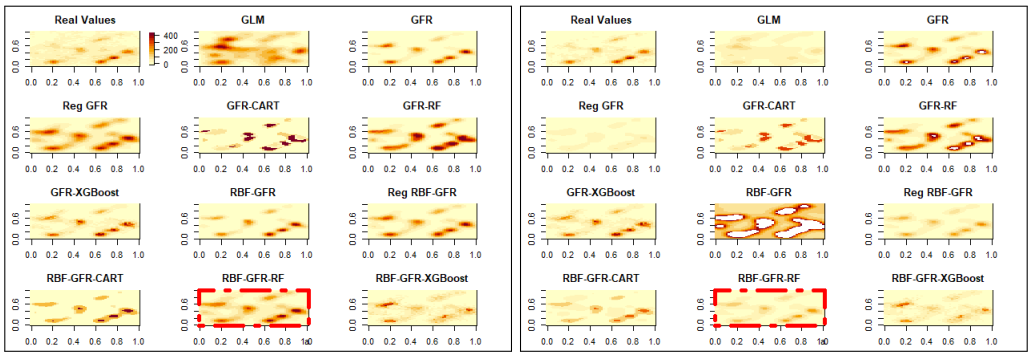


Figure A.23: A heat map of abundance and geographical predictions of the abundance of sample instance # 2 from the second simulated dataset in terms of geographical dimensions: latitude and longitude for the ground truth and the various models, as shown in Table 3.1. The two panels differ in colour range. In the left panel, the same output range is used for all models, while in the right panel, the colour range encompasses the whole range of model outputs and may be different for different models, as the minimum and maximum values for which colours should be plotted are limited by the minimum and maximum numbers of the true values. Model outputs that larger than the maximum value of the truth are thus treated as missing values and are shown in white. The map with red borders is the best predictive model based on out-of-sample R^2.

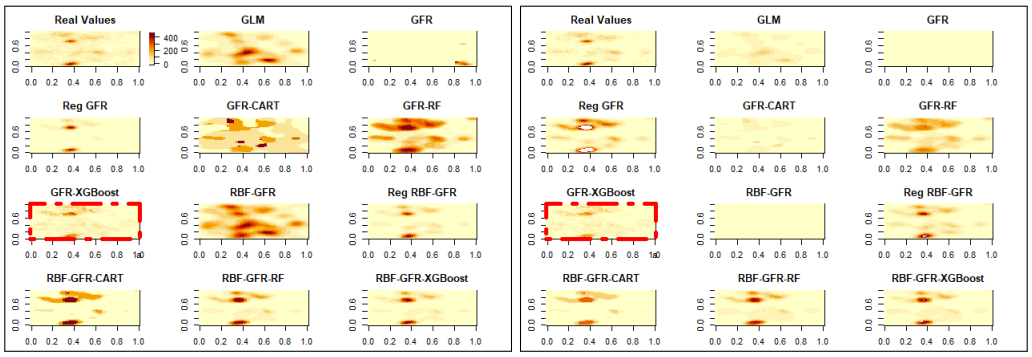


Figure A.24: A heat map of abundance and geographical predictions of the abundance of sample instance # 3 from the second simulated dataset in terms of geographical dimensions: latitude and longitude for the ground truth and the various models, as shown in Table 3.1. The two panels differ in colour range. In the left panel, the same output range is used for all models, while in the right panel, the colour range encompasses the whole range of model outputs and may be different for different models, as the minimum and maximum values for which colours should be plotted are limited by the minimum and maximum numbers of the true values. Model outputs that larger than the maximum value of the truth are thus treated as missing values and are shown in white. The map with red borders is the best predictive model based on out-of-sample R^2.

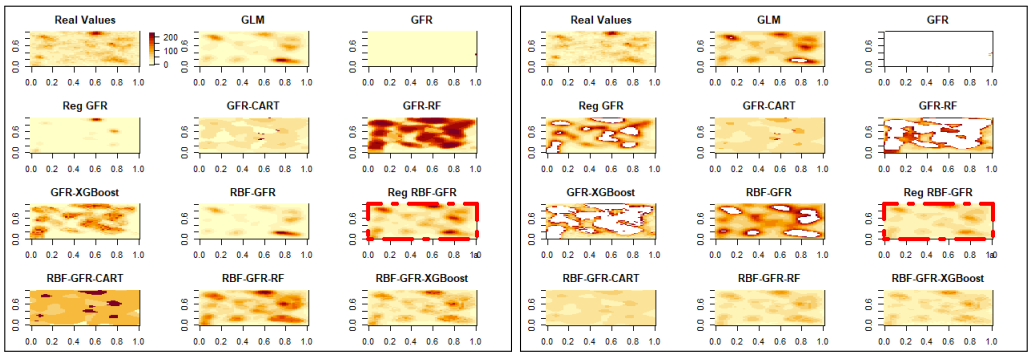


Figure A.25: A heat map of abundance and geographical predictions of the abundance of sample instance # 5 from the second simulated dataset in terms of geographical dimensions: latitude and longitude for the ground truth and the various models, as shown in Table 3.1. The two panels differ in colour range. In the left panel, the same output range is used for all models, while in the right panel, the colour range encompasses the whole range of model outputs and may be different for different models, as the minimum and maximum values for which colours should be plotted are limited by the minimum and maximum numbers of the true values. Model outputs that larger than the maximum value of the truth are thus treated as missing values and are shown in white. The map with red borders is the best predictive model based on out-of-sample R^2.

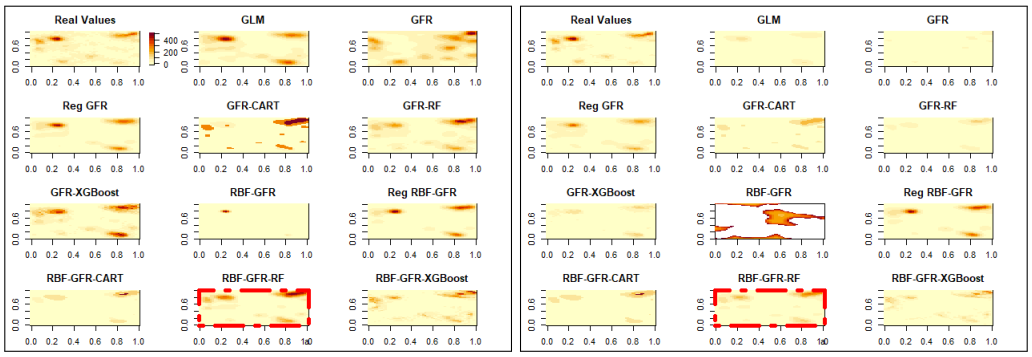


Figure A.26: A heat map of abundance and geographical predictions of the abundance of sample instance # 6 from the second simulated dataset in terms of geographical dimensions: latitude and longitude for the ground truth and the various models, as shown in Table 3.1. The two panels differ in colour range. In the left panel, the same output range is used for all models, while in the right panel, the colour range encompasses the whole range of model outputs and may be different for different models, as the minimum and maximum values for which colours should be plotted are limited by the minimum and maximum numbers of the true values. Model outputs that larger than the maximum value of the truth are thus treated as missing values and are shown in white. The map with red borders is the best predictive model based on out-of-sample R^2.

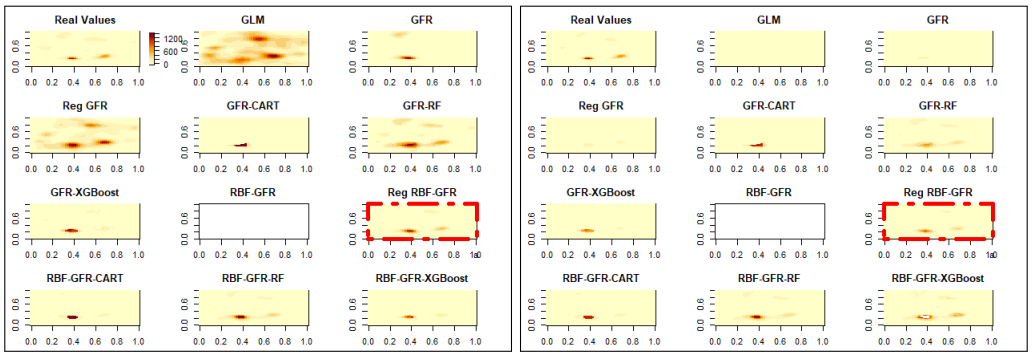


Figure A.27: A heat map of abundance and geographical predictions of the abundance of sample instance # 10 from the second simulated dataset in terms of geographical dimensions: latitude and longitude for the ground truth and the various models, as shown in Table 3.1. The two panels differ in colour range. In the left panel, the same output range is used for all models, while in the right panel, the colour range encompasses the whole range of model outputs and may be different for different models, as the minimum and maximum values for which colours should be plotted are limited by the minimum and maximum numbers of the true values. Model outputs that larger than the maximum value of the truth are thus treated as missing values and are shown in white. The map with red borders is the best predictive model based on out-of-sample R^2.

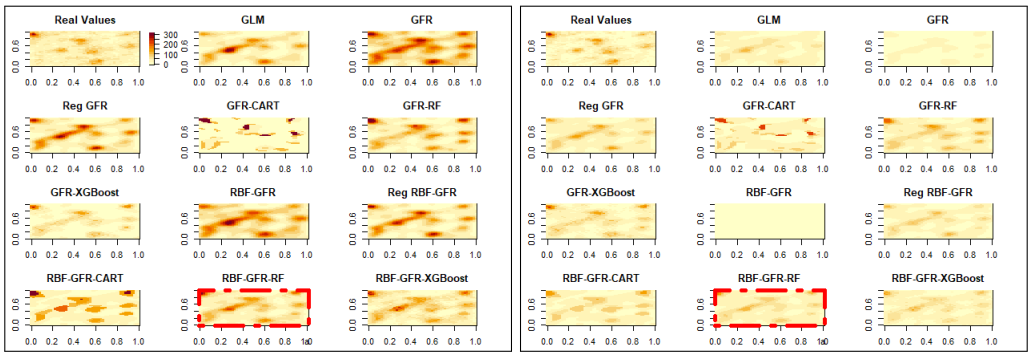


Figure A.28: A heat map of abundance and geographical predictions of the abundance of sample instance # 12 from the second simulated dataset in terms of geographical dimensions: latitude and longitude for the ground truth and the various models, as shown in Table 3.1. The two panels differ in colour range. In the left panel, the same output range is used for all models, while in the right panel, the colour range encompasses the whole range of model outputs and may be different for different models, as the minimum and maximum values for which colours should be plotted are limited by the minimum and maximum numbers of the true values. Model outputs that larger than the maximum value of the truth are thus treated as missing values and are shown in white. The map with red borders is the best predictive model based on out-of-sample R^2.

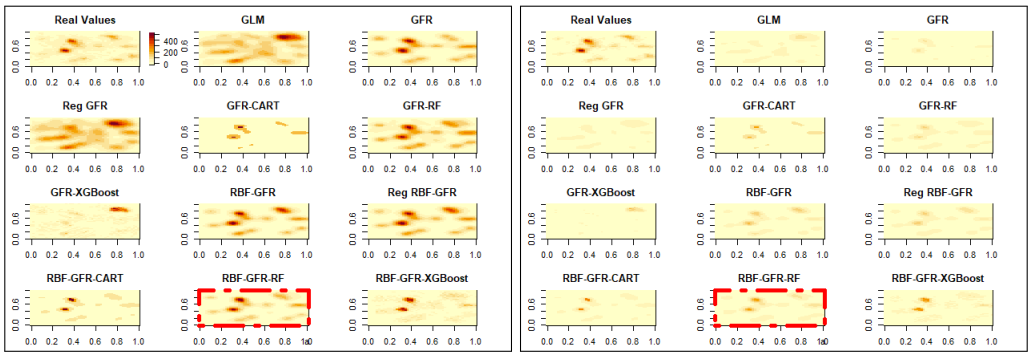


Figure A.29: A heat map of abundance and geographical predictions of the abundance of sample instance # 15 from the second simulated dataset in terms of geographical dimensions: latitude and longitude for the ground truth and the various models, as shown in Table 3.1. The two panels differ in colour range. In the left panel, the same output range is used for all models, while in the right panel, the colour range encompasses the whole range of model outputs and may be different for different models, as the minimum and maximum values for which colours should be plotted are limited by the minimum and maximum numbers of the true values. Model outputs that larger than the maximum value of the truth are thus treated as missing values and are shown in white. The map with red borders is the best predictive model based on out-of-sample R^2.

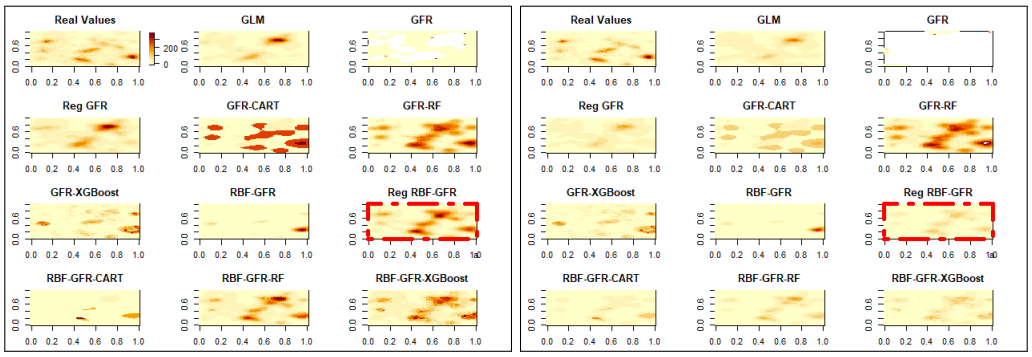


Figure A.30: A heat map of abundance and geographical predictions of the abundance of sample instance # 20 from the second simulated dataset in terms of geographical dimensions: latitude and longitude for the ground truth and the various models, as shown in Table 3.1. The two panels differ in colour range. In the left panel, the same output range is used for all models, while in the right panel, the colour range encompasses the whole range of model outputs and may be different for different models, as the minimum and maximum values for which colours should be plotted are limited by the minimum and maximum numbers of the true values. Model outputs that larger than the maximum value of the truth are thus treated as missing values and are shown in white. The map with red borders is the best predictive model based on out-of-sample R^2.

A.10 Normality Assumption Check of Out-of-sample R^2 Scores

Fig. A.31 shows the histograms used for the normality assumption check of the out-of-sample R^2 scores distribution from the standard (GLM) model for all species in the BBS dataset before including other species' abundance as additional covariates in the left and after including other species' abundance in the right panel.

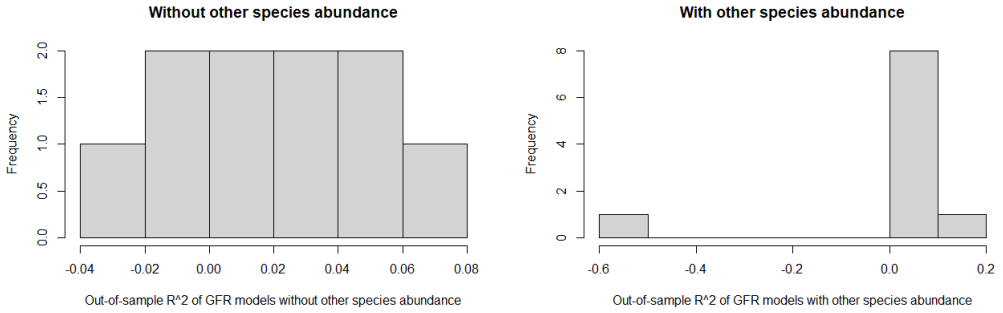


Figure A.31: Histogram of normality assumption check of out-of-sample R^2 scores distribution from the standard (GLM) model for all species in the BBS dataset before including other species' abundance as additional covariates in the left and after including other species' abundance in the right panel.

A.11 Discrepancy Between the Land Cover Covariates in Training and Testing Satasets of the BBS Dataset.

The scatter plots in Fig. A.32 show the agreement of the entropy scores in the training and testing sets in the BBS dataset, where most of the points are scattered around a line of equal performance. Figs. A.33 and A.34 illustrate how the training set is different from the testing set of the BBS dataset, as described in Section 7.5.2.2.

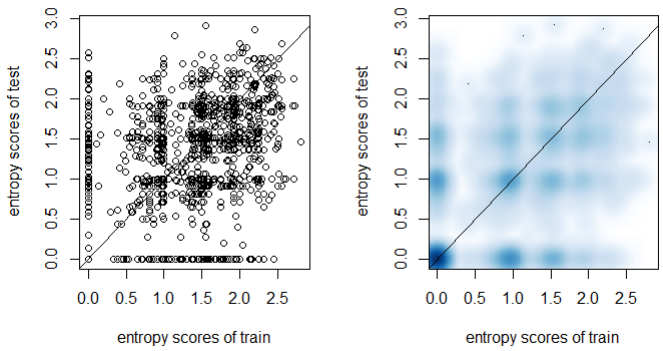


Figure A.32: Scatter plot of entropy scores in the training set in the x-axis vs entropy scores of the test set in the y-axis where the majority of the points are scattered around the line of equal performance.

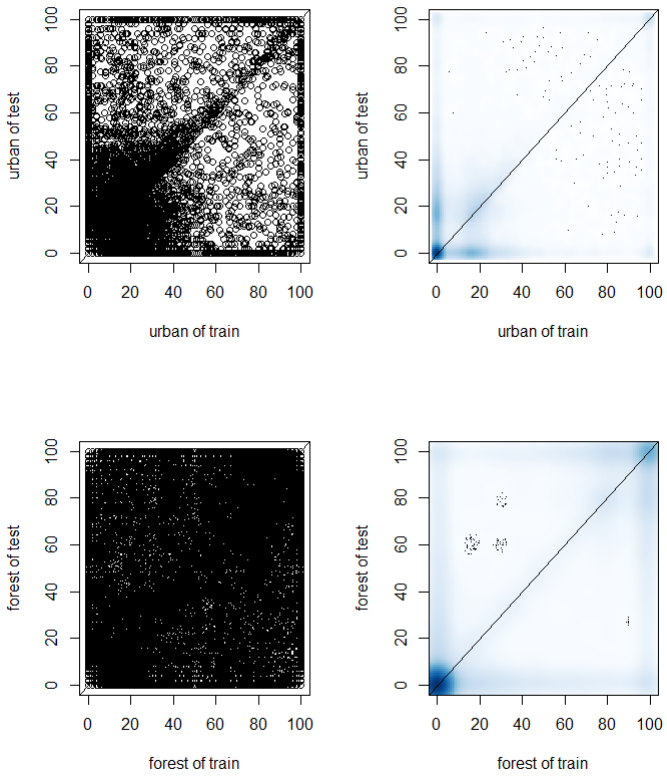


Figure A.33: Scatter plots of urban and forest covariates in the training set in the x-axis vs urban and forest covariates of the test set in the y-axis and the line is the equal performance line where there is a discrepancy between these land cover covariates in training and testing datasets.

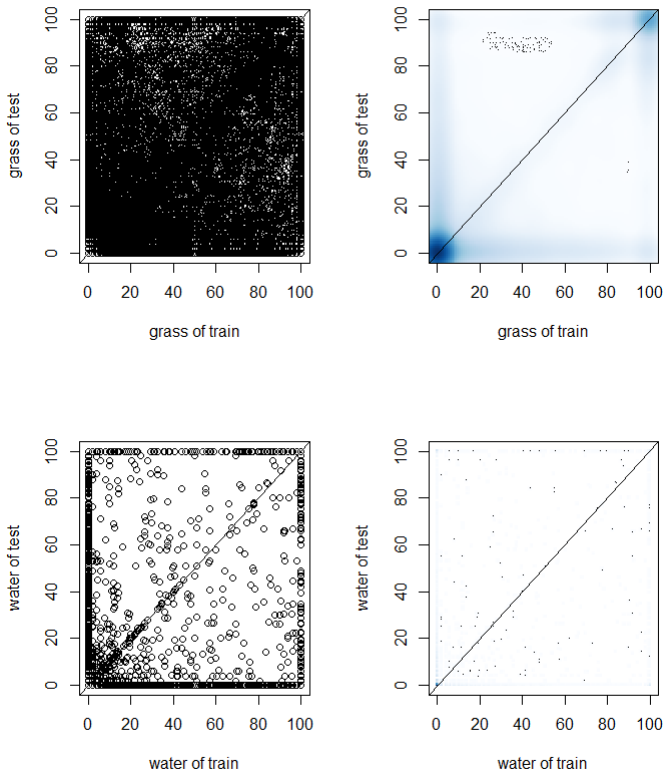


Figure A.34: Scatter plots of grass and water covariates in the training set in the x-axis vs grass and water covariates of the test set in the y-axis and the line is the equal performance line where there is a discrepancy between these land cover covariates in training and testing datasets.

Bibliography

G. Aarts, M. MacKenzie, B. McConnell, M. Fedak, and J. Matthiopoulos. Estimating space-use and habitat preference from wildlife telemetry data. *Ecography*, 31(1):140–160, 2008.

G. Aarts, J. Fieberg, and J. Matthiopoulos. Comparative interpretation of count, presence–absence and point methods for species distribution models. *Methods in Ecology and Evolution*, 3(1):177–187, 2012.

H. Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974. doi: 10.1109/TAC.1974.1100705.

S. Aldossari, J. Matthiopoulos, and D. Husmeier. Statistical modelling of habitat selection. In *The 35th International Workshop on Statistical Modelling*. Servicio Editorial de la Universidad del País Vasco, 2020.

S. Aldossari, D. Husmeier, and J. Matthiopoulos. Generalized functional responses in habitat selection fitted by decision trees and random forests. In *The 3rd International Conference on Statistics: Theory and Applications (ICSTA'21)*. Avestia Publishing, 2021. doi: 10.11159/icsta21.125.

S. Aldossari, D. Husmeier, and J. Matthiopoulos. Transferable species distribution modelling: Comparative performance of generalised functional response models. *Ecological Informatics*, 71:101803, 2022.

M. Austin. Spatial prediction of species distribution: an interface between ecological

theory and statistical modelling. *Ecological modelling*, 157(2-3):101–118, 2002. doi: https://doi.org/10.1016/S0304-3800(02)00205-3.

M. Austin. Species distribution models and ecological theory: a critical assessment and some possible new approaches. *Ecological modelling*, 200(1-2):1–19, 2007.

V. Bahn and B. J. McGill. Testing the predictive performance of distribution models. *Oikos*, 122(3):321–331, 2013. doi: https://doi.org/10.1111/j.1600-0706.2012.00299.x.

M. Barbet-Massin, Q. Rome, C. Villemant, and F. Courchamp. Can species distribution models really predict the expansion of invasive species? *PloS one*, 13(3):e0193085, 2018. doi: https://doi.org/10.1371/journal.pone.0193085.

A. M. Barbosa, R. Real, and J. M. Vargas. Transferability of environmental favourability models in geographic space: the case of the iberian desman (galemys pyrenaicus) in portugal and spain. *Ecological modelling*, 220(5):747–754, 2009. doi: https://doi.org/10.1016/j.ecolmodel.2008.12.004.

R. J. Barker and J. R. Sauer. Statistical aspects of point count sampling. *In: Ralph, C. John; Sauer, John R.; Droege, Sam, technical editors. 1995. Monitoring bird populations by point counts. Gen. Tech. Rep. PSW-GTR-149. Albany, CA: US Department of Agriculture, Forest Service, Pacific Southwest Research Station: p. 125-130*, 149, 1995.

A. Barnwal, H. Cho, and T. Hocking. Survival regression with accelerated failure time model in xgboost. *Journal of Computational and Graphical Statistics*, pages 1–11, 2022.

P. Batz, A. Ruttor, and M. Opper. Approximate bayes learning of stochastic differential equations. *Physical Review E*, 98(2):022109, 2018.

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*, 365 (1550):2245–2254, 2010. doi: https://doi.org/10.1098/rstb.2010.0083.

G. Biau and E. Scornet. A random forest guided tour. *Test*, 25(2):197–227, 2016.

C. M. Bishop. *Neural Networks for Pattern Recognition*. Clarendon Press, 1995.

C. M. Bishop. *Pattern recognition and machine learning*. Springer, 2006.

W. M. Block and L. A. Brennan. The habitat concept in ornithology. In *Current ornithology*, pages 35–91. Springer, 1993.

A. A. Boiarov and O. N. Granichin. Stochastic approximation algorithm with randomization at the input for unsupervised parameters estimation of gaussian mixture model with sparse parameters. *Automation and Remote Control*, 80(8):1403–1418, 2019.

M. S. Boyce and L. L. McDonald. Relating populations to habitats using resource selection functions. *Trends in ecology & evolution*, 14(7):268–272, 1999. doi: https://doi.org/10.1016/S0169-5347(99)01593-1.

M. S. Boyce, P. R. Vernier, S. E. Nielsen, and F. K. Schmiegelow. Evaluating resource selection functions. *Ecological modelling*, 157(2-3):281–300, 2002.

L. Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001. doi: https://doi.org/10.1023/A:1010933404324.

L. Breiman, J. Friedman, and R. Olshen. *Clasification and Regression Trees*. Wadsworth, 1984.

L. Brillouin. *Science and information theory*. Courier Corporation, 2013.

R. Bryll, R. Gutierrez-Osuna, and F. Quek. Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets. *Pattern recognition*, 36(6):1291–1302, 2003.

A. C. Cameron and F. A. Windmeijer. R-squared measures for count data regression models with applications to health-care utilization. *Journal of Business & Economic Statistics*, 14(2):209–220, 1996.

G. Casella and R. L. Berger. *Statistical inference*. Cengage Learning, 2021.

V. Cazalis, S. Belghali, and A. S. Rodrigues. Using a large-scale biodiversity monitoring dataset to test the effectiveness of protected areas at conserving north-american breeding birds. *BioRxiv*, page 433037, 2019.

D. S. Chen and R. C. Jain. A robust backpropagation learning algorithm for function approximation. *IEEE Transactions on Neural Networks*, 5(3):467–479, 1994.

T. Chen and C. Guestrin. XGBoost: a scalable tree boosting system. *KDD*, pages 785–794, 2016.

T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, et al. Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4, 2015.

D. J. Currie and S. Venne. Climate change is not a major driver of shifts in the geographical distributions of north american birds. *Global Ecology and Biogeography*, 26(3):333–346, 2017.

G. N. Daskalova, I. H. Myers-Smith, A. D. Bjorkman, S. A. Blowes, S. R. Supp, A. E. Magurran, and M. Dornelas. Landscape-scale forest loss as a catalyst of population and biodiversity change. *Science*, 368(6497):1341–1347, 2020.

T. G. Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting and randomization. *Machine learning*, 32:1–22, 1998.

D. A. Donald, Y. Everingham, L. McKinna, and D. Coomans. Feature selection in the wavelet domain: adaptive wavelets. Elsevier, 2009.

C. F. Dormann. Promising the future? global change projections of species distributions. *Basic and applied ecology*, 8(5):387–397, 2007. doi: https://doi.org/10.1016/j.baae.2006.11.001.

P. Du, A. Samat, B. Waske, S. Liu, and Z. Li. Random forest and rotation forest for fully polarized sar image classification using polarimetric and spatial features. *ISPRS Journal of Photogrammetry and Remote Sensing*, 105:38–53, 2015.

J. Duque-Lazo, H. Van Gils, T. Groen, and R. Navarro-Cerrillo. Transferability of species distribution models: The case of phytophthora cinnamomi in southwest spain and southwest australia. *Ecological Modelling*, 320:62–70, 2016.

J. Ehrlén and W. F. Morris. Predicting changes in the distribution and abundance of species under environmental change. *Ecology letters*, 18(3):303–314, 2015. doi: https://doi.org/10.1111/ele.12410.

J. Elith and J. R. Leathwick. Species distribution models: ecological explanation and prediction across space and time. *Annual review of ecology, evolution, and systematics*, 40:677–697, 2009.

J. Elith*, C. H. Graham*, R. P. Anderson, M. Dudík, S. Ferrier, A. Guisan, R. J. Hijmans, F. Huettmann, J. R. Leathwick, A. Lehmann, et al. Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 29(2):129–151, 2006.

M. R. Evans, K. J. Norris, and T. G. Benton. Predictive ecology: systems approaches, 2012.

J. J. Faraway. *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. CRC press, 2016.

D. A. Fordham, H. R. Akçakaya, J. Alroy, F. Saltré, T. M. Wigley, and B. W. Brook. Predicting and mitigating future biodiversity loss using long-term ecological proxies. *Nature Climate Change*, 6(10):909–916, 2016. doi: https://www.nature.com/articles/nclimate3086.

C. Fraley, A. E. Raftery, T. B. Murphy, and L. Scrucca. mclust version 4 for r: normal mixture modeling for model-based clustering, classification, and density estimation. Technical report, Technical report, 2012.

F. Frankel and R. Reid. Big data: Distilling meaning from data. *Nature*, 455(7209):30–30, 2008.

Y. Gao and Q. Dai. *View-based 3-D object retrieval*. Morgan Kaufmann, 2014.

J. Gareth, W. Daniela, H. Trevor, and T. Robert. *An introduction to statistical learning: with applications in R*. Spinger, 2013.

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.

J. E. Gentle, W. K. Härdle, and Y. Mori. *Springer Handbooks of Computational Statistics*. Springer, 2012.

P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Machine learning*, 63 (1):3–42, 2006.

C. S. Gillies, M. Hebblewhite, S. E. Nielsen, M. A. Krawchuk, C. L. Aldridge, J. L. Frair, D. J. Saher, C. E. Stevens, and C. L. Jerde. Application of random effects to the study of resource selection by animals. *Journal of Animal Ecology*, 75(4):887–898, 2006. doi: https://doi.org/10.1111/j.1365-2656.2006.01106.x.

I. M. R. Godvik, L. E. Loe, J. O. Vik, V. Veiberg, R. Langvatn, and A. Mysterud. Temporal scales, trade-offs, and functional responses in red deer habitat selection. *Ecology*, 90 (3):699–710, 2009.

A. S. Goldberger et al. Econometric theory. *Econometric theory.*, 1964.

Y. Haddou, R. Mancy, J. Matthiopoulos, S. Spatharis, and D. M. Dominoni. Widespread extinction debts and colonization credits in united states breeding bird communities. *Nature ecology & evolution*, 6(3):324–331, 2022.

H. Han, X. Guo, and H. Yu. Variable selection using mean decrease accuracy and mean decrease gini based on random forest. In *2016 7th ieee international conference on software engineering and service science (icsess)*, pages 219–224. IEEE, 2016.

J. Han, J. Pei, and H. Tong. *Data mining: concepts and techniques*. Morgan kaufmann, 2022.

J. M. Hanowski and G. J. Niemi. A comparison of on-and off-road bird counts: Do you need to go off road to count birds accurately?(una comparación de conteos dentro-de y fuera-de caminos:¿ hay que alejarse de los caminos para contar aves con exactitud?). *Journal of Field Ornithology*, pages 469–483, 1995.

R. Harré. *Cognitive science: A philosophical introduction*. Sage, 2002.

T. Hastie and R. Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(4):757–779, 1993. doi: https://doi.org/10.1111/j.2517-6161.1993.tb01939.x.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2008.

T. Hastie, J. Qian, and K. Tay. An introduction to glmnet, 2016.

R. K. Heikkinen, M. Marmion, and M. Luoto. Does the interpolation accuracy of species distribution models come at the expense of transferability? *Ecography*, 35(3):276–288, 2012.

M. Higashi and G. J. Klir. Measures of uncertainty and information based on possibility distributions. *International journal of general systems*, 9(1):43–58, 1982.

J. D. Holbrook, L. E. Olson, N. J. DeCesare, M. Hebblewhite, J. R. Squires, and R. Steenweg. Functional responses in habitat selection: clarifying hypotheses and interpretations. *Ecological Applications*, 29(3):e01852, 2019.

C. S. Holling. Some characteristics of simple types of predation and parasitism1. *The canadian entomologist*, 91(7):385–398, 1959.

J. E. Houlahan, S. T. McKinney, T. M. Anderson, and B. J. McGill. The priority of prediction in ecological understanding. *Oikos*, 126(1):1–7, 2017. doi: https://doi.org/10.1111/oik.03726.

J. Huang, H. Fang, and X. Fan. Decision forest for classification of gene expression data. *Computers in biology and medicine*, 40(8):698–704, 2010.

D. E. Huber and C. G. Healey. Visualizing data with motion. In *VIS 05. IEEE Visualization, 2005.*, pages 527–534. IEEE, 2005.

M.-A. R. Hudson, C. M. Francis, K. J. Campbell, C. M. Downes, A. C. Smith, and K. L. Pardieck. The role of the north american breeding bird survey in conservation. *The Condor: Ornithological Applications*, 119(3):526–545, 2017.

M. Iturbide, J. Bedia, and J. M. Gutiérrez. Background sampling and transferability of species distribution model ensembles under climate change. *Global and Planetary Change*, 166:19–29, 2018.

A. Kassambara. *Machine learning essentials: Practical guide in R*. Sthda, 2018.

H. K. Kindsvater, N. K. Dulvy, C. Horswill, M.-J. Juan-Jordá, M. Mangel, and J. Matthiopoulos. Overcoming the data crisis in biodiversity conservation. *Trends in ecology & evolution*, 33(9):676–688, 2018. doi: https://doi.org/10.1016/j.tree.2018.06.004.

S. Kullback. *Information theory and statistics*. Courier Corporation, 1997.

J. J. Lawler, D. White, R. P. Neilson, and A. R. Blaustein. Predicting climate-induced range shifts: model differences and model reliability. *Global change biology*, 12(8): 1568–1584, 2006.

C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of experimental social psychology*, 49(4):764–766, 2013. doi: https://doi.org/10.1016/j.jesp.2013.03.013.

A. Liaw and M. Wiener. Package 'randomforest'. *University of California, Berkeley: Berkeley, CA, USA*, 2018. doi: 10.1023/A:1010933404324.

P. K. Lira, M. de Souza Leite, and J. P. Metzger. Temporal lag in ecological responses to landscape change: Where are we now? *Current Landscape Ecology Reports*, 4(3): 70–82, 2019.

R. J. Lopez. L'hôpital's rule. *Maple via Calculus: A Tutorial Approach*, pages 88–90, 1994.

S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

R. MacArthur. Fluctuations of animal populations and a measure of community stability. *Ecology*, 36(3):533–536, 1955. ISSN 00129658, 19399170. URL `http://www.jstor.org/stable/1929601`.

D. J. MacKay. Bayesian interpolation. *Neural computation*, 4(3):415–447, 1992.

D. I. MacKenzie, J. D. Nichols, J. A. Royle, K. H. Pollock, L. L. Bailey, and J. E. Hines. *Occupancy estimation and modeling: inferring patterns and dynamics of species occurrence*. Elsevier, 2017.

U. Manikpuri and Y. Yadav. Image enhancement through logarithmic transformation. *International Journal of*, 2014.

D. L. Marchisio and R. O. Fox. Solution of population balance equations using the direct quadrature method of moments. *Journal of Aerosol Science*, 36(1):43–73, 2005.

V. Maris, P. Huneman, A. Coreau, S. Kéfi, R. Pradel, and V. Devictor. Prediction in ecology: promises, obstacles and clarifications. *Oikos*, 127(2):171–183, 2018. doi: https://doi.org/10.1111/oik.04655.

J. Matthiopoulos. *How to be a quantitative ecologist: the'A to R'of green mathematics and statistics*. John Wiley & Sons, 2011.

J. Matthiopoulos, M. Hebblewhite, G. Aarts, and J. Fieberg. Generalized functional responses for species distributions. *Ecology*, 92(3):583–589, 2011. doi: https://doi.org/10.1890/10-0751.1.

J. Matthiopoulos, J. Fieberg, G. Aarts, H. L. Beyer, J. M. Morales, and D. T. Haydon. Establishing the link between habitat selection and animal population dynamics. *Ecological Monographs*, 85(3):413–436, 2015. doi: https://doi.org/10.1890/14-2244.1.

J. Matthiopoulos, C. Field, and R. MacLeod. Predicting population change from models based on habitat availability and utilization. *Proceedings of the Royal Society B*, 286 (1901):20182911, 2019. doi: https://doi.org/10.1098/rspb.2018.2911.

J. Matthiopoulos, J. Fieberg, and G. Aarts. Species-habitat associations: Spatial data, predictive models, and ecological insights, 2020a.

J. Matthiopoulos, J. Fieberg, G. Aarts, F. Barraquand, and B. E. Kendall. Within reach? habitat availability as a function of individual mobility and spatial structuring. *The American Naturalist*, 195(6):1009–1026, 2020b. doi: https://doi.org/10.1086/708519.

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*, 100(1):112–124, 2003. doi: https://doi.org/10.1034/j.1600-0706.2003.12056.x.

P. McCullagh and J. Nelder. Generalised linear models.,(chapman and hall ltd: London.), 1983.

G. J. McLachlan and S. Rathnayake. On the number of components in a gaussian mixture model. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 4(5): 341–355, 2014.

G. J. McLachlan, S. X. Lee, and S. I. Rathnayake. Finite mixture models. *Annual review of statistics and its application*, 6:355–378, 2019.

S. Menard. *Applied logistic regression analysis*. Number 106. Sage, 2002.

R. Miao, P. N. Ghosh, M. Khanna, W. Wang, and J. Rong. Effect of wind turbines on bird abundance: A national scale analysis based on fixed effects models. *Energy Policy*, 132: 357–366, 2019.

K. Miura. An introduction to maximum likelihood estimation and information geometry. *Interdisciplinary Information Sciences*, 17(3):155–174, 2011.

C. Molnar. *Interpretable machine learning*. Lulu. com, 2020.

N. Mouquet, Y. Lagadeuc, V. Devictor, L. Doyen, A. Duputié, D. Eveillard, D. Faure, E. Garnier, O. Gimenez, P. Huneman, et al. Predictive ecology in a changing world. *Journal of Applied Ecology*, 52(5):1293–1310, 2015. doi: https://doi.org/10.1111/1365-2664.12482.

K. P. Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012. doi: https://doi.org/10.1080/09332480.2014.914768.

A. Mysterud and R. A. Ims. Functional responses in habitat use: Availability influences relative use in trade-off situations. *Ecology*, 79(4):1435–1441, 1998. doi: https://doi.org/10.1890/0012-9658(1998)079[1435:FRIHUA]2.0.CO;2.

A. Mysterud and R. A. Ims. Relating populations to habitats. *Trends in ecology & evolution*, 14(12):489–490, 1999. doi: 10.1.1.707.9728.

J. I. Myung and M. A. Pitt. Model comparison methods. *Methods in enzymology*, 383: 351–366, 2004.

T. T. Nguyen, H. D. Nguyen, F. Chamroukhi, and G. J. McLachlan. Approximation by finite mixtures of continuous density functions that vanish at infinity. *Cogent Mathematics & Statistics*, 7(1):1750861, 2020.

V. Nikulin, G. J. McLachlan, and S. K. Ng. Ensemble approach for the classification of imbalanced data. In *Australasian Joint Conference on Artificial Intelligence*, pages 291–300. Springer, 2009.

N. C. Oza and S. Russell. *Online ensemble learning*. University of California, Berkeley, 2001.

R. S. Paton and J. Matthiopoulos. Defining the scale of habitat availability for models of habitat selection. *Ecology*, 97(5):1113–1122, 2016. doi: https://doi.org/10.1890/14-2241.1.

K. Pearson. Method of moments and method of maximum likelihood. *Biometrika*, 28 (1/2):34–59, 1936.

B. G. Peterjohn. Some considerations on the use of ecological models to predict species' geographic distributions. *The Condor*, 103(3):661–663, 2001.

A. T. Peterson, M. Papes, and D. A. Kluza. Predicting the potential invasive distributions of four alien plant species in north america. *Weed Science*, 51(6):863–868, 2003.

B. Petitpierre, O. Broennimann, C. Kueffer, C. Daehler, and A. Guisan. Selecting predictors to maximize the transferability of species distribution models: Lessons from cross-continental plant invasions. *Global Ecology and Biogeography*, 26(3):275–287, 2017.

S. J. Phillips, R. P. Anderson, and R. E. Schapire. Maximum entropy modeling of species geographic distributions. *Ecological modelling*, 190(3-4):231–259, 2006. doi: https://doi.org/10.1016/j.ecolmodel.2005.03.026.

P. Pintelas and I. E. Livieris. Special issue on ensemble learning and applications, 2020.

J. Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

R. Polikar. Ensemble learning. In *Ensemble machine learning*, pages 1–34. Springer, 2012.

M. Poongodi, M. Malviya, C. Kumar, M. Hamdi, V. Vijayakumar, J. Nebhen, and H. Alyamani. New york city taxi trip duration prediction using mlp and xgboost. *International Journal of System Assurance Engineering and Management*, 13(1):16–27, 2022.

A. M. Prasad, L. R. Iverson, and A. Liaw. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems*, 9(2):181–199, 2006.

P. PRISM. Prism climate data (prism climate group, oregon state university), 2019. http://prism.oregonstate.edu.

D. Raghavarao, J. B. Wiley, and P. Chitturi. *Choice-based conjoint analysis: models and designs*. Chapman and Hall/CRC, 2010. doi: https://doi.org/10.1201/9781420099973.

M. Rakhra, P. Soniya, D. Tanwar, P. Singh, D. Bordoloi, P. Agarwal, S. Takkar, K. Jairath, and N. Verma. Crop price prediction using random forest and decision tree regression:-a review. *Materials Today: Proceedings*, 2021.

J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer, 2005.

C. F. Randin, T. Dirnböck, S. Dullinger, N. E. Zimmermann, M. Zappa, and A. Guisan. Are niche-based species distribution models transferable in space? *Journal of biogeography*, 33(10):1689–1703, 2006. doi: https://doi.org/10.1111/j.1365-2699.2006.01466.x.

S. Raschka, Y. H. Liu, V. Mirjalili, and D. Dzhulgakov. *Machine Learning with PyTorch and Scikit-Learn: Develop machine learning and deep learning models with Python*. Packt Publishing Ltd, 2022.

I. W. Renner and D. I. Warton. Equivalence of maxent and poisson point process models for species distribution modeling in ecology. *Biometrics*, 69(1):274–281, 2013. doi: 10.1111/j.1541-0420.2012.01824.x.

C. Ricotta. Bridging the gap between ecological diversity indices and measures of biodiversity with shannon's entropy: comment to izsák and papp. *Ecological Modelling*, 152 (1):1–3, 2002.

D. R. Roberts, V. Bahn, S. Ciuti, M. S. Boyce, J. Elith, G. Guillera-Arroita, S. Hauenstein, J. J. Lahoz-Monfort, B. Schröder, W. Thuiller, et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8): 913–929, 2017.

V. F. Rodriguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo, and J. P. Rigol-Sanchez. An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS journal of photogrammetry and remote sensing*, 67:93–104, 2012.

L. Rokach. Genetic algorithm-based feature set partitioning for classification problems. *Pattern Recognition*, 41(5):1676–1700, 2008.

K. V. Rosenberg, P. J. Blancher, J. C. Stanton, and A. O. Panjabi. Use of north american breeding bird survey data in avian conservation assessments. *The Condor: Ornithological Applications*, 119(3):594–606, 2017.

C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

N. Russell, L. Cribbin, and T. B. Murphy. upclass: An r package for updating model-based classification rules. *Cran R-Project Org*, 2014.

O. Sagi and L. Rokach. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1249, 2018.

K. L. Sainani. Explanatory versus predictive modeling. *PM&R*, 6(9):841–844, 2014.

O. E. Sala, F. Stuart Chapin, J. J. Armesto, E. Berlow, J. Bloomfield, R. Dirzo, E. Huber-Sanwald, L. F. Huenneke, R. B. Jackson, A. Kinzig, et al. Global biodiversity scenarios for the year 2100. *science*, 287(5459):1770–1774, 2000.

J. R. Sauer, B. G. Peterjohn, and W. A. Link. Observer differences in the north american breeding bird survey. *The Auk*, 111(1):50–62, 1994.

R. E. Schapire. The boosting approach to machine learning: An overview. *Nonlinear estimation and classification*, pages 149–171, 2003.

G. Schwarz. Estimating the dimension of a model. *The annals of statistics*, pages 461–464, 1978.

A. M. Sequeira, P. J. Bouchet, K. L. Yates, K. Mengersen, and M. J. Caley. Transferring biodiversity models for conservation: opportunities and challenges. *Methods in Ecology and Evolution*, 9(5):1250–1264, 2018. doi: https://doi.org/10.1111/2041-210X.12998.

R. P. Sheridan, W. M. Wang, A. Liaw, J. Ma, and E. M. Gifford. Extreme gradient boosting as a method for quantitative structure–activity relationships. *Journal of chemical information and modeling*, 56(12):2353–2360, 2016.

W. B. Sherwin and N. Prat i Fornells. The introduction of entropy and information methods to ecology by ramon margalef. *Entropy*, 21(8):794, 2019.

G. Shmueli. To explain or to predict? *Statistical science*, 25(3):289–310, 2010.

A. Sih and B. Christensen. Optimal diet theory: when does it work, and when and why does it fail? *Animal behaviour*, 61(2):379–390, 2001.

T. J. Smith and C. M. McKenna. A comparison of logistic regression pseudo r2 indices. *Multiple Linear Regression Viewpoints*, 39(2):17–26, 2013.

E. Sober. Instrumentalism, parsimony, and the akaike framework. *Philosophy of Science*, 69(S3):S112–S123, 2002.

P. Sollich and A. Krogh. Learning with ensembles: How over-fitting can be useful. *Neural Information Processing Systems (NIPS)*, 8:190–196, 1996.

J. P. Stevens. *Intermediate statistics: A modern approach*. Routledge, 2013.

G. W. Stewart. On the early history of the singular value decomposition. *SIAM review*, 35 (4):551–566, 1993.

H. A. Sturges. The choice of a class interval. *Journal of the american statistical association*, 21(153):65–66, 1926.

G. Sundblad, M. Härmä, A. Lappalainen, L. Urho, and U. Bergström. Transferability of predictive fish distribution models in two coastal systems. *Estuarine, coastal and shelf science*, 83(1):90–96, 2009.

M. Sunnåker, A. G. Busetto, E. Numminen, J. Corander, M. Foll, and C. Dessimoz. Approximate bayesian computation. *PLoS computational biology*, 9(1):e1002803, 2013.

G. Tallis and R. Light. The use of fractional moments for estimating the parameters of a mixed exponential distribution. *Technometrics*, 10(1):161–175, 1968.

G. Tessarolo, J. M. Lobo, T. F. Rangel, and J. Hortal. High uncertainty in the effects of data characteristics on the performance of species distribution models. *Ecological Indicators*, 121:107147, 2021. doi: https://doi.org/10.1016/j.ecolind.2020.107147.

T. M. Therneau, E. J. Atkinson, et al. An introduction to recursive partitioning using the rpart routines. Technical report, Technical report Mayo Foundation, 1997.

L. G. Torres, P. J. Sutton, D. R. Thompson, K. Delord, H. Weimerskirch, P. M. Sagar, E. Sommer, B. J. Dilley, P. G. Ryan, and R. A. Phillips. Poor transferability of species distribution models for a pelagic predator, the grey petrel, indicates contrasting habitat preferences across ocean basins. *PLoS One*, 10(3):e0120014, 2015. doi: https://doi.org/10.1371/journal.pone.0120014.

A. Townsend Peterson, M. Papeş, and M. Eaton. Transferability and model evaluation in ecological niche modeling: a comparison of garp and maxent. *Ecography*, 30(4): 550–560, 2007.

H. Travers, M. Selinske, A. Nuno, A. Serban, F. Mancini, T. Barychka, E. Bush, R. A. Rasolofoson, J. E. Watson, and E. Milner-Gulland. A manifesto for predictive conservation. *Biological Conservation*, 237:12–18, 2019. doi: https://doi.org/10.1016/j.biocon.2019.05.059.

R. E. Ulanowicz. Information theory in ecology. *Computers Chemistry*, 25(4):393–399, 2001. ISSN 0097-8485. doi: https://doi.org/10.1016/S0097-8485(01)00073-0. URL https://www.sciencedirect.com/science/article/pii/S0097848501000730.

T. Václavík and R. K. Meentemeyer. Invasive species distribution modeling (isdm): are absence data and dispersal constraints needed to predict actual distributions? *Ecological modelling*, 220(23):3248–3258, 2009.

W. Vanreusel, D. Maes, and H. Van Dyck. Transferability of species distribution models: a functional habitat approach for two regionally threatened butterflies. *Conservation biology*, 21(1):201–212, 2007.

W. N. Venables and B. D. Ripley. *Modern applied statistics with S-PLUS*. Springer Science & Business Media, 2013.

J. Verner. Assessment of counting techniques. In *Current ornithology*, pages 247–302. Springer, 1985.

D. Warton and G. Aarts. Advancing our thinking in presence-only and used-available analysis. *Journal of Animal Ecology*, 82(6):1125–1134, 2013. doi: 10.1111/1365-2656.12071.

D. I. Warton and L. C. Shepherd. Poisson point process models solve the" pseudo-absence problem" for presence-only data in ecology. *The Annals of Applied Statistics*, pages 1383–1402, 2010. doi: 10.1214/10-AOAS331.

S. J. Wenger and J. D. Olden. Assessing transferability of ecological models: an under-appreciated aspect of statistical validation. *Methods in Ecology and Evolution*, 3(2): 260–267, 2012.

S. J. Wenger, D. J. Isaak, J. B. Dunham, K. D. Fausch, C. H. Luce, H. M. Neville, B. E. Rieman, M. K. Young, D. E. Nagel, D. L. Horan, et al. Role of climate and invasive species in structuring trout distributions in the interior columbia river basin, usa. *Canadian Journal of Fisheries and Aquatic Sciences*, 68(6):988–1008, 2011.

G. O. Wogan. Life history traits and niche instability impact accuracy and temporal transferability for historically calibrated distribution models of north american birds. *PLoS One*, 11(3):e0151024, 2016.

S. N. Wood. *Generalized additive models: an introduction with R*. CRC press, 2017.

G. Xuan, W. Zhang, and P. Chai. Em algorithms of gaussian mixture model and hidden markov model. In *Proceedings 2001 international conference on image processing (Cat. No. 01CH37205)*, volume 1, pages 145–148. IEEE, 2001.

L. Yang, S. Jin, P. Danielson, C. Homer, L. Gass, S. M. Bender, A. Case, C. Costello, J. Dewitz, J. Fry, et al. A new generation of the united states national land cover database: Requirements, research priorities, design, and implementation strategies. *ISPRS journal of photogrammetry and remote sensing*, 146:108–123, 2018.

K. L. Yates, P. J. Bouchet, M. J. Caley, K. Mengersen, C. F. Randin, S. Parnell, A. H. Fielding, A. J. Bamford, S. Ban, A. M. Barbosa, et al. Outstanding challenges in the transferability of ecological models. *Trends in ecology & evolution*, 33(10):790–802, 2018. doi: https://doi.org/10.1016/j.tree.2018.08.001.

H. Yu, T. Xie, S. Paszczyñski, and B. M. Wilamowski. Advantages of radial basis function networks for dynamic system design. *IEEE Transactions on Industrial Electronics*, 58 (12):5438–5450, 2011.

C. Zhang and Y. Ma. *Ensemble machine learning: methods and applications*. Springer, 2012.

Z.-H. Zhou. *Ensemble methods: foundations and algorithms*. CRC press, 2012.

D. Zurell, F. Jeltsch, C. F. Dormann, and B. Schröder. Static species distribution models in dynamically changing systems: how good can predictions really be? *Ecography*, 32 (5):733–744, 2009. doi: https://doi.org/10.1111/j.1600-0587.2009.05810.x.