FACTORS AFFECTING VARIABLE IMPORTANCE ESTIMATIONS FROM SPECIES DISTRIBUTION MODELS: A VIRTUAL ECOLOGIST APPROACH

NIVEDITA VARMA HARISENA
Enschede, The Netherlands, February 2019

SUPERVISORS:
dr. Ir.T. A. Groen
dr. A.G.Toxopeus

FACTORS AFFECTING VARIABLE IMPORTANCE ESTIMATIONS FROM SPECIES DISTRIBUTION MODELS: A VIRTUAL ECOLOGIST APPROACH

NIVEDITA VARMA HARISENA
Enschede, The Netherlands, February 2019

Thesis submitted to the Faculty of Geo-Information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation.
Specialization: Natural Resource Management

SUPERVISORS:
dr. Ir.T. A. Groen
dr. A.G.Toxopeus

THESIS ASSESSMENT BOARD:
dr. Y.A.Hussin
dr. B. Naimi (External Examiner, ETH Zurich)

DISCLAIMER

This document describes work undertaken as part of a programme of study at the Faculty of Geo-Information Science and Earth Observation of the University of Twente. All views and opinions expressed therein remain the sole responsibility of the author, and do not necessarily represent those of the Faculty.

ABSTRACT

Aim: The aim of this study was to investigate three factors of 'spatial dependency' in input parameters namely: 1) spatial autocorrelation in predictor variables; 2) types of species response curve geometries; 3) varying sampling densities and analyse their effect on model-independent variable importance estimations derived from species distribution models.

Method: This study uses simulated data of both the environmental predictors and the species response curves. Twenty-five levels of spatial autocorrelation (SAC) in combination with four types of species based on response types (linear, unimodal and combinations of these) and three levels of sampling density were analysed. The simulations were also run for two scenarios of relative SAC (0% background and 12.5% background). The choice of models includes eight models: Generalised Linear model (GLM); Generalised Linear Mixed Model (GLMM) with spatial random effect; Generalised additive model (GAM); Maximum Entropy (MaxEnt); Random Forest (RF); Boosted Regression Trees (BRT); Support Vector Machines (SVM) and Artificial neural networks (ANN). From each of the models, the variable importance (based on a model-independent randomisation technique) is calculated on an independent simulated test dataset, along with the reporting of the area under the Receivers Operating Characteristic curve, kappa and the autocorrelation in the residuals.

Results: The results showed that for all four species all the models estimated biased importance towards the highly autocorrelated predictors, but the magnitude of bias was higher for the linear response species. The threshold relative SAC within which there was no bias was also narrower for the said species. Another significant result is the importance bias towards the unimodal responses when compared to a linear response from all the models. Changing sampling densities did not have an observable effect. The RF and SVM were the most robust amongst all the models.

Main conclusions: The type of response curve geometry, which are mainly dictated by species characteristics (i.e. narrow or wide-ranging species), along with the relative SAC of the covariates were the most determining factors for bias in relative variable importance estimations due to spatial autocorrelation in predictors. Therefore, species response characteristics along with the relative spatial structure of predictor variables must be given due consideration for making proper inferences about variable importance from species distribution models.

ACKNOWLEDGEMENTS

First and foremost I would like to thank my first supervisor dr. Thomas Groen for his immeasurable guidance and for providing me with the creative freedom to try out new ideas; always guiding and reorganising my thoughts but never overriding them, for this I am highly thankful. I would also like to thank dr. Toxopeus for his unconditional support in the completion of this thesis. I would also like to thank the other staff at ITC, NRM dept for their valuable comments and guidance throughout the course.

I would like to thank my friends here at ITC, especially my batchmates in NRM Dept., for their warm camaraderie and lively company. Without them, my term here would have been as dreary as the cloudy, sunless days of winter. Issamladin Mohammed, you deserve special mention because I would have been lost without your constant confidence boosts. My floormates Isaac Oliech, Eko Kustiyanto, Mary Mukuba, Calyne Khamila, Silas Afwamba and Max Korir and the many others who have brought me out of isolation with lively cooking sessions. Eko, your dinner parties were akin to heaven. Mary your chapatis and ginger tea helped me recover from the many blues of life. And Isaac, Calyne, Silas and Max, hearty conversations with you made my day.

I would also like to extend my thanks to my fellow country-people, Sheeba Lawrence, Mitali Joshi, Richa Maheshwari, Ashish Vivekar and Harinish Palanirajan. Your constant company kept me close to home.

And finally, no acknowledgement is complete without extending my gratitude to my family. I would like to thank my brother, without him I would not have been here, my mother without whose unconditional backing I would have been forever lost, my father, without whom my philosophy to life would have been lacking and of course Vidyasagar S., for his constant emotional support. Without them, nothing in life was ever and will ever be possible. I thank them wholeheartedly for being there for me, always.

Thank you.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1. Background

Predictive modelling of species distributions is an important tool to analyse the impact of a changing environment on plant and animal communities (Guisan and Zimmermann 2000). Similar to urban area demarcations and the strict spatial extents of city plans, the natural environment also has certain 'niches' where the largest populations survive. Understanding the spatial patterns of these areas can also help gain insight into the environmental conditions in which the species survive and thus enable proper management and planning for their futures (Franklin and Miller 2010). This is where Species distribution modelling or ecological niche modelling finds its own proverbial home.

One of the major contributions of species distribution modelling is in the estimation of possible causal relationships between the environment and the species. As several studies now intend to define the ecological niche ('realised' niche) of a species beyond simply its geographical ranges, species distribution modelling is being increasingly used to define what variables (abiotic and biotic) are significant as well as important for the species (Austin et al. 1990, Austin 2002, Berdugo et al. 2019, Meineri et al. 2012). Efforts have also been made to estimate 'generalisable' and 'simpler' models that can define the species niche across different spatiotemporal frames(Duque-Lazo 2013). To do so, many studies revert to reducing the wide gamut of possible explanatory variables to a shorter array of most 'important' ones. Therefore, measures of variable importance are commonly used in ranking species-environment relationships. This is also sometimes referred to as sensitivity analysis of the model to different predictors (Wei et al., 2015). Model-independent Variable Importance Measures (VIM's) quantify how a certain input or absence of it (from a pool of candidate variables) affects the output of a model, either in its accuracy or in terms of increase/decrease of uncertainty (Thuiller et al. 2009, Wei et al. 2015). This kind of analysis provides us with meaningful relations that can aid in simplifying models for practical purposes i.e. to improve transferability or for cost and time effectiveness(Duque-Lazo 2013). Beyond this, the use of P-values, standardised coefficients and partial response curves can also help in comparison of variable influence across models (Naimi and Araújo 2016). Thus, as species distribution/ecological niche modelling is now being increasingly used to estimate important species-environment relationships, it is imperative to understand the statistical assumptions behind the models so as to make an effective contribution to ecological theory (Austin, 2006; Austin, 2002; Guisan et al., 2006).

One of the confounding aspects often neglected in ecology is that of spatial dependency of covariates. Where the biotic components like dispersal are considered to be a manifestation of spatial patterns, the contribution of spatial dependence of other abiotic covariates like temperature and humidity are often overlooked (De Knegt et al. 2010). Even in the abiotic components there is a spatial pattern that entails a possible conflict with basic statistical assumptions. Spatial statistics, which is the basis for species distribution modelling, tackles many such spatial dependency concepts like edge effects, spatial outliers, modifiable areal unit problem etc (Fotheringham et al., 2000). Spatial autocorrelation (SAC) is one such concept that disrupts the fundamental assumption of classical statistical inference: that of independence of observations. Observations scattered in a certain spatial extent are prone to certain structure that creates a covariance pattern between neighbouring points. This stems out of a fundamental law of geography known as Tobler's law, which states that "Everything is related to everything else, but near things are more related than distant things" (Harvey J. Miller 2004). In statistical terms the true degree of freedom is actually lower than the no. of observations due to pseudo-replication of data (nearer things have the same value); thus the estimated significance is biased (Legendre 1993, Lennon 2000) . Therefore, due to the reasons mentioned above, SAC in models can affect the determination of important species-environment relationships (Dormann 2007).

SAC in species observations can occur mainly in two ways: First due to inherent endogenous processes (e.g. dispersal or species interaction; as mentioned above) that are distance related, and secondly due to exogenous conditions i.e. a structure in the predictor variables that species respond to uniformly or linearly (De Knegt et al. 2010). SAC is usually described in input data or model residuals by using global measures such as Moran's I, Geary's c or the semi-variogram; whereas local associations can be assessed using Local Indicator of Spatial Association (LISA) (Naimi 2015). Analysing these measures can help delimit the nuances that plague statistical models, hampering proper ecological inference.

A second main contender that structures the spatial dependency of covariates, and thus, engenders statistical misinterpretation of ecological causal mechanisms, are the varying species response geometries. Species responses to any one variable is rarely a result of an independent ecological process and, thus, variations and complexities are inherent in its basic structure (Austin 2002, Oksanen and Michin 2002). Therefore even though in theory one would expect a unimodal response curve for most environmental conditions (as per the niche theory by Hutchinson 1957), still most computed response curves from observations can take various forms from linear to asymmetric/skewed curves (Rydgren et al 2003). This 'skewness' can also be due to limitations in sampling along environmental gradients, owing to which we might be able to observe only a part of these curves (truncated)(Austin and Nicholls 1997). Such variations can further bias the detection of variable importance from correlative models (Austin 2006). Traditionally unimodal responses are considered quite common, and any skew in it is regarded as a physiological effect of geographic limitations or interactive ecological (biotic) processes, though in the modern data-rich world the skewed distributions can also be an effect of interplay of geographic extent and data resolution along with species characteristics (De Knegt et al. 2010, Rydgren et al. 2003). The effects of different kinds of response curves on model inference have been immensely studied with model accuracy, complexity and overfitting being the highlight (Bell and Schlaepfer 2016), yet the independent effects of the different geometries of response curves on the estimation of variable importance have not yet been explored.

And lastly, the sampling density is another crucial parameter as it forms the bridge between the population characteristics and the model inputs. Amongst many other aspects, the sampling density directly influences the level of autocorrelation in the covariate data (spatial dependency). Many studies have highlighted the use of lower sampling densities to counter spatial autocorrelation in samples (Hawkins et al. 2007, Legendre and Fortin 1989). However, this can lead to loss of a lot of information as the data is thinned out (Fortin and Dale 2005) and thus can increase the variability in results due to the lower degrees of freedom for modelling. Therefore, the effect of changing sampling densities in estimations of model-independent variable importance is a useful parameter to investigate and has been identified as one of the objectives in this study.

1.2. Problem statement

Spatial autocorrelation and its applications in ecology have been highlighted in numerous studies since the 1980s (Legendre 1993, Legendre and Fortin 1989). Along with the statistical problems of bias introduced in the previous paragraphs, many studies have discussed the practical problems of using datasets with considerable spatial autocorrelation. Lennon (2000) first pointed out that the chance of highly spatially autocorrelated variables being modelled as falsely significant is high. This means that the importance rankings based on significance can be a reiteration of the levels of autocorrelation in the variables. Thus, the variable with the highest spatial structure poses as a red-shifted positively biased parameter. This finding was also reiterated by P. Segurado, Araújo, & Kunin, 2006. However, following up on this study, Diniz-Filho et al. (2003) and Hawkins et al. (2007) argued that the positive shift for correlated macro-scale variables are in fact an ecological mechanism and is not a spurious association. Nevertheless, this might just be a specific case, and the chance of small scale mechanisms being brushed under the rug is possible and thus important ecological interpretations can be lost (De Knegt et al. 2010, Pedro Segurado and Araújo 2004).

Unlike the debate on the effect of autocorrelation, the effect of different geometries of species-environment relationships (response curve geometries) and their effect on predictor variable significance/importance have only been given limited attention. Two studies that explicitly analyse this relationship are Meynard & Quinn, 2007 and Santika & Hutchinson, 2009. Both used a range of geometries including linear, unimodal, beta and threshold responses to analyse their effect on model predictions. Santika & Hutchinson, 2009 identified that in a situation where the linearly responded predictor (linear covariate) was simulated to be more dominant, the univariate model including only the linear covariate showed low values of AUC with higher variability, when compared to that from a unimodal covariate, in a few modelling methods (BIOCLIM). Meynard and Quinn 2007, also noted a bias against linear response curves in the predictive power (AUC, sensitivity, kappa) of the models. This present study differs from these in terms of estimating biases in model accuracy, as well as model independent variable importance, within a fully specified model (and not a univariate one). Therefore, the relative effects of different geometries in the same model (in combination with the effect on spatial autocorrelation) can be assessed.

Returning to the issue of spatial autocorrelation we look at another line of debate regarding the use and benefits of spatial models to account for autocorrelation errors in both coefficient estimation and its significance (Bahn et al. 2006; Keitt, Bjørnstad et al., 2002). Some studies have shown that the ranking of variable importance, based on significance, changes when using spatial vs non-spatial methods (Diniz-Filho et al. 2003, Kühn 2007). These could be evidences of important small-scale mechanisms being highlighted that were initially lost in non-spatial models. However, many studies have reported that not all spatial methods give better unbiased estimates and that there is high variability in output in terms of the different spatial models used; mainly because of the many model decisions required to account for the spatial pattern that can truly represent the spatial ecological process (Betts et al 2007, Dormann et al 2007, Kissling and Carl 2008). A study by De Knegt *et al.* (2010) found that spatial methods are more accurate for predictions, however in terms of coefficient estimation they can fail to account for large scale processes (usually exogenous SAC) that are correlated to the spatial error term, thus biasing against the broad scale environmental variable in favour of smaller scale spatial patterns (usually endogenous SAC). Thus the choice of models depends highly on the ecology of the species involved and the scale of analysis (Levy 1992). Another study by Bini *et al.* (2009) used data from 97 real cases (as opposed to the many simulated datasets used in the cases above) with varying levels of SAC to analyse the same and could not find any conclusive evidence to back up the claim both about the effect of SAC and the usefulness of spatial models. Therefore, in light of the many contradictory results, there is a need to conclusively investigate this relation based on both spatial and non-spatial models.

Although the debate has shifted to question the importance of spatial models over non-spatial ones, the original argument about the increasing importance of highly autocorrelated explanatory variables still holds in the inflation of significance of estimates. The impact of this on estimated model independent variable importance values, using different types of modelling methods, needs to be investigated along with the influence of different response curve geometries and sampling densities. This is where the current study finds its main purpose.

One of the best ways to analyse the statistical inferences of models is by using a virtual species distribution model (VSDM). VSDM is an efficient method developed by (Hirzel *et al.,* 2001) to address inherent issues of spatial statistical models. This involves the use of a set of 'virtual' species in a predefined environment amidst other controlled settings like defined species responses, levels of autocorrelation and sampling schemes. The true distribution is known here, and thus the characteristics of models and related statistical measures can be properly investigated without other confounding variables (unlike real datasets that can be affected by numerous hidden or unknown aspects) (Miller, 2014; Naimi, 2015). The use of such a method can help determine the discrepancies in variable importance with varying spatial autocorrelation. Thus, the study can help expand our understanding of the vagaries associated with variable importance estimations from species distribution models and provide a better perspective on the influence of spatial autocorrelation, response curve geometry and sampling density on them.

1.3. Research Objectives

The main objective of the study is to analyse the effect of spatial autocorrelation (SAC), response curve geometry and sampling density on the estimation of predictor variable importance in correlative species distribution modelling.

1.3.1. The sub-objectives are: -

1. To analyse how varying levels of exogenous SAC affect variable importance estimation from generalised linear, spatial generalised linear and machine learning methods.
2. To analyse the effect of different geometries of response curves on variable importance estimations from generalised linear, spatial generalised linear and machine learning methods.
3. To analyse the effect of changing the sampling density levels on variable importance estimations from generalised linear, spatial generalised linear and machine learning methods.
4. To compare the various models in their overall accuracy and their robustness at estimating accurate variable importance.

1.4. Research Questions

The main research questions are:

1) How do varying levels of exogenous SAC affect the variable importance estimation from generalised linear, spatial generalised linear and machine learning methods?
2) How do different combinations of response curve geometries affect variable importance estimation generalised linear, spatial generalised linear and machine learning methods?
3) What is the effect of sampling density on the variable importance estimation of differently responding autocorrelated variables from generalised linear, spatial generalised linear and machine learning methods?

4) What is the most accurate and robust modelling method, amongst generalised linear, spatial generalised linear and machine learning methods, at estimating accurate variable importance?

1.5. Research hypotheses

From the review of current literature with regards to the topic, the main hypotheses for the research are:

1) H_{1-} There is an increasing trend in estimated variable importance as the percentage exogenous SAC increases.

2) H_{1-} The variable importance of linear geometries of response curves is lower than those from unimodal geometries.

3) H_{1-} The lower the sampling density, the lower the change in variable importance due to the increase in SAC.

4) H_{1-} The variable importance estimations from the spatial generalised linear mixed models are the least affected by changing levels of SAC when compared to non-spatial models.

2. METHOD:

The overall framework of the study has been captured in Fig.1. The main steps include the simulation of the covariates and the species followed by the running of the various models, after which the variable importance values are extracted or computed and finally analysed using graphs and boxplots. All computations are to be done on R statistical software that is freely available online (R Core team 2017).

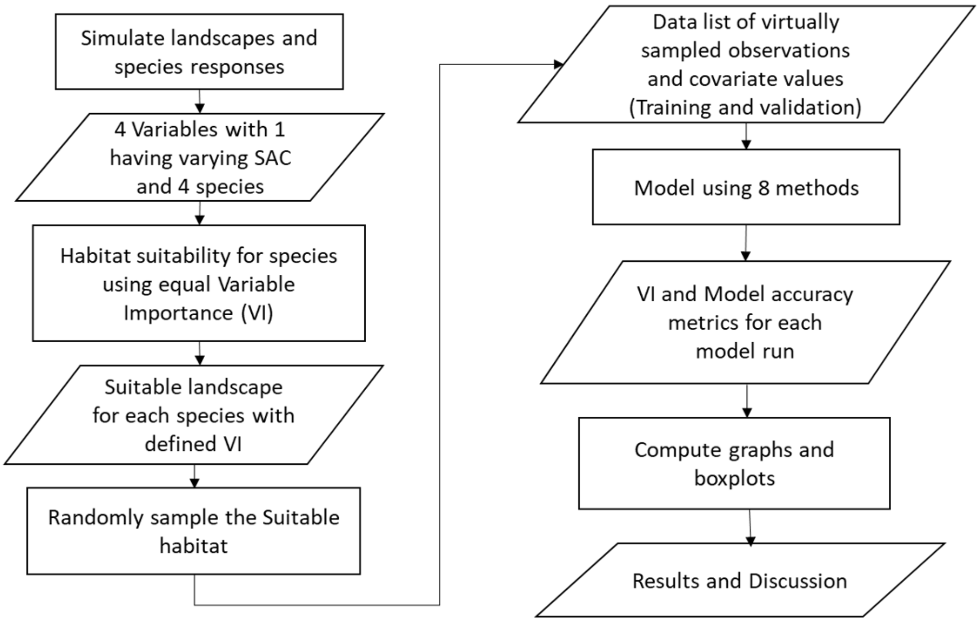

Simulate landscapes and species responses
4 Variables with 1 having varying SAC and 4 species
Habitat suitability for species using equal Variable Importance (VI)
Suitable landscape for each species with defined VI
Randomly sample the Suitable habitat
Data list of virtually sampled observations and covariate values (Training and validation)
Model using 8 methods
VI and Model accuracy metrics for each model run
Compute graphs and boxplots
Results and Discussion

Figure 1: Flowchart showing the overall methodology of the study

2.1. Setting up the experimental design

The experiment has been set up to cater to the main objective of assessing the relative variable importance estimations. To assess the relative effects of different combinations of predictor characteristics, four independent covariates were simulated. The choice of no. of covariates was fixed at four so as to accommodate the various combinations of response curves and autocorrelation within each species type (see section 2.1.3 for details). Further for each variable 20 realisations were simulated and were used consistently across all model runs. As mentioned before four types of species (A, B, C and D) are formulated, each with a varying response combination to the input variables. Finally, to check the effect of autocorrelation, one of the covariates, and its 20 realisations, is made gradually more spatially autocorrelated. Two scenarios are maintained throughout the experiments involving the baseline fixed autocorrelation of the covariates used for comparison (background variables) against the covariate with varying autocorrelation. Three of the simulated covariates (V1, V2 and V3) are kept constant at two cases: 1) 0% SAC and 2) 12.5% SAC. The fourth covariate (V4) is simulated with a varying autocorrelation from 0% to 30%, with 25 levels of autocorrelation in between, to make the steps as gradual as possible. A basic schematic of the steps involved in the experimental design can be seen in Fig.2.

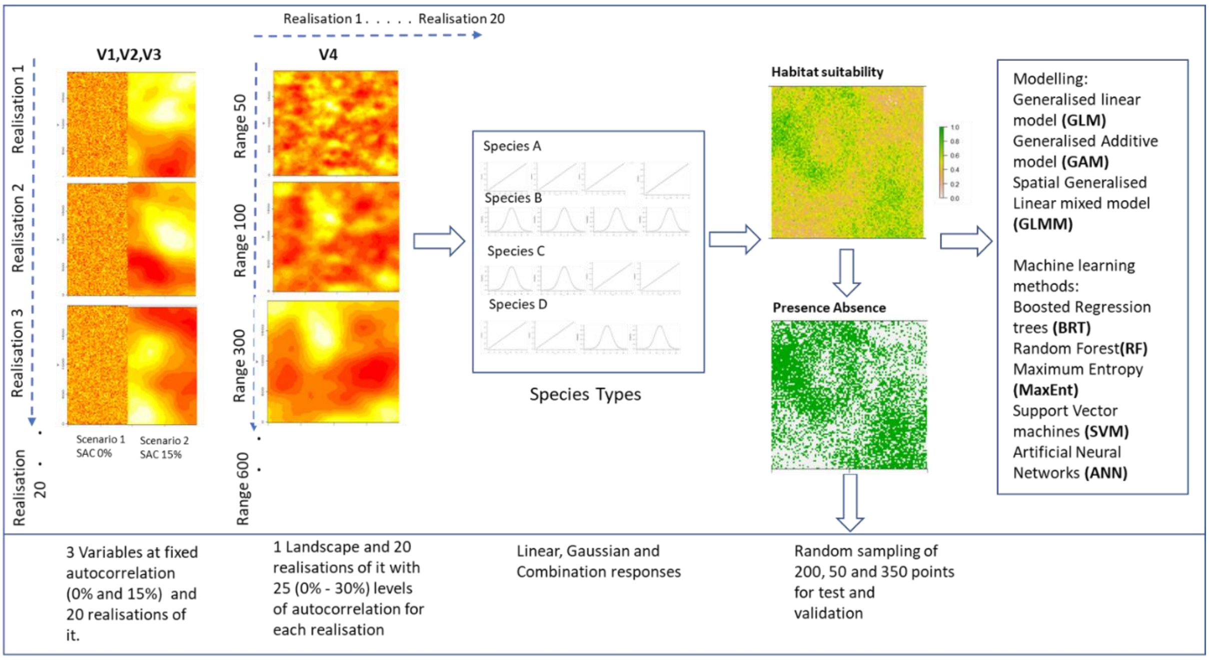

Realisation 1 . . . . . Realisation 20
V1,V2,V3
V4
Realisation 1
Realisation 2
Realisation 3
Realisation 20
Range 50
Range 100
Range 300
Range 600
Scenario 1 SAC 0%
Scenario 2 SAC 15%
Species A
Species B
Species C
Species D
Species Types
Habitat suitability
Presence Absence
Modelling:
Generalised linear model (GLM)
Generalised Additive model (GAM)
Spatial Generalised Linear mixed model (GLMM)
Machine learning methods:
Boosted Regression trees (BRT)
Random Forest(RF)
Maximum Entropy (MaxEnt)
Support Vector machines (SVM)
Artificial Neural Networks (ANN)
3 Variables at fixed autocorrelation (0% and 15%) and 20 realisations of it.
1 Landscape and 20 realisations of it with 25 (0% - 30%) levels of autocorrelation for each realisation
Linear, Gaussian and Combination responses
Random sampling of 200, 50 and 350 points for test and validation

Figure 2: Experimental design of the study

2.1.1. Simulating the virtual landscapes (covariates)

All the different variables were simulated on a grid representing a 100 by 100 pixel area. For the purpose of simulation, the pixel size had to be limited to 20 by 20 metres, so as to be able to define the autocorrelated range as a Euclidean distance. However, the analysis and reporting of results only mention the autocorrelation as a percentage of the extent to allow for scale-free interpretation of the results.

The explanatory variables are simulated as unconditional gaussian random fields, based on the above-mentioned grid, that have been given a certain covariance structure (Beguería and Pueyo 2009, Dormann et al. 2007, Nychka et al. 2017). Therefore, initially for all the variables, a covariance matrix is defined where the correlation between two pixels is defined as:

$$exp(-\frac{D_{ij}}{theta})$$

where $D_{.ij}$ is the Euclidean distance between $x1_{[i,]}$ and $x2_{[j,]}$ and theta is the range of autocorrelation.
This simulates an autocorrelated surface roughly with variograms as shown in Fig. 3.

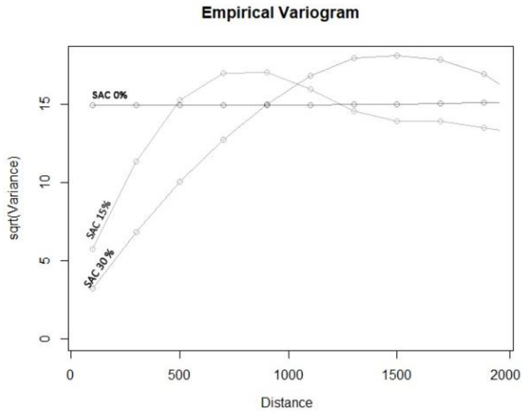


Figure 3: Variogram plot for 3 levels of SAC (0%,12.5% and 30%)

The covariance function defines stationary and isotropic autocorrelation to limit the variability in the results, though this might be a limitation to the exact application of the study to real data. Theta was defined as varying from 0% to 30% (1 to 600 units of the total 2000 units extent), as a saturation of the matrix was seen after the 30% mark as seen in the variogram in Fig.3. This could also be a manifestation of the method used here, as an unconditional simulation using a variogram directly instead of a covariance matrix maximises at around 70% autocorrelation range. The study has adopted this specific method due to time limitations as method 1 took exponentially increasing time to make landscapes for higher autocorrelation levels (as the neighbourhood distances used to calculate each pixel increased). The differences in detail is evident as seen in Fig.4, yet for this study, the amount of unique pixel by pixel calculation demanded in the latter method (method 1 in Fig.4) was not required since the models were to run on sampled datasets that further limits the amount of covariate information captured.

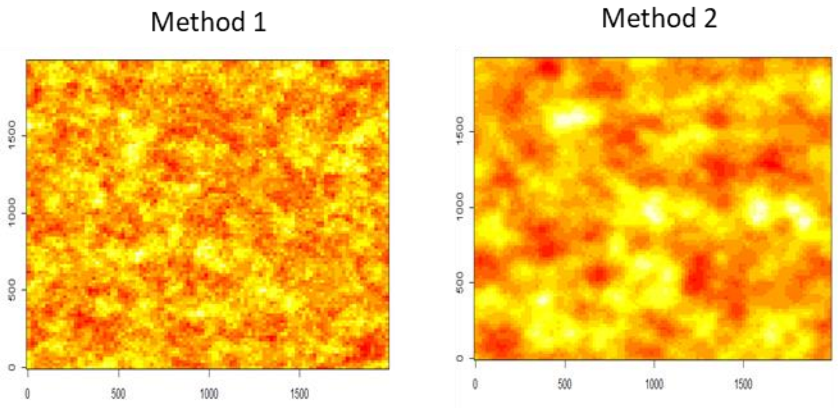


Figure 4: Simulated covariates using a) Method 1: unconditional simulation using defined exponential variogram ('gstat' package in R); b) Method 2: Unconditional simulation using defined exponential covariance matrix ('fields' package in R)

The covariance matrix of covariates V1, V2 and V3 were defined with two scenarios of theta values: 1 (0%) and 250 (12.5%), such that effectively the relative variable importance can be calculated for when the comparative baseline variables are low in autocorrelation and also when they are high. The fourth variable (V4) incorporates variable theta values ranging from 0(%) to 600(30%) with 25 values in between (see Fig.2). Therefore, a gradual but increasing effect of autocorrelation can be analysed.

To make different realisations of the same level of autocorrelation, each covariance matrix was then multiplied with 20 realisations of a Gaussian random error derived from rnorm~ (0,30). So, for the same level of autocorrelation different 'geometries' of the covariate are produced while preserving the basic variogram structure (see Fig.5). This implied that a dataset of 20 iterations (of the same conditions) can be used to derive the confidence interval of the differences perceived in the variable importance estimation

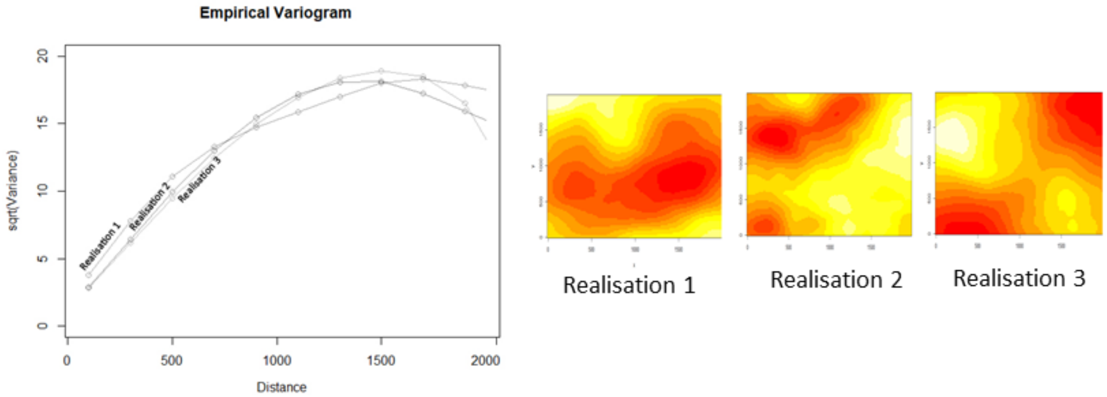


Figure 5: Variogram for different realisations of the same level of autocorrelation and the rasters that are created alongside.

(see section 2.3). So, in the end, the total number of simulated random fields include: 20 realisations of Variable 1,2 and 3 each; 20 realisations of variable 4 at each step of autocorrelation range, i.e. a total of 60 + 500 simulated input variables.

2.1.2. Simulating virtual species

For simulating the virtual species, it is imperative to define the species response to each covariate. Therefore, based on species characteristics, the area of occurrence (suitable ranges) within a covariate range is defined in the environmental space, which has implications directly for the geographical range of the species.

For this study, two different responses were chosen namely: monotonic (linear increasing) and unimodal. Many studies have debated the popular use of unimodal curves to model species responses (Austin 2002), yet for the simplicity of this study and to minimise the number of permutations, the choice of response curves were limited to these two basic geometries (see Fig. 6).

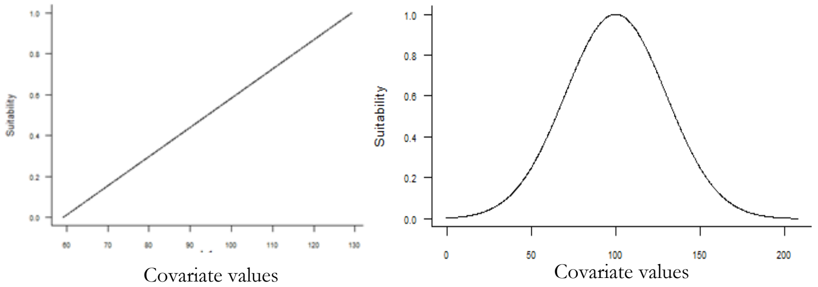


Figure 6: The two basic response curves used in the study: a) Monotonic (linear); b) Symmetric unimodal

Other possibilities of defining a species response like beta distributions or bimodal distributions can also be a case in the real world (Rydgren et al. 2003), yet the study does not go into an exhaustive possibility of these.

2.1.2.1. The unimodal response

The presence of an environmental optimum (indicator value) for a species, identified as its ecological niche, has been around in ecology for a long time (Hutchinson 1957) and is used in many current studies also (Graham and Duda 2011, Jamil et al. 2014, Santika and Hutchinson 2009). The application of the theory predicts a unimodal response curve for the species in response to environmental gradients, which can be identified by three parameters: the optimum or preferred environment, the tolerance or width of the curve and the maximum suitability achieved at environmental optimum. For this study, the curve has a preferred environment around 100 units (the range varying from 40~70 to 130~ 140), the tolerance (standard deviation) is maintained at a precise 15% of the total extent unless experimentally forced to alter and the maximum suitability is maintained at 1 to ensure all variables have equal importance.

2.1.2.2. The monotonic response

The monotonic response or a linearly increasing response to an environment was modelled with a constant slope at 45 degrees inclination from either axis. Though not widely favoured in ecological literature, these responses have still been used to model a form of a truncated version of the optimum curve described above (Guo 2014, Rydgren et al. 2003), since it is not possible to always have full coverage of the entire extents of species occurrence.

Therefore, a roster of species (A, B, C & D) was simulated based on the above-mentioned response curves and the varying combination of these with an autocorrelated covariate, the details of which will be shown in the next section.

2.1.3. Schematic of virtual species generated

The main four species generated include the following: (also see Fig.7)

1) Species A that has monotonic (positive linear) response to all the variables, but one of the variables is varied in its level of autocorrelation.
2) Species B that has unimodal (precise gaussian) response to all the variables, but one the variables is varied in its level of autocorrelation.
3) Species C has a unimodal response to the first two variables, monotonic response to latter two of which one of the monotonically responded variables varied in its level of autocorrelation.
4) Species D has a monotonic response to the first two variables, unimodal response to latter two of which one of the unimodally responded variables varied in its level of autocorrelation.

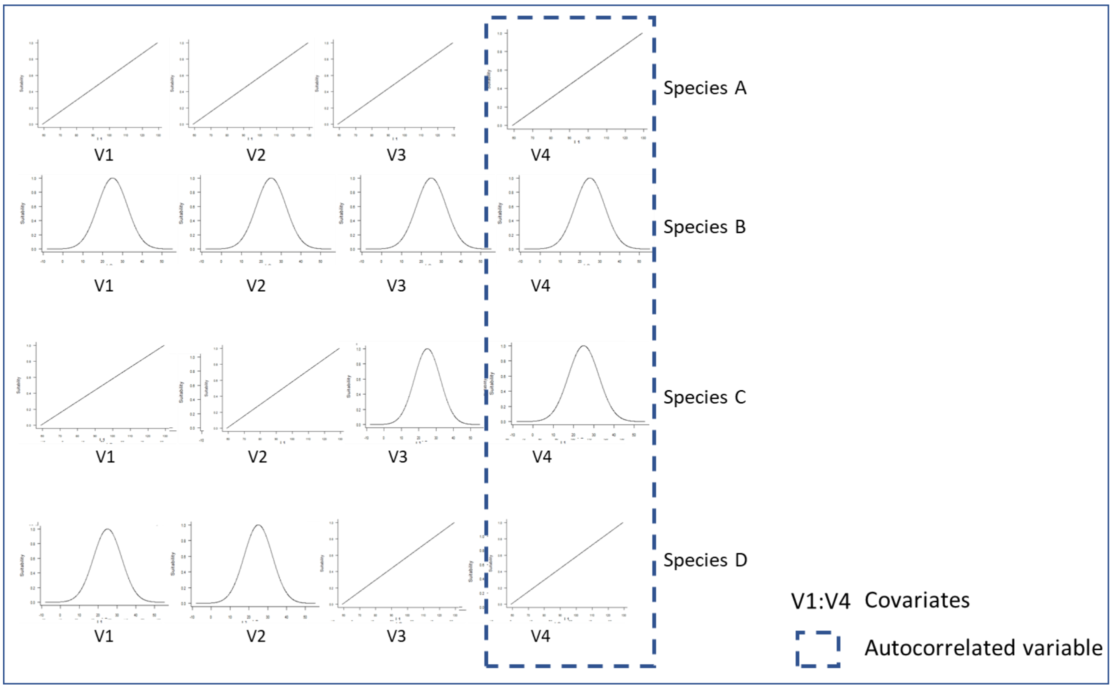


Figure 7: Response curve combinations for the four species (A, B, C, and D)

Species A and B were generated to analyse the effect of autocorrelation differently on unimodal and linear response curves; therefore each species has a single type of response to all four covariates, thus isolating the effect of relative autocorrelation on variable importance estimation for both monotonic and unimodal curves individually. Species C and D were generated to analyse the impact of combinations of response curves and autocorrelation in the estimation of variable importance and to analyse the relative variable importance of such two differently responded variables.

2.1.4. Virtual species to cater to three different 'sub-experiments'

Beyond this standard narrative of the four species and the autocorrelated variables, a few other species were also generated to test the effect of changing geometries lying between the linear and the unimodal, as well as to test the effect of imprecise unimodal curves. Details of these 'sub-experiments' have been shown herewith. All sub-experiments were conducted on covariates with 0% autocorrelation.

2.1.4.1. Species generation to test geometries between linear and unimodal

Since truncation of covariate range extents can occur at any cut-point, response curves in between the linear and the symmetric unimodal are highly probable (Austin and Nicholls 1997). Therefore, to test the effect of such geometries and to analyse the effect of gradually shifting the response curve geometry from linear to unimodal the following set of species were generated. This was done by fixing a unimodal response curve as shown in section 2.1.2.1 and the varying the value range of the input variable to create truncated forms of the response curves. Nine levels of partial response curves were generated (10%,20%,30…..90%) with the 10th representing the full unimodal response (see Fig,8). Ten different species was generated using these ten forms of response curves for one covariate (L1- 10%; L2- 20%; L3……L10- 100%) and a constant fixed linear response for the second (background) covariate.

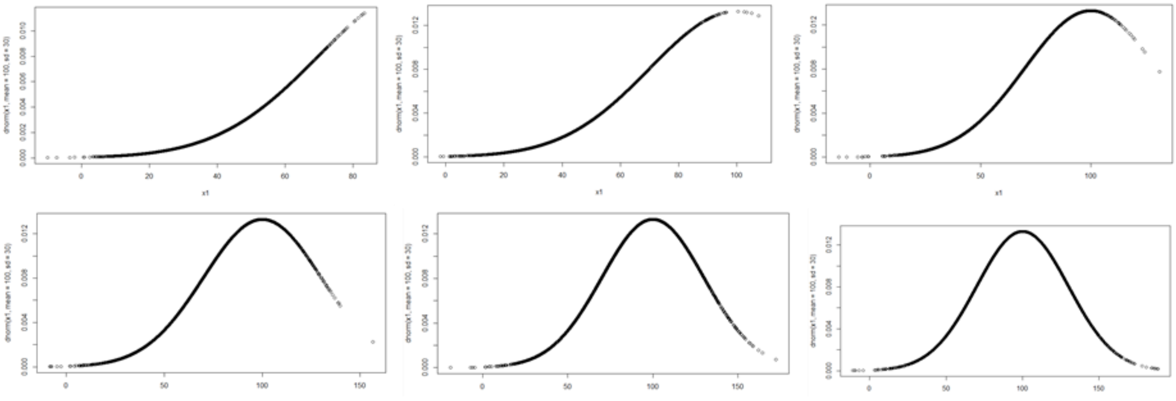


Figure 8: Series of geometries L1: L10 (10 in count of which 6 are shown), each representing a new species

2.1.4.2. Species generation to test the effect of varying width of unimodal response curves

One of the ways to characterise a species unimodal response is by their tolerance levels to different environmental gradients. This can be represented as the 'precision' or statistically the standard deviation of the Gaussian (symmetric unimodal) curve. To test the effect of this parameter on estimated variable importance, a species with changing (high to low) levels of tolerance to four different covariates W1 to W4 (as seen in Fig.9) was generated.

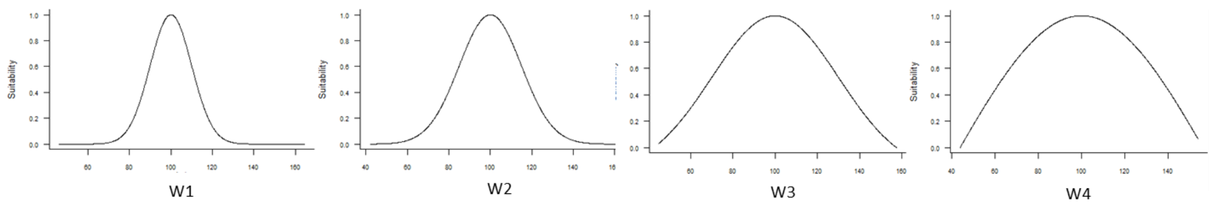


Figure 9: Four response curves with changing tolerance from low to high (left to right); used to generate one species and test relative variable importance estimates

2.1.4.3. Species generation to test the effect of the combination of tolerance and autocorrelation extremes

Since it is rare for a covariate in the environment to not be spatially autocorrelated, it is imperative that the relative effects of different autocorrelation and tolerance levels combined are analysed, in terms of bias in estimated variable importance. Therefore, the third and final of the sub-experiments requires the generation of a species whose characteristics include an additive combination of: a) High tolerance response on covariate with low autocorrelation; b) High tolerance response on covariate with high autocorrelation; c) Low tolerance response on covariate with high autocorrelation; and d) Low tolerance response on covariate with high autocorrelation. The details of this are shown in Fig. 10.

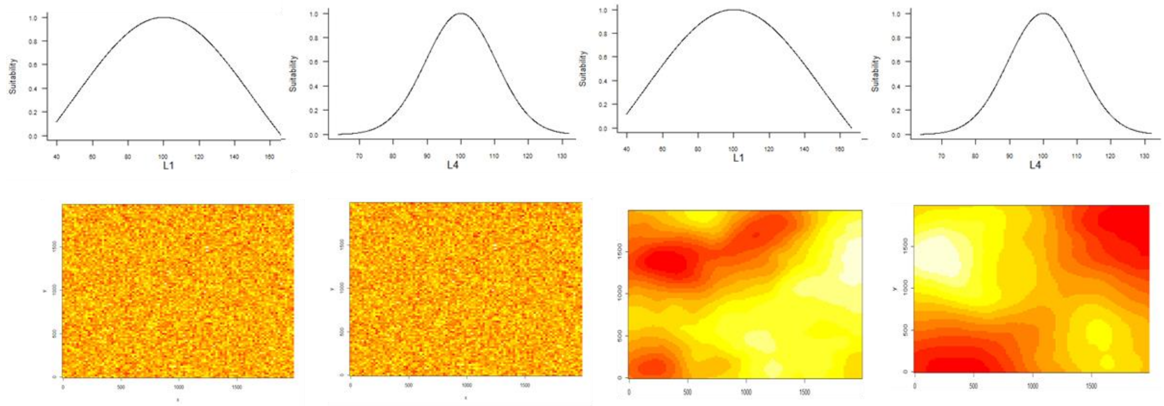


Figure 10: Response curves and covariates showing (from left to right): a) High tolerance on covariate with low autocorrelation; b) High tolerance on covariate with high autocorrelation; c) Low tolerance on covariate with high autocorrelation; and d) Low tolerance on covariate with high autocorrelation

After the generation of species (by defining different combinations of response curves) the individual responded covariates are then added up to get a habitat suitability, which is explained in the next section.

2.1.5. Simulating the habitat suitability

For each species scenario (and the 20 realisations of each) a habitat suitability for the corresponding species is defined as per the equation:

$$\text{H.S} = \sum_{i=1}^{4} w_i \; x_i$$

Where w_i is the predefined variable importance and is kept equal to reduce the complexity of the research (therefore all variables are defined to be equally important) and x_i is the habitat suitability as per each variable (i.e. the variable as transformed by the species response curve). The assumption here is that the equal importance of the responses to the variables implies an equal importance for the variables themselves as the responses are constant across variables. The variables are rescaled to a scale (0,1) therefore the final suitability map mimics a probability distribution map with each cell containing a certain probability of having the species or not. The suitability is represented as a probability map as shown in Fig. 11.

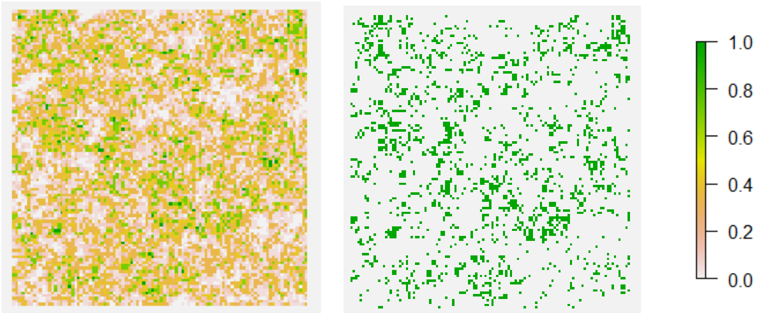


Figure 11: Example of probability map and corresponding presence-absence map of species using unimodal response function

To isolate the different parameters the most simplistic additive suitability was imagined and the additional effects of interactions and collinearity, that so often confuse real datasets, have been disregarded. Though in practical applications the effect of these is unavoidable, but for the sake of the present experiment, the covariates defining the species are independent and isolated.

The suitable habitat is then converted to a presence-absence map. The main parameters that define this conversion are alpha (slope), beta (threshold) of the logistic transformation (See Fig.12) and the prevalence of the species. Here again, to reduce the confounding factors, the prevalence is coerced to be limited around 50%.

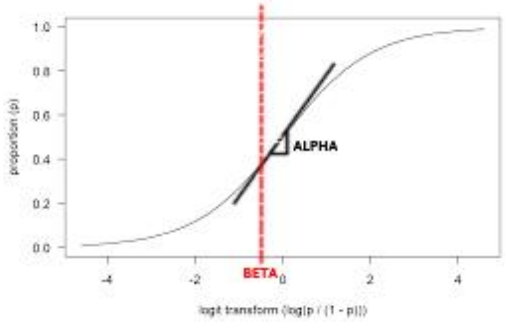


Figure 12: Parameters (alpha and beta) affecting the logistic conversion of suitability (probability) to presence-absence

Therefore, the model varies the beta factor (the threshold of identifying presence/absence) to maintain the prevalence at 50%, as the variables differ in their autocorrelation levels. Realistically, the prevalence of a species is defined by many factors including species characteristics, biotic factors like competition or dispersal limits and then it is likely that under normal conditions the prevalence of a species will be skewed beyond the 50% used here. For e.g. Fig 13 shows how the prevalence changes for the different species (the four colours) and the for different levels of relative autocorrelation when the threshold is fixed at 0.5.

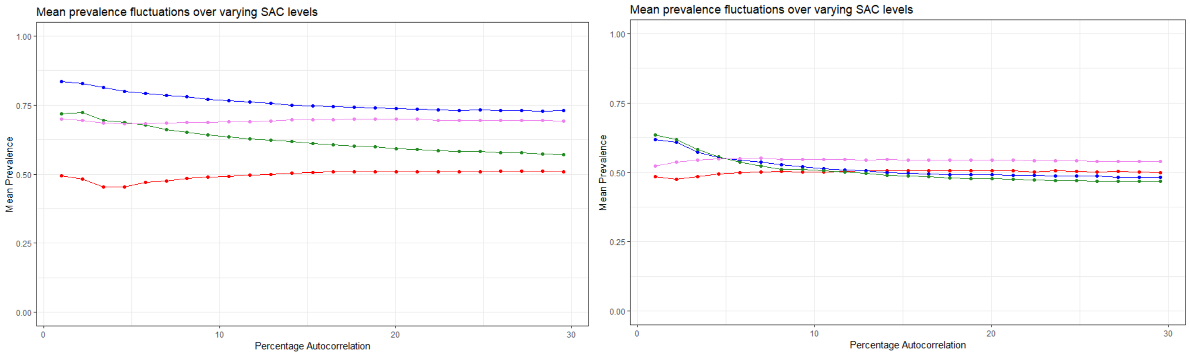


Figure 13: The effect of increasing autocorrelation in prevalence of species when the threshold (beta) = 0.5, for the 2 cases: a)V1,V2,V3 at 0% autocorrelation; b) V1,V2,V3 at 12.5% autocorrelation; Red line- species A, Violet – Species D, Green- Species C, Blue- Species B.

As can be seen from Fig. 13. the species with unimodal responses to landscapes tend to have skewed positive prevalence, that decreases with increasing autocorrelation whereas the ones with linear responses have a constant prevalence at 50%. This is because the optimum suitable areas in unimodal responses match with the normal distribution of the pixels of the input covariate, so the landscapes have a mean value that is ideal for the survival of the species. This could be a condition that is possible in reality; thus skewed prevalence is more the norm than an exception.

However, many studies have highlighted the effect of a skewed prevalence on model accuracy due to unbalanced proportions of presences and absences (Meynard and Quinn 2007, Sor et al. 2017a). Therefore, to allow for the experiment to isolate the effects of response curves and autocorrelation, the prevalence for all the species was limited at 50% (Sor et al. 2017a)(See Fig. 14)

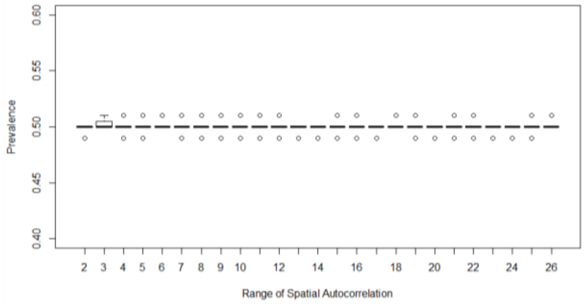


Figure 14: Bar plot showing the coerced prevalence values limiting itself to ~50%

When the prevalence of the species dataset is constrained, the beta values (threshold) vary across each realisation of species presence-absence. To aid a transparent reporting scheme of the input parameters the changing beta values were recorded (as shown in Fig.15).

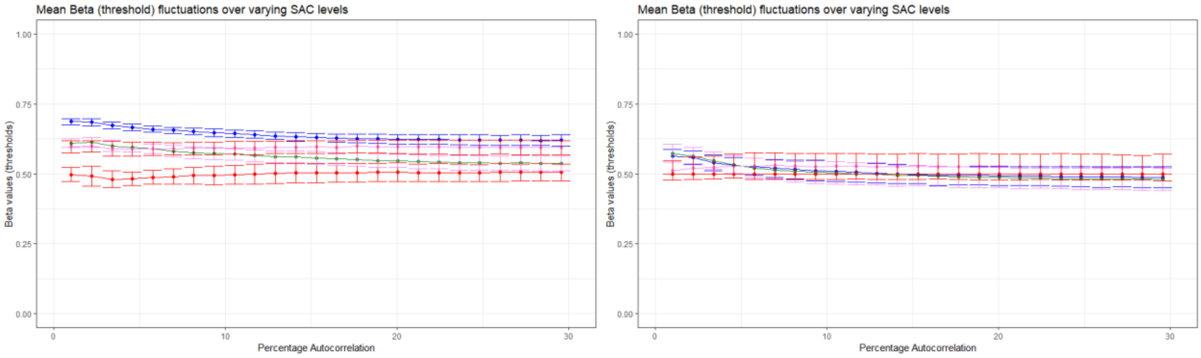


Figure 15: Plots showing variation of beta (threshold) values when the prevalence was kept constant at 50%; for the 2 cases: a)V1,V2,V3 at 0% autocorrelation; b) V1,V2,V3 at 12.5% autocorrelation; Red line- species A, Violet – Species D, Green- Species B, Blue- Species C

2.1.6. Sampling the area for presence/absence points

Random samples (200 points) are taken from each of the presence-absence to create a training dataset and an additional 200 points are sampled independently and randomly to make a test dataset. This implies a sampling density of 2%. 20 sets of random sample locations (one for each realisation) are initially taken and kept constant for the different levels of SAC and the different species. Test and training samples are also taken at 0.05% (50 points) and 3.5% (350 points) to help in analysing the effect of varying sampling densities.

Thus, section 2.1 details out the many steps involved in setting up the virtual environment, and an overview of the tools involved is given in Table 1.

Table 1: Overview of R packages used for setting up the virtual ecologist experiment (Leroy et al 2016, Nychka et al 2017)

Method	Parent' package	Required packages for method
Generate autocorrelated landscape	'fields'	'gstat' ; 'raster'; 'psych'
Generate virtual species	'virtualspecies'	'raster'

Generate samples	'sp'	null

2.2. Choice of models to run

The study incorporates eight modelling techniques, each run at their default functionalities as provided by the various modelling packages (see Table.2). They are: Generalised Linear models (logistic regression); Generalised Additive models; Boosted Regression Trees, Random Forest, Support Vector Machines, Random Forest, Artificial Neural Networks and a spatial version of a Generalised Linear Mixed Models.

Logistic regression (GLM)- A generalised approach for a binomial distribution using the logistic transformation of the 0,1 response. This is an effective method that provides results comparable to a simple linear regression(Guisan et al. 2002).

$$\ln(\frac{p}{1-p}) = b_0 + b_1 L_1 + b_2 L_2 + b_3 L_3 + b_4 L_3$$

Where L_1, L_2, L_3, L_4 are the four variables and b_1, b_2, b_3, b_4 are the slope coefficients of each variable, b_0 is the intercept; p is the probability of a presence. Square terms are incorporated wherever the variable response was unimodal. GLM were run using the stats package available in R.

To allow for better fit to the non-linear responses the **Generalised Additive Models (GAM)** model is also used. This model basically fits a non-linear (smooth) function on each variable before linearly adding each term; therefore, adding flexibility to the model (Guisan et al. 2002). The GAM was run through the 'biomod2' package, using default settings (smoothing splines and no interactions between covariates).

$$\ln(\frac{p}{1-p}) = b_0 + f_1(L_1) + f_2(L_2) + f_3(L_3) + f_4(L_3)$$

For the spatial model, the **Generalised Linear Mixed Model (GLMM)** is used that accounts for the spatial component as a spatial random effect. For the logistic version of the LMM, only the GLMMPQL (a penalised quasi-likelihood to accommodate binomial data) seemed the applicable method as referred from Dormann et al., 2007. This seemed a good choice since most other spatial methods (autoregressive methods, SAR, CAR) etc require lattice data. Also as per Betts et al. (2007), the GLMM is better for computing inferences on parameter estimations, whereas the other spatial models might produce spurious estimates, though their predictive performance can be higher. The basic form of the spatial GLMM is:

$$\mathrm{y} = \mathrm{X}\beta + \mathrm{Z}u + \varepsilon$$

Where y is the logit link transformed response, X is the coefficients of the fixed effects (landscape variables), u is the spatial random effect modelled as an exponential curve on the basis of x,y location details of each point and Z is the estimated coefficient for it, ε is the residual error. Therefore, the model only has fixed effects (for the four covariates) and a spatial random effect.

For the machine learning methods, a series of models frequently used in Species Distribution Modelling has been used including **Boosted Regression Trees** (BRT), **Random Forest** (RF), **Maximum Entropy** (MaxEnt), **Support Vector Machines** (SVM) and **Artificial Neural Networks** (ANN). These models have been used simply because of the popular use of them in Variable Importance investigations, and all runs are at default parameters.

Boosted Regression Trees (BRT)- This algorithm is known in ecology for being able to model complex non-linear functions. It uses an additive (boosting) mechanism to add more flexibility to the normal regression trees method (which uses binary recursive splits to estimate the class mean of parameters) (J. Elith et al. 2008, Naimi 2015). The default model settings in the SDM package were used, which are 'n.trees' initial =100; 'bag.fraction'=0.5; 'learning rate' of 0.1; and total no. of trees = 1000.

Random Forest (RF)- This algorithm uses bootstrap samples from the dataset to fit a no, of regression trees. These trees are fitted on the samples and the estimates used to define the complexity and variable importance internally in the model (Breiman 2001, Naimi 2015). The default settings used here are 'no. of trees' = 1000, no. of variables chosen at each split = 1. Due to the added stochasticity in the model, at large number of trees, the problems of overfitting are typically less giving more accurate results.

Maximum Entropy (Maxent)- A presence only model (uses pseudo-absences to compensate) that tries to minimise a 'gain' function (similar to deviance) to maximise the entropy (uniform distribution of uncertainty in geographical space), along with accommodation of constraints which force the output probability distributions to be similar to the mean values of the input covariate values at sampled presence points (Hastie et al. 2010, Phillips et al. 2006, Phillips and Schapire 2004). The default setting used here is 'no. of iterations' = 500, default prevalence of 0.5 and all possibility of features (auto features). No pseudo-absence points were inputted; therefore, background points were taken from the sample dataset itself.

Support Vector Machines (SVM)- The method identifies critical elements (support vectors) with which it defines an optimum hyperplane that maximises the margin (or separation 'street' between the classes). This method can be used for linear separations as well as non-linear for which 'kernels' (to map the non-linear function to a linear output) are used (Drake et al. 2006). To avoid overfitting, a cost constraint is also defined. The default settings used for this model was epsilon=0.1, cost c=1, Gaussian radial basis kernel with hyperparameter sigma = 0.22.

Artificial Neural Networks (ANN)- Neural networks are highly flexible algorithms that inform correlative relationships between variables as an output of multiple hidden layer interactions (each a composition of weighted nodes). This is also called a feedforward network. These are known to be very useful in determining complex distributions of data but are known to overfit the datasets also if the structure of the network is not properly tuned (Rocha et al. 2017, Sor et al. 2017a, Thuiller 2003). Default settings used: no. of cross-validations- 5; maximum no. of iterations = 200; size and decay functions are optimised by the cross-validations based on model AUC.

This study does not go into an exhaustive understanding of how different machine learning methods work, but simply provides a template of results at reported default settings, that can be used practically since the main aim of the paper is not a comparison of the individual models.

Table 2. details out the packages used in R to implement all the models. The eight models are run on each of the 20 realisations of each for the four species, and on each of the 25 levels of changing autocorrelation of variable four (V4).

Table 2: List of R packages used to implement the models (Naimi and Araújo 2016, R Core team 2017, Ripley and Venables 2002, Thuiller et al. 2009)

Method	Parent' package	Required packages for method
GLM	'stats'	NULL
GLMM-PQL	'MASS'	NULL
GAM	'biomod2'	'gam'; 'mgcv'
BRT	'sdm'	'gbm'
RF	'sdm'	'randomForest'
MaxEnt	'sdm'	'maxent.jar'
SVM	'sdm'	'kernlab'
ANN	'biomod2'	'nnet'

2.3. Estimating model independent Variable Importance

Model-independent methods of variable importance are used to compute the estimated variable importance as per different models (Duque-Lazo 2013, Naimi and Araújo 2016). This method of variable importance calculations basically randomises one of the variables (V1, V2, V3 or V4) and makes predictions on the 'tampered' dataset. These predictions are checked for correlation ("Pearsons") with the predictions from the untampered datasets. Since the higher the correlation, the lower should be the importance of the variable, as tampering with its values did not necessarily create much difference in the predictions, 1- the correlation coefficient (r) is calculated and considered as a measure of variable importance. Therefore, it is a method to check the sensitivity of a model to each of its variables and since it can be computed independently of model runs it is called a model-independent variable importance assessment. To preserve the autocorrelation structure in the randomised version of the covariate, the method used in this study was to simply swap each of the 20 realisations of the covariate (section 2.1.1) with each other before making the 'tampered' predictions.

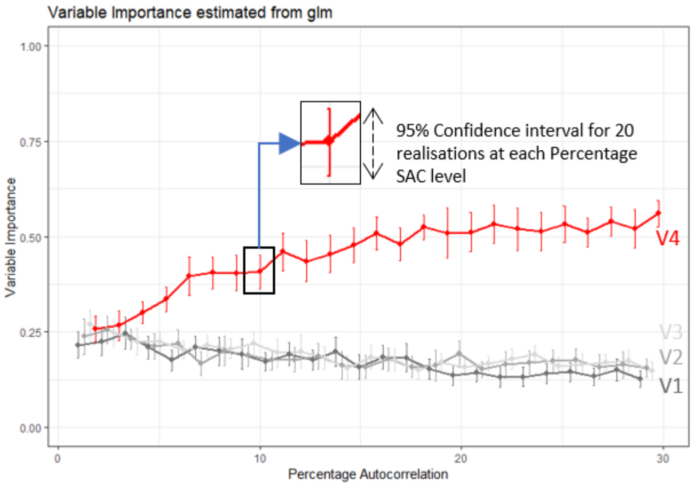


Figure 16: Variable Importance graphs with 95% confidence intervals for covariates V1, V2, V3 and V4 as estimated from a single model (GLM); the red line represents covariate with varying levels of autocorrelation; the grey lines represent the covariates with fixed autocorrelation

Variable importance estimations are computed for each run of the different models for different species and autocorrelation levels at two baseline scenarios (see section 2.4). Graphs are computed with data at 95% confidence intervals (from 20 realisations) for each covariate variable importance and each model. These graphs, presented as line graphs, can help show the changes in variable importance with respect to different levels of SAC (see Fig.16).

Since the variable importance was assumed to be equal initially (equally weighted in the suitability equation), it is expected that an optimum model (not sensitive to SAC) would still report equal variable importance under the influence of autocorrelation and for the different species and thus be robust under the effects of SAC. To ensure that no hidden aspect was defining the variable importance, test runs were conducted on covariates with similar responses and the same levels of autocorrelation. The results from these tests showed similar variable importance estimated for all the covariates for species A and B. Thus no hidden parameter affected the variable importance estimations. This can also be seen at the 0% SAC level of Fig. 16, where on the left end V1, V2, V3, V4 are significantly similar. ANOVA tests were done on the data to validate the differences found. These results cater to the first and second objective of the study.

2.4. The two scenarios- Background SAC at 0% and Background SAC at 12.5%

Since this study seeks to find an answer to the effect of different parameters on the relative variable importance, it was important to set two baseline scenarios to understand the effects of relative SAC between the covariates. The first scenario incorporates a background SAC at 0%, which implies that the three background variables are spatially uncorrelated. The second scenario incorporates a background SAC at 12.5%, which is almost halfway across the total range (30%), which implies that the three background variables are autocorrelated at a theta (range) that is 12.5% of the total extent. The two scenarios incorporate relative SAC percentages that vary from 0% to 30% (first scenario) and (-) 12.5% to (+) 17.5% (second scenario), where 0% relative SAC implies the autocorrelation in the covariates are the same. Thus, the second scenario assesses the model estimations at a smaller relative SAC.

To allow for better interpretations of the results, the Moran's I (calculated using the R package 'ape' by E. and K. 2018) of each variable (background and for each changing SAC level) is also reported, where 0% SAC range implies an average Moran's I of 0, 12.5% SAC implies 0.265 value for Moran's I and 30% SAC implies a Moran's I of 0.34. This can be useful in inferring about the relative spatial structure in the covariates where the relative difference in percentage SAC can be less practical (in terms of different shapes of spatial relationships etc). Since the levels of spatial autocorrelation were structured using the percentage SAC scale-free statistic, the Moran's I is not a directly controlled parameter, and rather is computed from the sampled covariates. Therefore, it can only act as an additional measure to understand the relative spatial structure of the covariates, and its accuracy can be hampered by the sampling density and the locations of the samples.

2.5. Measures to assess the Robustness of a model: Area between the Variable Importance curves

Robustness in the presence of SAC or varying response curve geometry for a model can be defined as the ability of the models to estimate equal relative variable importance consistently, even in the presence of the changing parameters. To assess this robustness of the variable importance estimations, for different levels of autocorrelation and different types of species, a useful metric would be to assess the area between the estimated importance curves (see Fig.17). The metric will calculate the area between the autocorrelated curve (variable importance of V4) with respect to V1, V2 and V3 individually for each of the 20 iterations. The final area estimate will be an average of the three areas, thus calculating a dataset of 20 average values for each model and each species. Therefore, the model that computes the lowest areas with least variability (within the 20 iterations) will be the most robust, i.e. will account for the least average bias in variable importance in the presence of autocorrelation and different species types.

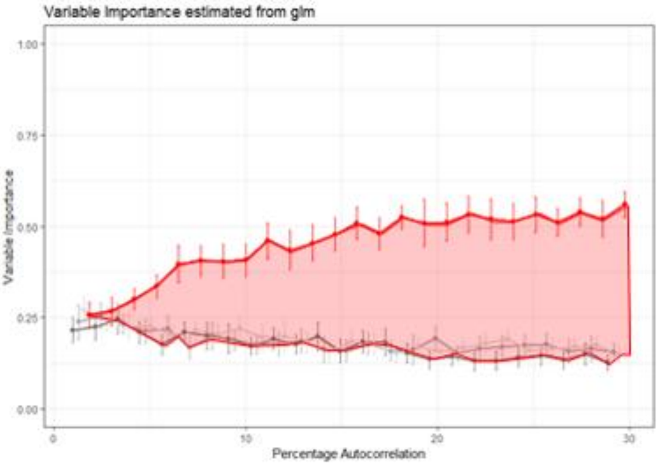


Figure 17: Schematic for the area between the curves of variable importance estimations; the light red region is the average area between the red line (V4) and each of the three grey lines (V1, V2, V3)

2.6. Model Accuracy measures: AUC, Kappa and Residual SAC

The area under the Receiver Operating Characteristic (ROC) curve is a popularly used model accuracy estimator (Allouche et al. 2006, Luoto et al. 2005, Meynard and Quinn 2007, Thuiller 2003). The ROC curve plots the graph between the sensitivity (true positives rate) and 1 – specificity (the false positive rate), for all values of possible thresholds. Therefore the AUC is regarded as a threshold independent measure, that mainly assesses the discriminatory power of the model in classifying presences and absences. AUC can be a good comparative tool to assess the different models in this study since the sample size was constant and the prevalence is maintained at an impartial 50% (Hanberry and He 2013). The AUC was computed directly from the model runs within the 'biomod2' and 'sdm' packages (Naimi and Araújo 2016, Thuiller et al 2009).

Besides the AUC another prominent statistic used in ecology to assess the accuracy of models is the Cohens Kappa statistic (Hirzel and Guisan 2002, Meynard and Quinn 2007, Naimi 2015, Sor et al. 2017a). The kappa weights the model accuracy with the probability of getting accurate prediction by chance. One of the negatives in using this metric is its reliance on a threshold. Since the beta (threshold) values were identified early on in the methodology, they can be used to estimate the 'true' kappa of the model. This measure can be a good complementary accuracy metric alongside AUC as a good estimate of both omission and commission error reported in one simple metric. Allouche et al., 2006, mentions the dependence of kappa on prevalence which does not pose an issue as the prevalence has been controlled to 50% in this study. The Kappa was computed using the 'PresenceAbsence' package in R (Freeman and Moisen 2008).

Since spatial autocorrelation is one of the main parameters of this study, it is important to check whether the models can, in fact, account for the autocorrelation in the covariates. Therefore, the residual SAC is calculated by running the Moran's I on the residuals of the models (Bini et al. 2009, Diniz-Filho et al. 2003, Hawkins et al. 2007, Naimi 2015).

3. RESULTS

3.1. Effect of autocorrelation on Variable importance estimates

The plots of the relative variable importance (y-axis) for the different ranges of autocorrelation (x-axis) for each of the four species types are shown in Appendices A1-4. Appendix A1, A2 represent the scenarios when the background SAC (SAC of variables V1, V2, V3) are at 0% while appendix A3-A4 represent the scenarios when the background SAC is 12.5% (less than half the highest autocorrelation range). Section 2.1 shows the results of Species A and B whereas Section 2.2 shows the results of Species C and D since the objectives of each are different. The most general/important results have been shown and addressed to, and the rest have been attached in Appendices. All variables have been defined as of equal importance irrespective of autocorrelation and species type (as mentioned in section 2.1.5).

3.1.1. Species A- All linear responses

The results from scenario 1 (background SAC at 0%) are shown first, with a comparison between GLM and RF as examples signifying generalised linear models and machine learning methods respectively. As can be seen Fig.18a&b, the effect of increasing autocorrelation exaggerates the variable importance of the autocorrelated variable drastically. The ANOVA measures of these graphs are shown in Appendix B1. Clearly, beyond two steps of autocorrelation (0-1.25 % of the total extent), the variable importance estimates from all the models (except ANN) become sensitive to autocorrelation showing a linear increase as the percentage SAC increases. In the case of ANN, the variable importance are significantly different only after a relative higher SAC of 3.75%. Nonetheless, the results imply that the models are highly sensitive to spatial autocorrelation and that a difference in Moran's I of 0.03 (0.13 in case of ANN) units between the covariates can inflate the importance of the covariate with higher SAC.

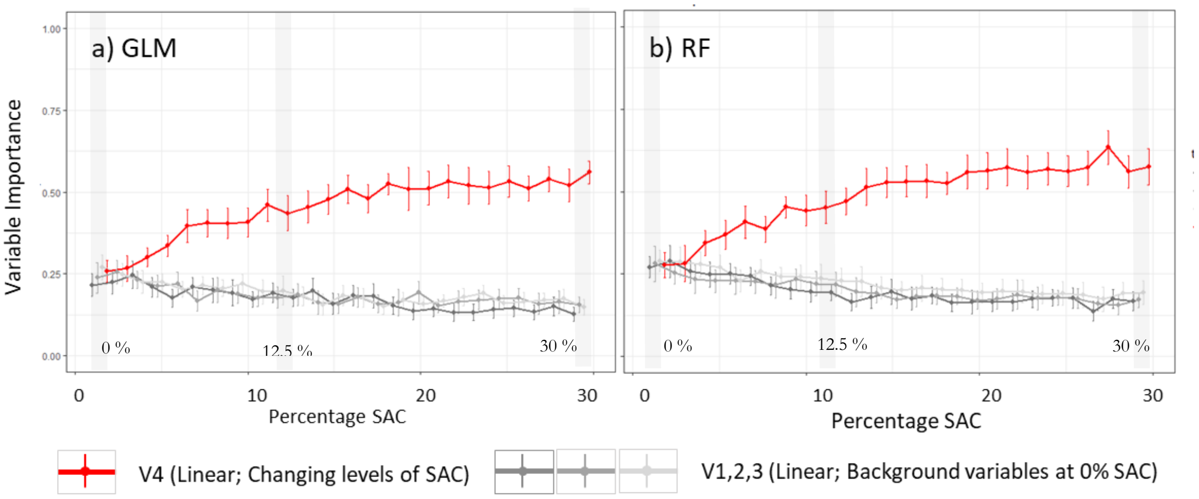


Figure 18: Variable Importance estimates from GLM and RF for Species A (at 95% confidence interval of 20 realisation values) and for Scenario 1 (background SAC 0%); The 3 wider grey vertical bands show areas of relative SAC of 0%,12.5%,30%: see numbers within the graphs (See Appendix A1 for the graphs from the other models)

In the second scenario (Fig. 19), the ANOVA values (see Appendix B3) show a significant difference of means when the V4 (red line in Figure) has an autocorrelation roughly less than 6.25% (relative decrease of 6.25 units in percentage SAC also) and greater than 25% (relative increase of 12.5 units in percentage SAC). In terms of Morans I, a difference of 0.06 in the lower spectrum and 0.08 units in the higher spectrum is enough to inflate the importance of the broader scale (higher sac) variable. These patterns are seen in all other models (machine learning and generalised linear-spatial and non-spatial) (see Appendix A3, B3), though the patterns in BRT and ANN are much more erratic with wider confidence intervals. Few cases of significant differences were also seen in the mid ranges for GLM, GLMM and GAM, though they were insignificant at 99% confidence interval.

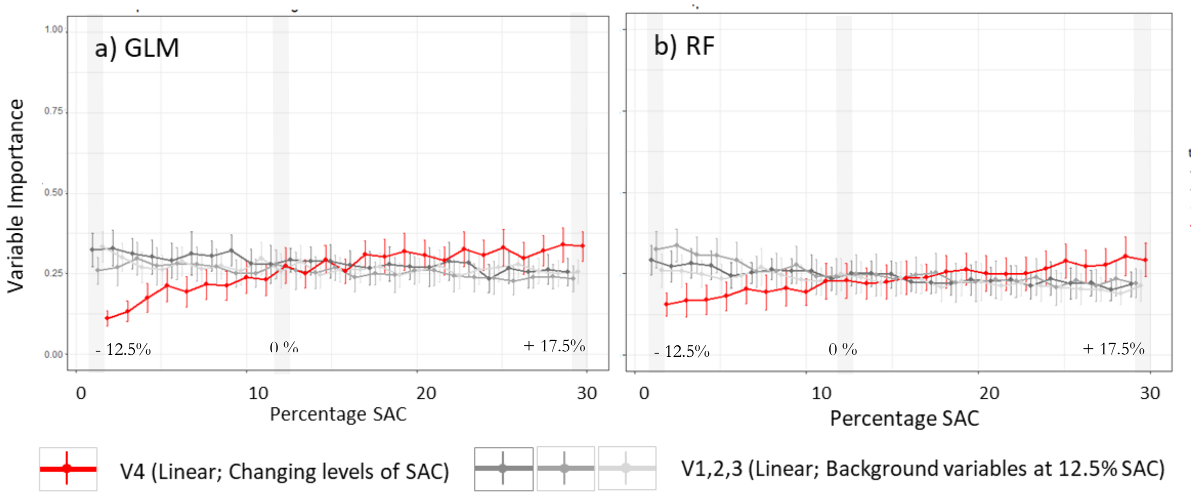


Figure 19: Variable Importance estimates from GLM and RF for Species A (at 95% confidence interval of 20 realisation values) and for Scenario 2 (background SAC 12.5%); The 3 wider grey vertical bands show areas of relative SAC of (-)12.5%,0%,(+)17.5% (from left to right); (See Appendix A3 for the graphs from the other models)

Therefore for species A which responds to all the variables linearly, autocorrelation affects the estimates significantly beyond certain relative SAC thresholds.

3.1.2. Species B- All Unimodal Responses

In cases where the species responds unimodally (i.e. it has an optimum environmental range) to the covariates, the effect of autocorrelation was not as prominent as seen in section 3.1.1, though still some amount of inflation/bias in estimates was seen at a lower magnitude.

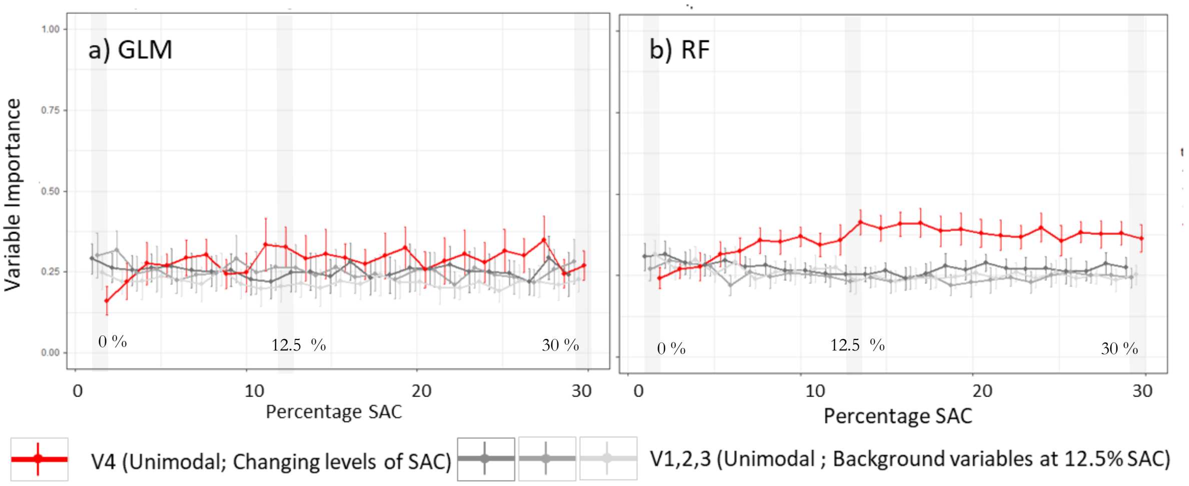


Figure 20: Variable Importance estimates from GLM & RF for Species B (at 95% confidence interval of 20 realisation values) and for Scenario 1 (background SAC 0%); The 3 wider grey vertical bands show areas of relative SAC of 0%,12.5%,30% (from left to right); (See Appendix A1 for the graphs from the other models)

As can be seen in Fig.20 a) the changes in autocorrelation levels do not create any significant differences in variable importance estimations from GLM's based on 95% ANOVA tests. However, in the case of RF (see Fig. 20 b) and other machine learning algorithms including GAM, except SVM and ANN, the estimated variable importance for the autocorrelated variable was significantly higher than the rest background variables beyond a threshold range of 5% SAC range, i.e. (relative increase in Moran's I of 0.13) is enough to inflate the importance of the higher autocorrelated one. (see Fig.20b and Appendix B1).

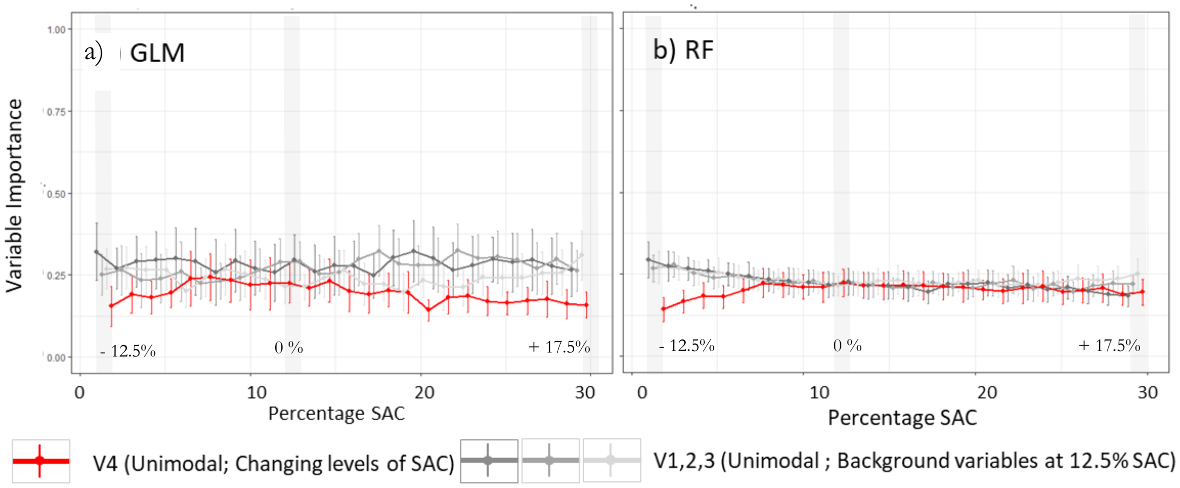


Figure 21: Variable Importance estimates from GLM & RF for Species B (at 95% confidence interval of 20 realisation values) and for Scenario 2; The 3 wider grey vertical bands show areas of relative SAC of (-)12.5%,0%,(+)17.5% (from left to right); (See Appendix A3 for the graphs from the other models)

However, in the case of a scenario 2 (see Fig. 21 a) where the background autocorrelation is at 12.5 %, the GLM performed poorly in estimating variable importance, showing a significant decrease in variable importance as autocorrelation increased. This is an unexpected result that implies that in cases where the background variables are significantly autocorrelated with percentage autocorrelation 12.5%, a relative increase >12.5 units of percentage SAC (>0.08 units of Moran's I) can decrease the variable importance from GLMs'. Whereas for the machine learning algorithms performed better, as seen in Fig.21b (example showing RF only, for the rest see Appendix A3), showing no significant differences beyond SAC ~5%, i.e. a 7.5 unit decrease in percentage SAC deflates the variable importance, but a corresponding 17.5% increase has no implications. This variability in thresholds seem counter-intuitive, but arises out of the mismatch between increasing percentage autocorrelation and corresponding Morans I. Therefore, the right inference is that autocorrelation does not bias model estimations of importance for unimodal responses if the relative spatial autocorrelation between them is within 0.09 units of Moran's I (both -7.5% and 17.5% SAC has Morans I <0.09), which is more comparable to the 0.13 units threshold found out in scenario 1.

3.2. Effect of the geometry of response curves on estimates of variable importance

Boxplots of varying variable importance (x-axis) for covariates changing gradually from linear to unimodal response (y-axis) are shown in Fig.22 (for details check section 2.1.4.1). When comparing the variable importance of a response curve that gradually changes from linear to unimodal, against a linear response curve, the estimated variable importance increased beyond L4 which represents 40% of the unimodal response curve (2.1.4.1) and despite a slight dip at L8,9 which is not pronounced enough to make an inference from, L10 (complete unimodal response) was invariably more important than L1 (linear) for all the models (See Appendix C3 for the rest of the methods)

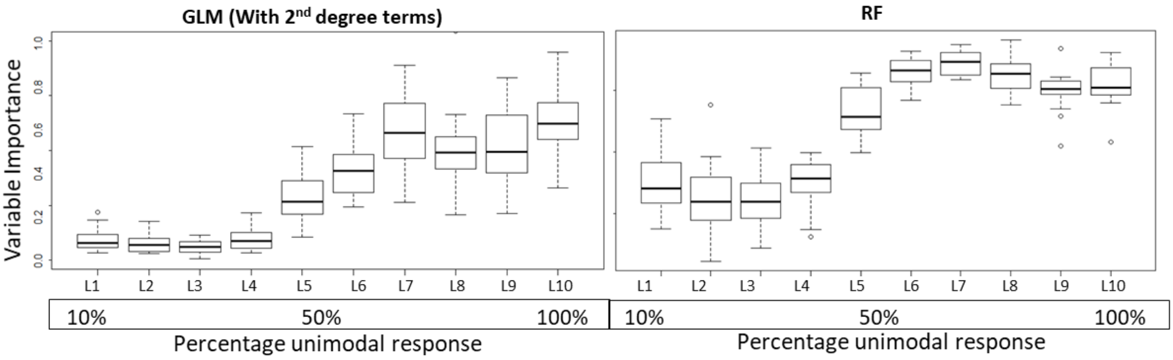


Figure 23: Boxplots showing variable importance estimations from GLM and RF showing the biased increase in estimations as the variable changes gradually from linear to unimodal

It can also be seen from Fig. 22; the GLM does not bias towards the unimodal as much as the machine learning methods (100% unimodal gains importance up to 0.8 to 1.0 in machine learning methods, whereas the GLM averages at 0.6). The relative increase in the magnitude of importance remains within 0.5~0.6 units for the unimodal over the linear for both the models.

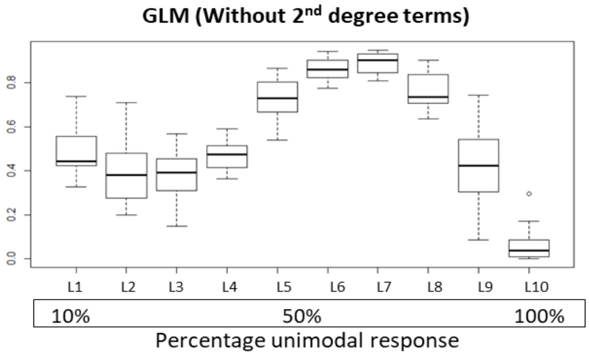


Figure 22: Boxplots showing variable importance estimations from GLM (without second degree term) showing the biased increase in estimations as the variable changes gradually from linear to unimodal

Fig. 23 shows the case of using a GLM without a second-degree term to analyse the variable importance bias for response curve geometry. As it can be seen, up to 70% of the unimodal response (which forms a slight 'S' shape; see section 2.1.4.1) the GLM shows increasing variable importance, after which the more unimodal the geometry, the less important the covariate.

3.2.1. Species C – Combined unimodal and linear response, former iterated with increasing spatial autocorrelation

In cases where the species had a combination of unimodal and linear responses to different covariates, the ANOVA was rarely insignificant (see appendix A2, A4), implying that all the variable importance estimates were dissimilar. For the first scenario (background SAC 0%; see Fig. 24), both the generalised linear models and the machine learning methods predict a biased variable importance for the unimodal responses across the entire spectrum of autocorrelation. The estimations for the linearly responded variable are consistently low. The higher the autocorrelation in the unimodal response the more inflated its importance as the relative SAC is higher than previously investigated Moran's I of 0.13 (see Appendix A2 for the rest of the methods). However, the higher magnitude of the increase in variable importance for the unimodal is unlike what was seen before, and there seems to be a trade-off between the two unimodal responses, with one increasing as the other decreases. This pattern was not found when all four responses were unimodal (see Fig. 20).

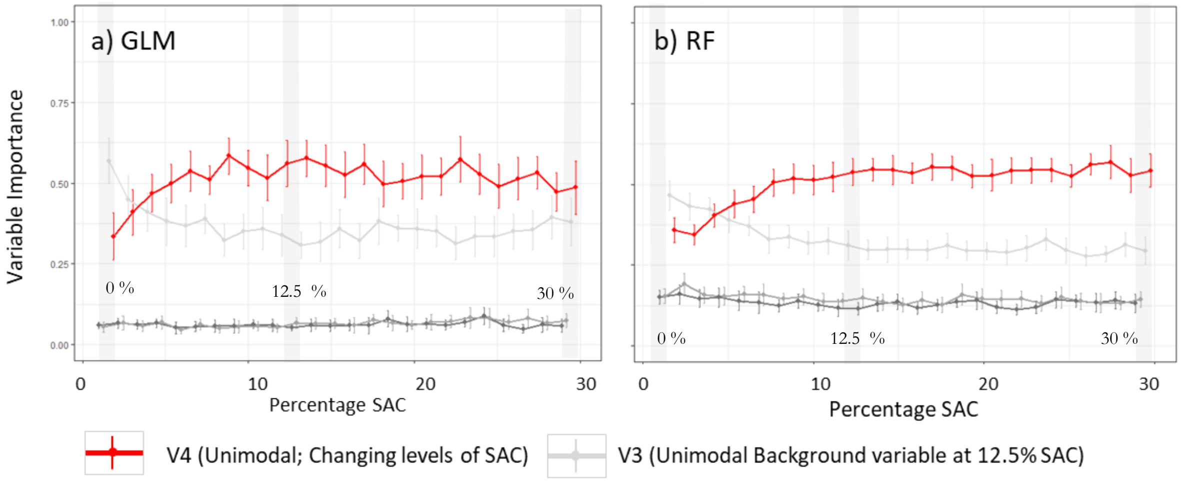


Figure 25: Variable Importance estimates from GLM & RF for Species C (at 95% confidence interval of 20 realisation values) and for Scenario 1 (background SAC 0 %); The 3 wider grey vertical bands show areas of relative SAC of (-)12.5%,0%,(+)17.5% (from left to right); (See Appendix A2 for the graphs from the other models)

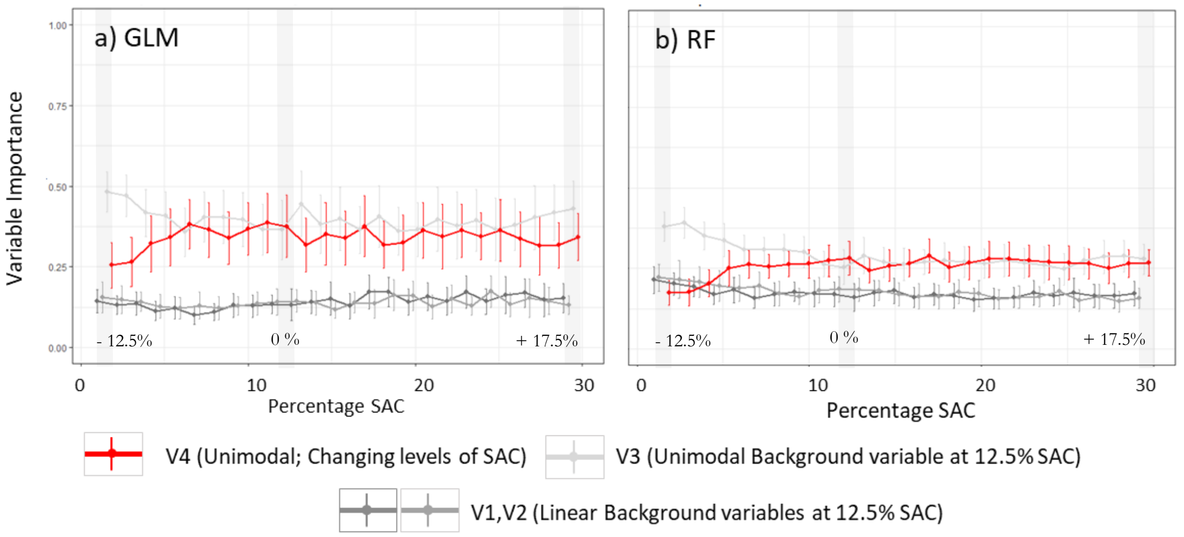


Figure 25: Variable Importance estimates from GLM & RF for Species C (at 95% confidence interval of 20 realisation values) and for Scenario 2 (background SAC 12.5%); The 3 wider grey vertical bands show areas of relative SAC of 0%,12.5%,30% (from left to right); (See Appendix A4 for the graphs from the other models)

For the second scenario also the same pattern applies, as can be seen from Fig.25. A few other inferences are worth considering, i.e. beyond 5% SAC (a relative SAC of 7.5 units) the red curve levels off implying that within this certain threshold of relative SAC the unimodal responses are robust. This peak also coincides with the grey unimodally responded covariate at background SAC of 12.5% in Fig. 25. This agrees with the results that were seen in section 2.1.2 for machine learning methods. Another important inference is that for machine learning methods a unimodal response with low SAC (<5%) and a linear one with higher SAC 12.5% (implying a relative SAC of ≤7.5 unit difference or ~≤0.09 units of Moran's I) are estimated at similar levels of variable importance (see coinciding red and dark grey lines in Fig 25b and check Appendix A4 for the other machine learning models).

Also, it can be seen from both Fig. 24 and 25 that the machine learning methods are less biased (smaller distance between the curves) than the GLM, GLMM and GAM.

3.2.2. Species D– Combined unimodal and linear response, latter iterated with increasing spatial autocorrelation

Similar to the results in section 3.2.1, none of the relative variable importance were similar, showing a permanent positive bias in the estimations towards the unimodal response (all the Anova results were significant) (Appendix B2, B4).

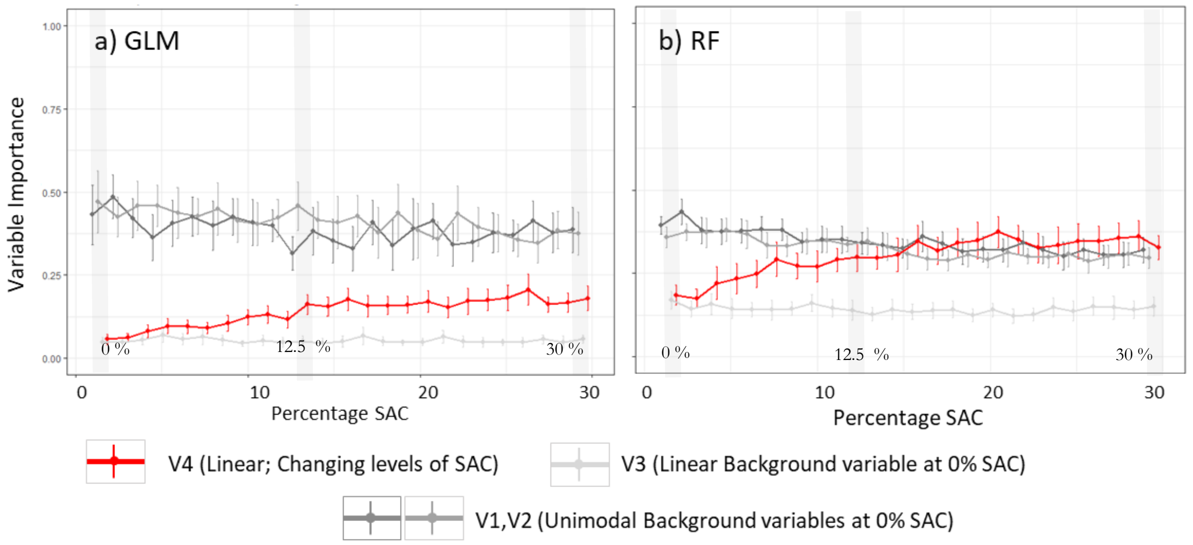


Figure 27: Variable Importance estimates from GLM & RF for Species D (at 95% confidence interval of 20 realisation values) and for Scenario 2 (background SAC 12.5%); The 3 wider grey vertical bands show areas of relative SAC of (-)12.5%,0%,(+)17.5% (from left to right); (See Appendix A2 for the graphs from the other models)

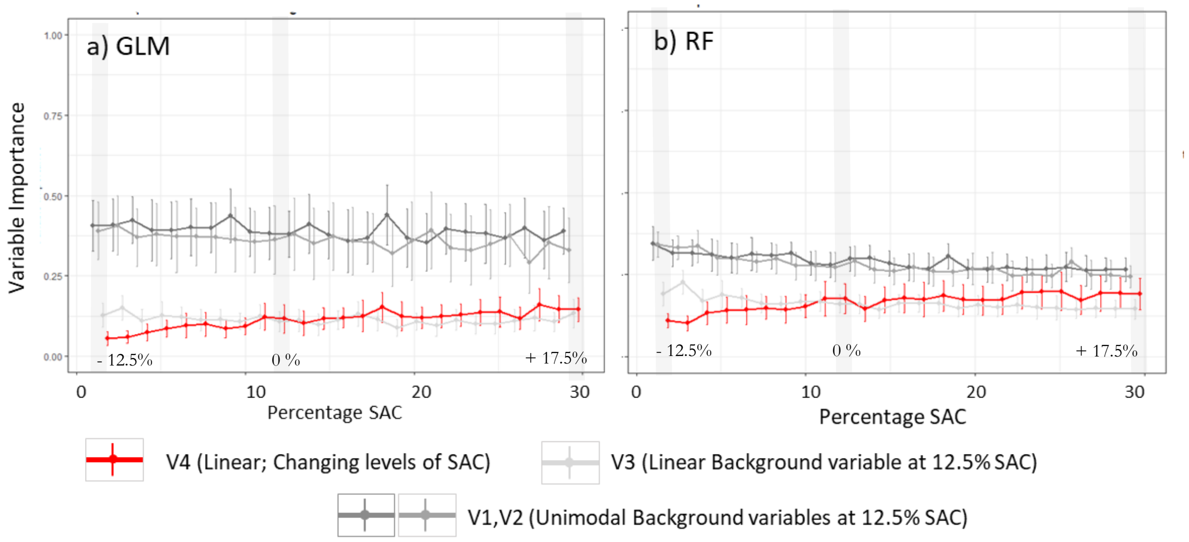


Figure 26: Variable Importance estimates from GLM & RF for Species D (at 95% confidence interval of 20 realisation values) and for Scenario 1 (background SAC 0%);The 3 wider grey vertical bands show areas of relative SAC of 0%,12.5%,30% (from left to right); (See Appendix A4 for the graphs from the other models)

Fig 26,27 shows the bias in variable importance for the different responses, with the importance of the linear variable increasing as the % SAC also increased. For the GLM's the differences between the curves are high for both 0% and 12.5% baseline scenario, whereas for the machine learning methods (especially RF), at maximum autocorrelation, the linearly responded variable is almost similar in importance estimates, if not more, than the corresponding background unimodally responded variables (see Fig. 26 b). Fig. 27 b shows a similar pattern, but since the background unimodal responses now have a SAC at 12.5%, even the maximum SAC of the linear response (30%) does not quite reach up to the unimodal response variables. Therefore a relative increase in percentage SAC by ten units (or a Moran's I of 0.25), in a linearly responded covariate over a unimodal one estimates equal variable importance from all machine learning methods (except BRT; see Appendix A4).

3.2.3. Effect of changing width (environmental tolerance) of response curves

In the previous experiments, the width of the unimodal curves (defined by the standard deviation) was kept constant throughout as the percentage SAC increased because in initial trials the effect of different standard deviations for the unimodally responded covariates sometimes outweighed the SAC effects to give confusing results. Therefore, to illustrate the effect of the width of unimodal response isolated (other parameters are constant) this sub-experiment was conducted. As can be seen from Fig.28, both the GLM and the RF shows a monotonic decrease in variable importance as the response becomes wider. For the results from the rest of the models see Appendix C2.

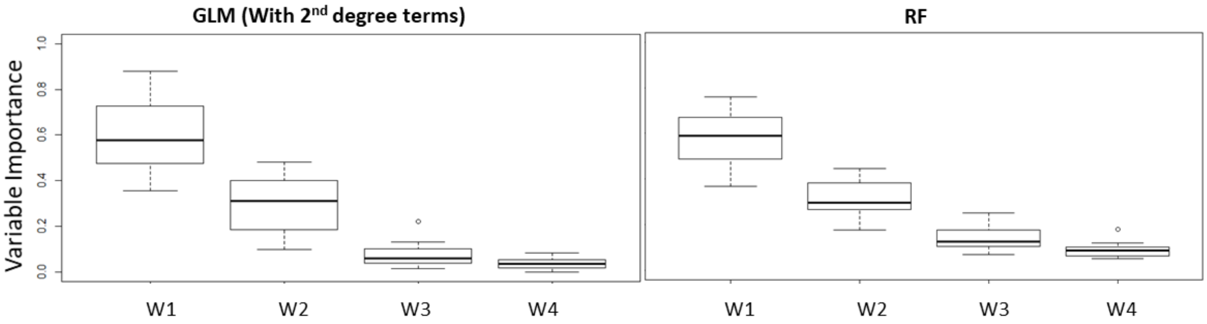


Figure 28: Boxplots of estimated variable importance from a) GLM and b) RF; where the width of the response curves of the variables changes as W1>W2>W3>W4 as per the response curve scheme shown in section 2.1.4.2

3.2.4. The combined effect of changing widths (environmental tolerance) and SAC

To analyse the combined effect of changing the width of response curves and SAC (therefore species characteristics and predictor characteristics) and to make a comparison of which effect outweighs the other, another sub-experiment with four landscapes as defined in section 2.1.4.3 was run across all the models.

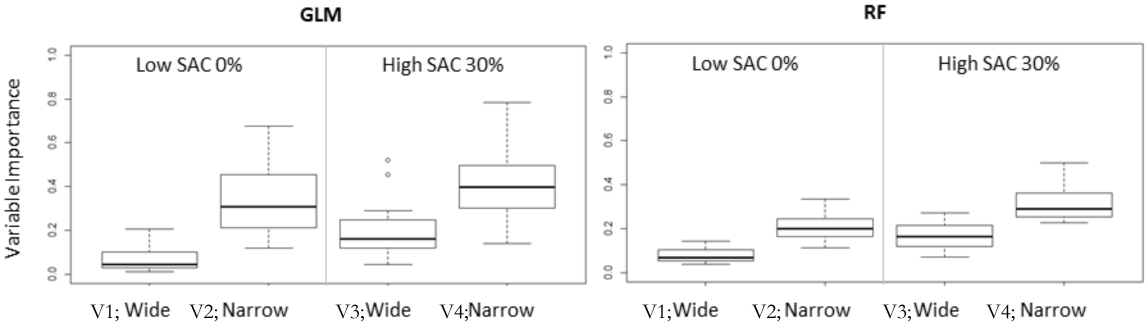


Figure 29: Boxplots showing estimations of variable importance from a) GLM, b) RF for four covariates V1:V4; where extremes of response curve widths and covariate SAC levels were incorporated.

As the Fig.29 shows, the models, as expected, tend to have a lower variable importance estimation for the low SAC covariate with a wide response curve and estimate high variable importance for high SAC covariate with a narrow response curve. Yet observably (though not as distinct in machine learning methods) the high precision overrides the high SAC (V2>V3) marginally in all the models; except ANN and BRT (See Appendix C3). This implies that covariates with lower autocorrelation can have higher importance if species response to it is narrow. BRT estimations have high variances and therefore, do not predict such a pattern. The ANN does not differentiate precision and autocorrelation (V1<V2=V3=V4).

3.3. Effect of changing the sampling density on autocorrelated variables with different responses

To test the effect of changing sampling densities within the same geographic extents, the study ran all the models for 3 levels of sampling densities (from 0.5% to 3.5%) in scenario 2 (background SAC 12.5%), but only for species A and B, as the relative effect of sampling density can be better inferred from if the response curves are kept constant (either all unimodal or all linear). Fig. 30 shows the three sampling densities as normal (3.5%), dashed (2%) and dotted (0.5%). Statistically, changing the sampling size from 200 (2%) to 50 (0.5%) points did not have any significant effect on the estimations of autocorrelation in term of bias correction. Though the points at which the means of the variable importance curves visually intersected were shifted (see Fig.30 where the red dashed line is shifted to the left relative to the blue lines), yet there were no statistically significant (ANOVA) differences.

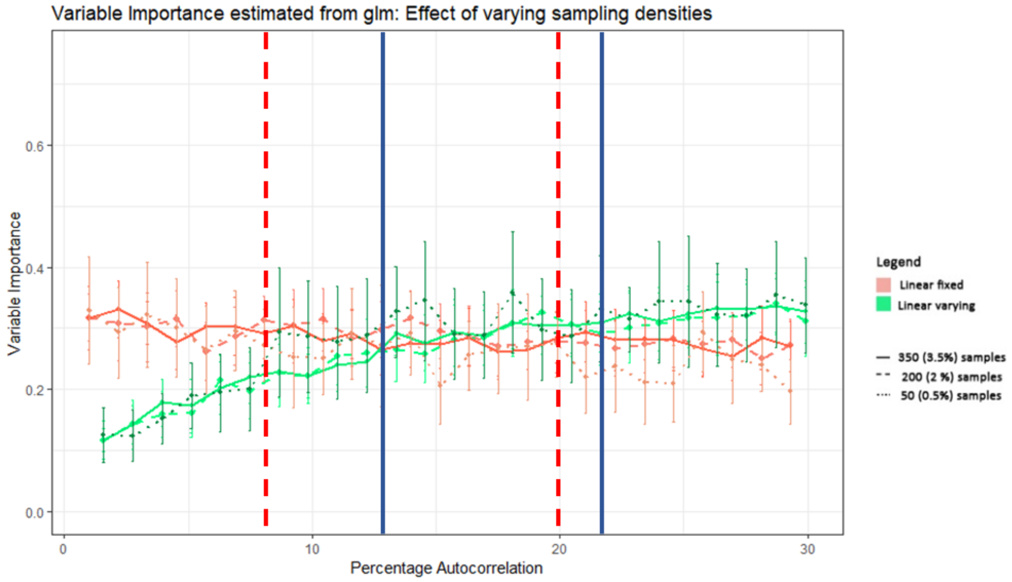


Figure 30: Variable importance estimates of species A (at 95% confidence interval) from GLM for only two of the four variables for scenario 2 (background variables at 12.5%) ; Red vertical dashed line- intersection of means for 0.5% sampling density; Blue vertical line- intersection of means for 0.2% & 3.5% sampling density.

Similarly increasing the sample size to 350 (3.5%) did not have any significant effect. The only observable difference as can be seen in the Fig.30 is the variance in the mean estimations of the variable importance decrease as the sampling density increases, thus decreasing the confidence intervals, and thus making the flow of the lines smoother. The 50 (0.5%) sampling density is more erratic as compared to the other two. Similar patterns are found in species B for all modelling methods (except GAM and ANN) used, the graphs of which can be found in Appendix D. GAM and ANN show higher importance at 0.5% sampling density with wider confidence intervals which could imply that they are more at risk of overfitting when the sample dataset is not as complete.

3.4. Comparison of the various models

3.4.1. The area between variable importance estimate curves across varying ranges of autocorrelation

The average area between the variable importance curves estimated for a spatial autocorrelation range from 0 to 30%, provides a comparable statistic to measure the average dissimilarity (and hence bias) in the variable importance estimations. The smaller the area between the curves the more robust the estimations of variable importance in the presence of autocorrelation and differing species responses. The outputs are shown both for scenario 1 (Fig.31) and scenario 2 (Fig.32). As can be seen from the Figures, overall the RF and SVM perform better than the rest of the models, while the BRT, Maxent and ANN perform the worst with higher absolute areas, increased variability within the 20 iterations and a higher number of outliers. Amongst the types of species, species B (all unimodal) performs overall better with lower areas both in scenario 1 and 2. Moreover, amongst the scenarios themselves, the higher relative SAC scenario (scenario 1) showed larger areas; therefore there is a positive relationship between relative SAC and magnitude of bias in relative variable importance.

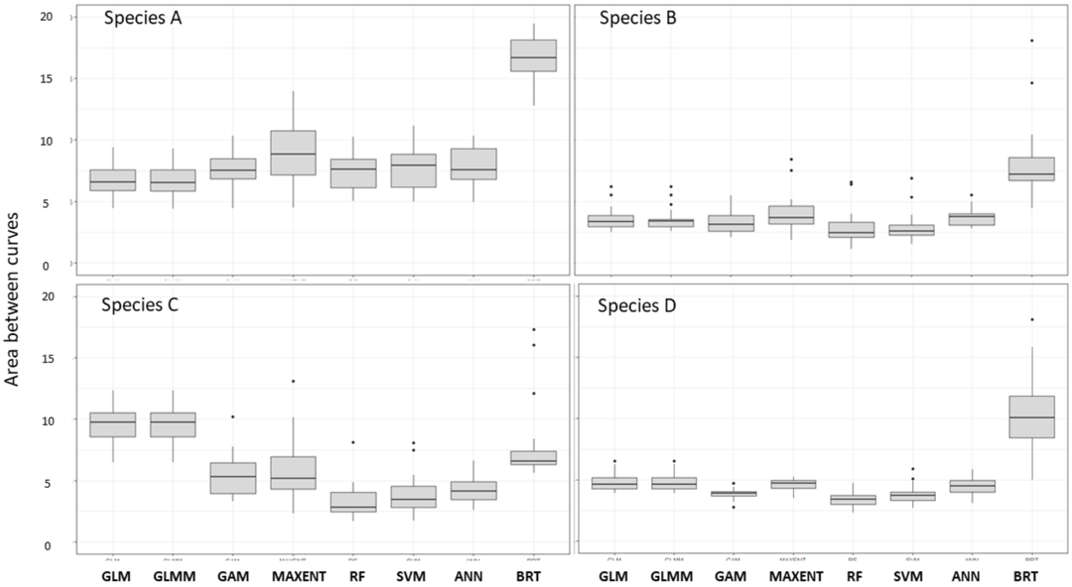


Figure 31: Bar plots showing average area between the variable importance curves for species A, B, C, D; across a range of autocorrelation for all the eight models; and for scenario 1 (background SAC 0%).

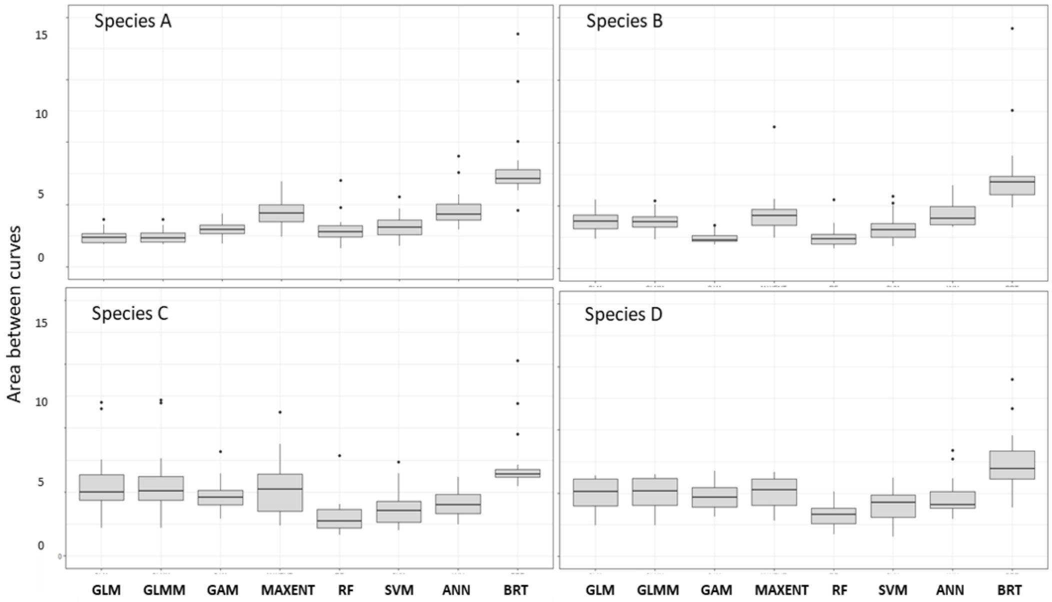


Figure 32: Bar plots showing average area between the variable importance curves for species A, B, C, D; across a range of autocorrelation for all the eight models; and for scenario 2 (background SAC 12.5%).

3.4.2. Model accuracy estimates: AUC and Kappa

The boxplots for the AUC and Kappa of each model for the four different species, for scenario 2 (background SAC 12.5%) have been shown in Fig. 33 and 34. The results have been chosen for ease of reporting for the three levels of SAC (0%,12.5% and 30%) that adequately represents the spectrum. Only results of scenario 2 are shown as the results from scenario 1 are similar and redundant and thus not reported. The mean and standard deviations for the AUC and the Kappa from all eight models have been reported in the form of tables in Appendix E.

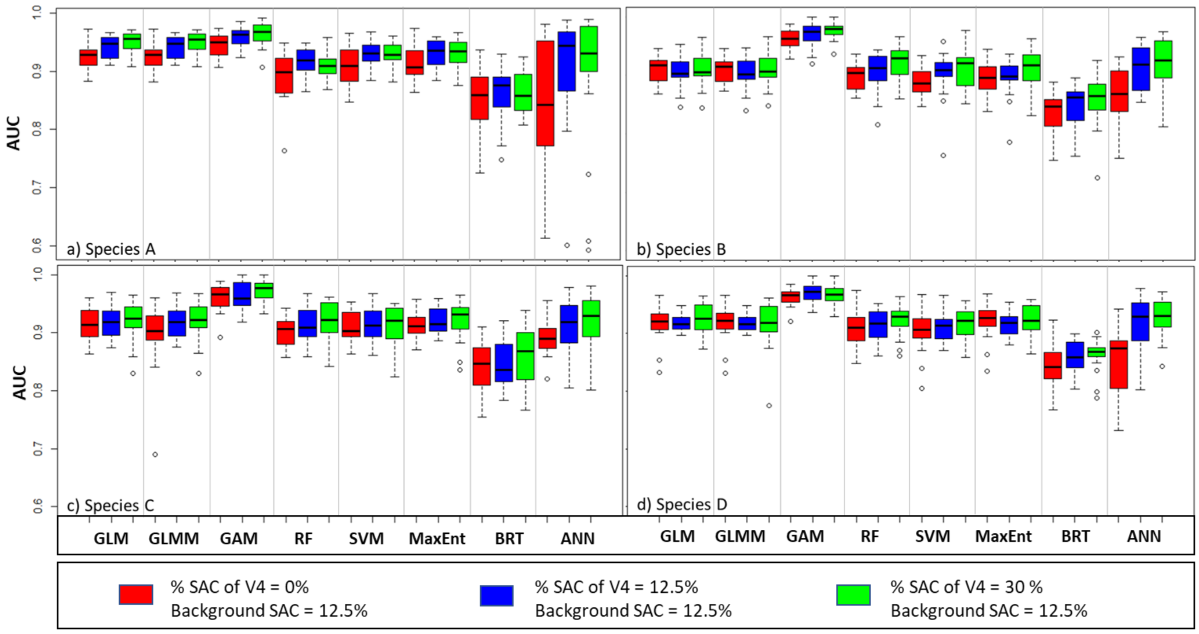


Figure 33: Boxplots showing of AUC values, for three levels of autocorrelation, from all the eight models and for all four species.

The AUC results for Species A and Species B (see Fig.33) show excellent discriminatory power for all the models across the spectrum of autocorrelation, with a marginal increase in AUC as the autocorrelation increases. Similar patterns are seen in AUC values for species C and D also (Fig.33). BRT and ANN show higher variability and overall lower mean AUC estimates, with larger increases in accuracy as autocorrelation increases.

For the Kappa values (Fig.34), more than a trend across the different ranges of autocorrelation, a ranking in the functioning of the models is evident. The values are less optimistic than the consistently excellent AUC values for all models, except for RF which has a constantly high Kappa of 0.98. Amongst the other models, the SVM and the GAM have optimal kappa values between 0.70 and 0.90, ANN, MaxEnt, GLM and GLMM have adequately positive kappa values between 0.60 and 0.70, whereas BRT has the lowest values ranging from 0.47 to 0.56. These patterns are seen across all the four species types, implying that species response curves types do not necessarily affect the model accuracy estimates.

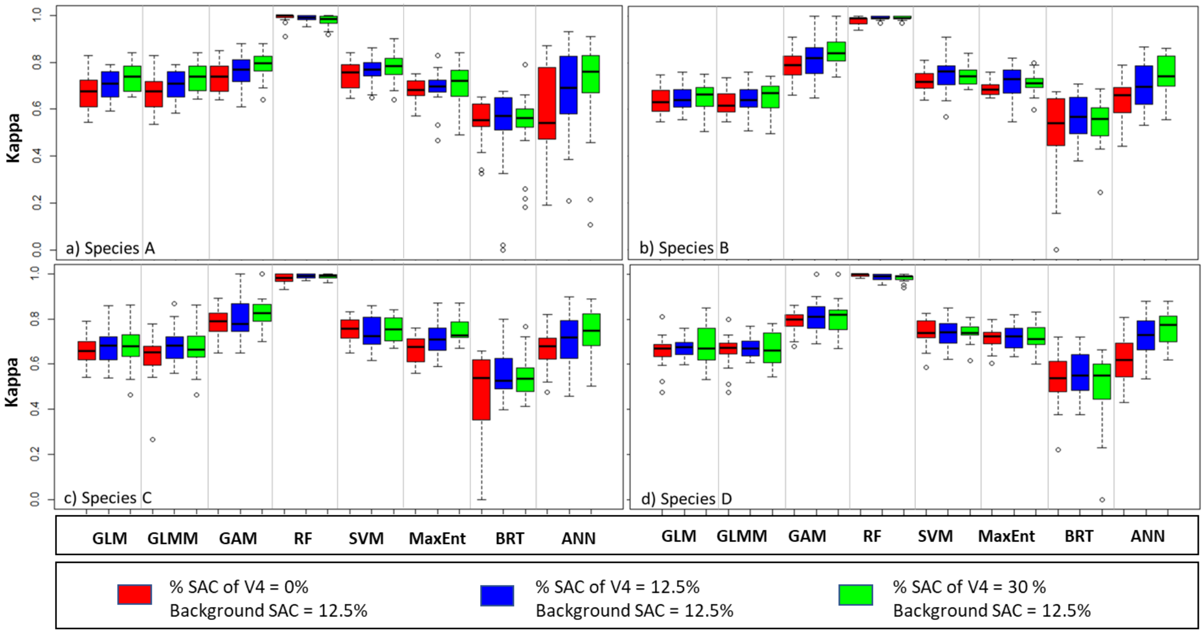


Figure 34: Boxplots showing of kappa values, for three levels of autocorrelation, from all the eight models and for all four species.

3.4.3. Autocorrelation in Residuals

Results from the Morans I on the residuals for each modal run for the 25 levels of autocorrelation were collected, from which the number of times a significant autocorrelation in the residuals was seen was noted. So for each level of autocorrelation and for the different models, the no. of times the residuals were significantly autocorrelated (within the 20 iterations) has been reported. (see Fig. 35).

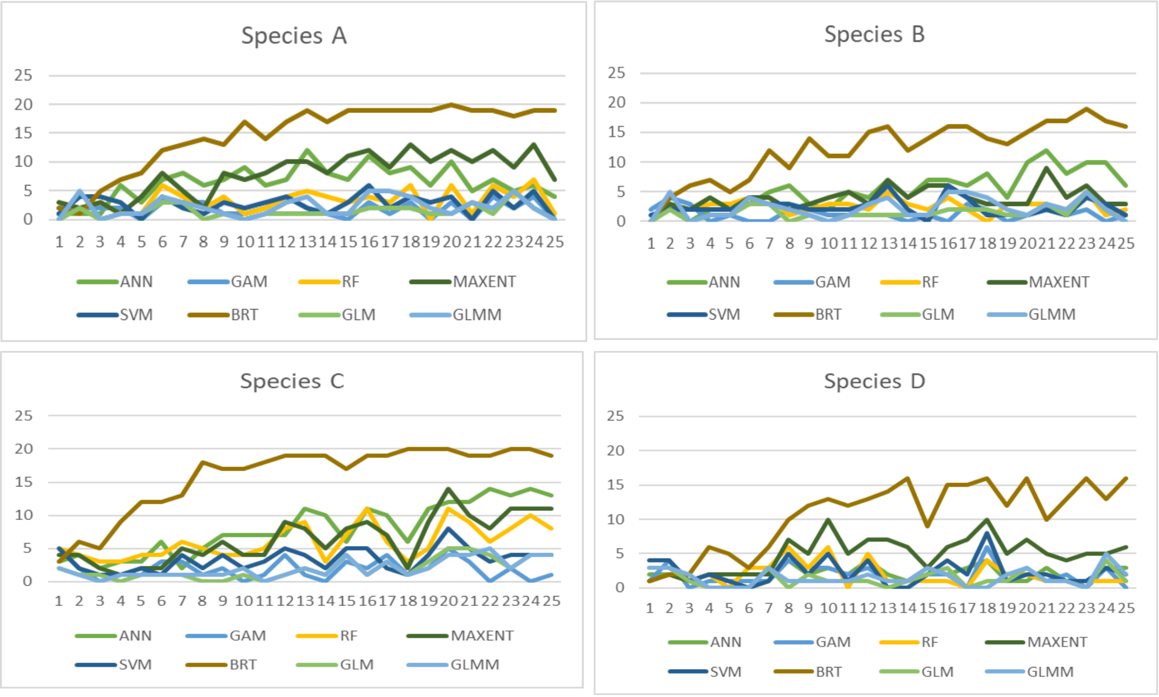


Figure 35: Line plots showing the number of model runs that produced autocorrelated residuals for different ranges of autocorrelation (0% to 30%) and for the different modelling methods.

The graphs in Fig.34. are plotted for the scenario where the covariates for comparison are at 0% autocorrelation only, just because an interesting trend is seen in the model residuals for ANN, MaxEnt and BRT and marginally for RF in Species C, whereas the autocorrelation in the explanatory variables increase so does the residual autocorrelation estimated from the models. This pattern is seen across all the species and is continued into the scenario where the covariates for comparison are at 12.5% autocorrelation, in which case the residuals are consistently autocorrelated for BRT, ANN and Maxent. Overall the results show a ranking of the form: BRT performs the worst with all 20 iterations having autocorrelated residuals; followed by MaxEnt at 15 to 20 of the iterations; and finally ANN at 10 to 15 of the iterations being autocorrelated in the residuals. (See Appendix F for plots of residual autocorrelation at 12.5% baseline autocorrelation).

4. DISCUSSION

4.1. The importance of species response curve geometry in the effect of spatial autocorrelation

One of the main research questions in this study was regarding the effect of spatial autocorrelation in the covariates on model-independent variable importance estimates. An exaggerating effect of spatial autocorrelation on the importance estimates was noticed in all model runs, and thus the red-shift coined by Lennon is evident in all cases but conditioned by the different species responses. For the unimodal response, the exaggerating effect is only noticeable beyond a certain threshold of the relative difference in Moran's I between the covariates (0.09~0.13; also >5 units of percentage SAC), which is higher than the threshold identified for linear responses (0.03~0.08; also >7.5 units of percentage SAC). Also, the magnitude of importance is much more exaggerated in linear responses than unimodal ones. This response specific bias can be explained by the differing geometry of the two curves and its effects on the covariate values whose own distribution vary due to the onset of spatial autocorrelation (Fig. 36).

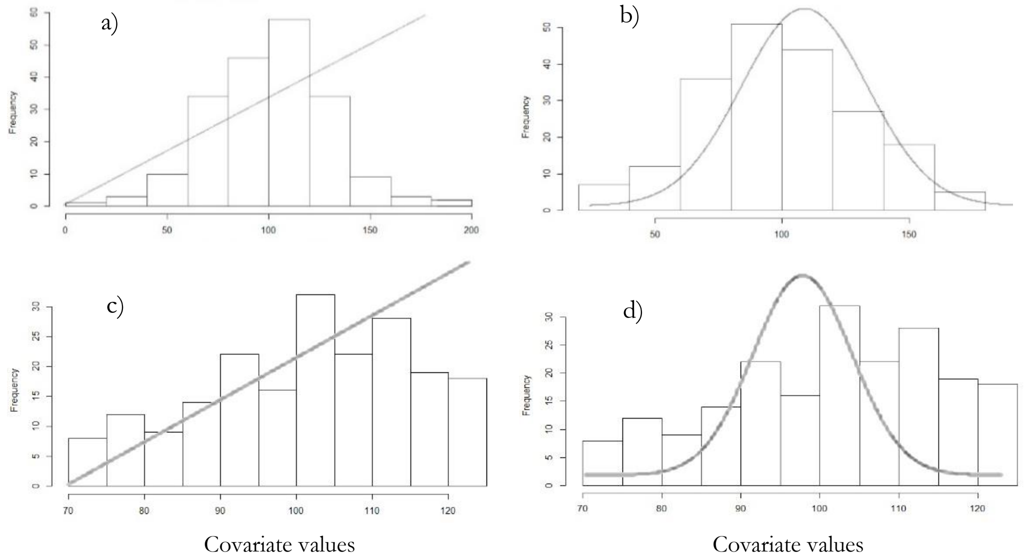


Figure 36: Histograms showing differences in sample distribution geometry w.r.t response curve geometry for a) top row: spatial autocorrelation-0%; b) bottom row: spatial autocorrelation-30%

Consider Fig. 36., which shows the histograms of the covariate values and the corresponding response curves. The top row shows the histogram of uncorrelated covariates and the bottom row shows that of the correlated covariates. Therefore Fig.a,b,c,d represents an uncorrelated linear covariate, an uncorrelated unimodal covariate, an autocorrelated linear covariate and an autocorrelated unimodal covariate respectively. These graphs, together with the knowledge of larger magnitude of inflated importance for correlated over the uncorrelated (c>a & d>b) and for the linear over the unimodal (a>b, c > d) indicates that the shape of the sampled dataset of the covariate with respect to that of the response curve is an important factor (as all other parameters are constant amongst the four cases). As the covariates become more autocorrelated the distribution of the covariate values move towards a non-normal (skewed) distribution. In such cases and for a linear response (Fig.36c), the suitability transformation implies a pseudo-replicated increase in data points that have a higher probability of presence/absence as opposed to datasets with more data in the mid ranges, resulting in the higher importance estimations for this covariate. For the unimodal geometry, the skewed distribution of the autocorrelated covariate has a higher chance of pseudo-replicating in confusing patterns that need not reinforce the covariates contribution to defining presences and absences, though beyond a certain threshold it affects it too, but at a lower magnitude. Hence it is evident from the results that the effect of autocorrelation is reduced (in magnitude and in terms of a higher 'relative SAC' threshold) when the response is unimodal and amplified when the response is linear. Therefore, as discussed, regarding the transformation of the spatial dependencies in the original dataset, the linear preserves the autocorrelation while the unimodal has a higher chance of overriding it. In line with this discussion, it is also relevant to note that the response curve specific behaviour could also be due to the lack of information in a binomial distribution of response data (0 or 1 based on probability) which makes the model more sensitive to the input characteristics of the covariates and might be less prominent in cases of continuous response datasets (normal/Poisson distributions) (F. Dormann et al 2007, Ripley and Venables 2002).

A few exceptions to the above-mentioned pattern would be in the case of generalised linear models with unimodally responded covariates (Species B), where an increase in relative SAC could decrease the variable importance. This could just be an exceptional case, but the reasons can be that as the autocorrelation increases, response curve shifts to match the skewed dataset, therefore the quadratic form initialised in the GLM is not able to fit properly to the model and thus as autocorrelation increases, the variable importance decreases. This effect is not seen in GAM and the other machine learning methods as they are not parametric.

Therefore, to infer the results for possible practical applications, it can be that within a given set of covariates, in a constant geographical extent, the ranking of importance of variables can differ for two different sets of species (A,B), where species A can represent a generalist type of species that survives over wider ranges of the covariate (hence the linear response), whereas species B represents a specialist type of species that survives at only very narrow preferred optima (hence unimodal response). In such cases, it is possible that spatial autocorrelation in the covariates is responsible for the differential covariate importance, as has been noticed regarding similar species (specialist vs generalist) in the same geographical extent and resolution(Peers et al. 2012).

Additionally, the superseding effect of species environmental tolerance (width), in the case of unimodal responses, is another important result (Section 3.2.4). As was seen that even a relative increase in SAC by 30% is overridden by the narrow width (higher precision) of a response curve (though marginally) for all models. Thus, backing up the fact that, both ecologically and statistically, the narrower your niche (the higher the constraint) and the lower your standard deviation (variation around the mean), the more predictive power the variable obtains. Therefore there is a tradeoff between the width of a species response curve and the autocorrelation of the covariate, where the higher SAC can be balanced by a decreasing precision, thus the overall importance can remain the same. This could also be why in many studies with real datasets the results are confounding (Bini et al. 2009).

4.2. The overriding importance of the unimodal over the linear

The second result that the study found was regarding the inherent bias for all the modelling methods for the non-linear response (unimodal) in agreement with the results of the studies by Meynard & Quinn (2007b) and Santika & Hutchinson (2009). Ecologically this can be explained by Liebig's law of the minimum, where the most constraining factor gains more importance (Austin 2006, Huston 2002). The unimodal response to a covariate adds greater constraints to the definition of a species in a given geography which implies that the pattern of presence-absence of the species might depend more on this covariate than others. Therefore, it is likely that the models are identifying real importance rankings. The only logical conclusion is that our initial assumption about a uniform weighting system that denotes equal baseline importance for all types of responses is wrong and is only valid if the responses are of the same geometry across the different covariates (all linear or all unimodal).

This acceptance of the higher estimated importance of the unimodal curve is ecologically sensible as the geometry of the response curve is a representation of the species-environment relationship and not just a statistical aspect. Unlike spatial autocorrelation the species response to a covariate is expected to remain constant at different spatiotemporal frames for the same species, given that the conditions like competition and predation are relatively similar (Guisan et al. 2006). Spatial autocorrelation, on the other hand, can differ for different landforms and thus is inconsistent across spatiotemporal frames. Accounting for this as an ecological mechanism can thus reduce the transferability of the model (Bell and Schlaepfer 2016). This argument also holds for the width of the response curve where it is ecologically sensible to consider the covariate that defines a narrow niche for the species to be more important (as seen in the results) as it also represents a tangible aspect of the species-environment relationship.

Another interesting result regarding the relative importance of linear vs unimodal responses, is that as the importance of the linear response increases, due to its spatial structure becoming more pronounced, at a relative increase of 10% SAC (0.25 Moran's I), the importance of the linearly responded covariate is similar to that of the unimodal response, for all the models and specifically for the RF and SVM. This is because of the high spatial structure in the variable amplifies its role as a constraining feature, and not the covariate itself. It would be interesting to research on simulated levels of explicit spatial structure (in the form of endogenous autocorrelation or simply a spatial contribution of a landscape feature) with separate environmental covariates that are correlated to the spatial counterpart (Hothorn et al. 2011); and analyse if the varying responses on the covariate further dictated relative importance levels, in which case we expect the linear to be always a less important feature (of ecological backing) and the rest to be a manifestation of the spatial geometry.

4.3. Effect of sampling density

The effect of sampling density in this study was not evident. This could be because of the limited 100 by 100 pixels extent of the basic grid for the covariates and the responses. No matter how low the sampling density, the spatial autocorrelation was still preserved in the sample (see Fig. 37 for Moran's I of different sampling densities). The method of simulating the random fields also limited the amount of information the landscapes held. Thus the gradual changes in the covariance structure did not correspond to the changing overall spatial correlation, as the Moran's I levelled off very quickly. Future studies on larger simulated extents might be able to document the actual effect of changing the sampling densities.

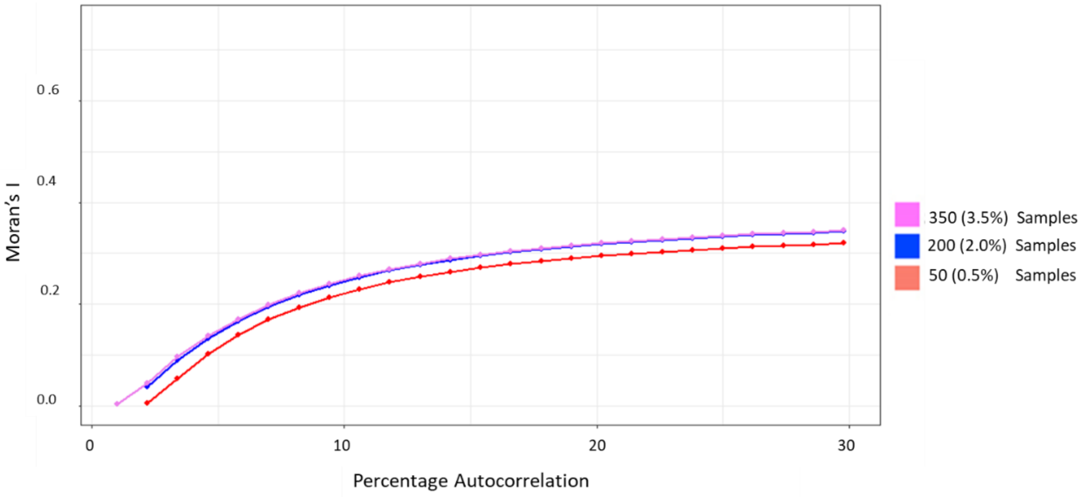


Figure 37: Moran's I for sampled datasets from 3 levels of sampling density (3.5%, 2% and 0.5%)

Traditionally reducing the sampling density is used as a measure against spatial autocorrelation (P. Segurado et al. 2006). Such practices of thinning the datasets are not always effective due to the ambiguity in the specification of the models (Fortin and Dale 2005). Therefore, when you 'thin' the dataset, you might lose out on much information regarding important covariates (of a smaller scale) while increasing the variability in the coefficient estimates, even though the residual sac (RSAC) can be limited. This again can hamper proper ecological inference. Though for models with properly specified inputs, with autocorrelation completely captured by the covariates alone (thus limited RSAC inherently), this method of reducing the sampling densities should work (Fortin and Dale 2005). The reason it did not work on this study is possibly due to the previously mentioned extent limitations.

4.4. Comparison of modelling methods

The relative variable importance was not consistent for any of the models nor for any of the species. For species C and D, the bias in variable importance estimations are considered ecologically correct. For species A and B, all models performed much better for species B than for the other species for reasons mentioned in sections 4.1, though the AUC and kappa values were similar. Amongst all the modelling methods, the machine learning methods, RF and SVM were the most robust and accurate (in terms of the area between the curves and accuracy estimates).Another interesting find is that in the presence of autocorrelation in the covariates, BRT, Maxent, ANN show an increase in autocorrelation in residuals with BRT performing the worst as above 10% SAC all the iterations showed residual SAC. It is difficult to judge the exact reasons for these models performing differently, as the thesis limits itself by applying the models at only one default setting. The weaker performance of these specific machine learning models might be due to overfitting issues or the lack of parameter optimisation. Maxent works on pseudo-background absences, and this study did not supply it with any, providing it with only the true absences from the dataset. For the BRT the learning rate of 0.1 might have decreased the flexibility of the model as now individual trees have more effect (J. Elith et al. 2008), while the ANN though intrinsically tuned by 'biomod2' (in no. of nodes and decay function) might have overfitted to the datasets, thus losing out on important spatial information of the covariates on the test dataset (Wenger and Olden 2012). Future studies can develop better comparisons by going into the details of chosen model types.

The GLM, GAM and spatial GLMM performed similarly in the robustness of the variable importance curves and in the accuracy metrics. Not more than 15% of the iterations produced significant autocorrelated residuals (2-3 mode runs of 20). The reason the spatial counterpart of the GLMM failed to perform better can be due to the full specification of the model, where all the spatial structure in the response curves was informed by the covariates, therefore the RSAC was minimum and the random spatial effect was continuously insignificant meaning that the spatial GLMM performed like a basic GLM. The only valid difference from the GLM and the Spatial GLMM is that the significance values (p-values) for the slope coefficients of the spatially autocorrelated covariate (V4) was less inflated from the GLMM; where all GLM p-values were significant at 99%, and the GLMM was significant at 95%. It can also be seen from Table.3 that the significance values for the autocorrelated covariate is more inflated for species A than for species B which agrees with the notion that unimodal responses are less affected by autocorrelation than linear responses, therefore the p-values of their coefficient estimates are less inflated (as discussed in section 4.1).

Table 3: p values indicating the significance of covariate V4 (autocorrelated variable) at the scenario of baseline sac=50% for species A & B, from GLM, spatial GLMM models.

	Species A		***Species B***	
	GLM	**GLMM**	**GLM**	**GLMM**
0% SAC	0.0019	0.0008	0.0034	0.0068
10% SAC	0.0000	0.0088	0.0004	0.0018
20% SAC	0.0000	0.0008	0.0025	0.0249
30% SAC	0.0001	0.0008	0.0092	0.0129

Therefore, running a spatial model is not efficient in these settings where the exogenous SAC is completely captured by the covariates (no significant RSAC), as the red-shift is still evident in model sensitivity (model independent variable importance). Spatial models are more effective in case of presence of endogenous SAC, that is a pattern in the response variable (presence-absence) that is not accounted for in the covariates (Beale et al. 2007). Though care must be taken to understand the exact causes of RSAC in a non-spatial model before using a spatial variant as many cases have reported an underestimation of the covariate influence when using spatial models (Dormann et al. 2007, Kissling and Carl 2008).

4.5. A critique on the randomisation method of variable importance

Model-independent variable importance estimations as defined by Thuiller et al. (2009) essentially assesses model sensitivity to any covariate, relative to another set of predictors, by analysing changes in the predictive power of the model when using a randomised version of the variable. As seen in the results, the variable importance estimates from none of the models were completely robust or consistent across the range of changing spatial autocorrelation. This could be because the measure is ultimately dependant on the coefficient estimates, the standard error and the significance of the variable, all of which are inflated in the presence of autocorrelation (Dormann 2007). Therefore, methods that include the randomisation of the covariate is highly prone to being misguided under spatial autocorrelation.

Randomisation methods cannot differentiate between the predictive capability of a variable owing explicitly to its spatial structure and the predictive ability of the covariate itself, even if the randomisation follows the same spatial correlation (variogram) structure. This is logical because the geometry of the spatial pattern is also a big part of the predictive power, and under the same variogram model and value ranges there are multiple geometries (or pattern in data) that can satisfy the spatial correlation structure (see Fig.5). Possibly a better way of randomising would have been in preserving the geometry and the correlation structure while shifting around the value ranges of the covariate.

Another method of calculating the actual importance can be if Monte Carlo approaches are used, as mentioned in Fortin & Dale, 2005, in which the randomisation (with only spatial correlation structure preserved) is repeated multiple times to get a distribution of the many possibilities of the importance, and then simple statistical tests can be used to check if the real computed variable importance is significant or not. It could also be that the 'Pearson' correlation is not sufficient in cases of spatially autocorrelated variables. The spatial cross-correlation metric as proposed by Chen (2015) can be an additional efficient tool in this case as it can help derive the 'direct correlation' between two variables beyond the spatial contributions. Araújo & Guisan, 2006 further discussed the inability of regression models to identify individual contributions of predictors in an absolute sense (compared to another set of predictors), the argument holds for model-independent variable importance measures also. Therefore, they favoured methods of hierarchical partitioning and variance partitioning that can provide robust measures for computing the unbiased contribution of each variable (spatial counterpart and otherwise), though not as useful for prediction (Heikkinen et al. 2005, Murray and Conner 2009). Hence, additional measures such as these can give robust backing in establishing efficient testable hypotheses for causal ecological relationships between the covariate and the species.

4.6. Endnote on species characteristics and relative scale of covariates

As predictor variables are increasingly being derived from remotely sensed imagery, the interplay between resolution, geographical extent and species response are more complex. Spatial autocorrelation can be one of the outcomes of the interaction between these three parameters. At any extent of observation, there will always be multiple factors that contribute to the occurrence of a species at multiple scales (Levy 1992). The broader the scale of a variable (like climatic factors), the more autocorrelated the covariate. Species responses to such covariates will usually be truncated, since not enough of the specific environmental range (for e.g. temperature ranges) is captured to identify a species optimum; therefore the relationship ends up being monotonic (for e.g. the warmer, the better) (Guo 2014). In such cases when truncated responses are used in relation to other types of response curves (that are not constraining enough to gain statistical importance), the red shift will pose as a significant issue, and true ecological mechanisms will not be captured (Austin et al 2010), which is why multiple studies have identified climatic variables as being increasingly important in variable importance rankings.

Autocorrelation can also arise due to mismatches between resolution and scale of species response (Atkinson 1993, De Knegt et al. 2010). In a certain geographic extent where complete global ranges of all the input covariates for a species is captured (species B in this study), a significant magnitude of relative SAC in the covariates could imply a decrease in the resolution of one of the images included (coarser resolution). The results of this study on such simulated environments suggest that exogenous (covariate) spatial autocorrelation inflates the variable importance to a lower magnitude if the complete and in-scale species response to the explanatory covariate is captured. Therefore making global case studies on biogeographical patterns using species distribution modelling less likely to be flawed (Jane Elith and Leathwick 2009).

However, in reality such models with a fully specified (no residual SAC) and to-scale species-covariates relationships are hard to find (every level of observation will have some or the other scale mismatches) for example important endogenous causes of autocorrelation (like dispersal) or small scale exogenous factors (like local topography, soil conditions) cannot be captured at broad scales and can give rise to residual SAC, just as broad-scaled variables captured at a smaller scale will give rise to the red-shift problem. Which is why (Austin et al. 2010) suggested that high-resolution global datasets must be used to understand the true effects of climatic variables. Therefore, it is difficult to find similar patterns in studies using real datasets if the relative scales with respect to species responses are further confounding (Bini et al. 2009, Hawkins et al. 2007). Nevertheless, the results of this study can help gain insight into the possible causes for a certain observed ranking of variable importance, that are statistically proved in a simulated environment, and thus it can help make better inferences.

4.7. Limitations in experimental design

Many aspects that have been controlled in this study to investigate the effect of spatial autocorrelation and species response curve geometry in an isolated setting are not replicable in reality. For example, it is not possible to coerce the prevalence of a species to a comfortable 50%. Skewed prevalences are commonly found, and lead to effects in model functioning that have not been documented in this study (Meynard and Quinn 2007, Sor et al. 2017b). The isotropic and stationary system of autocorrelation simulated is another such crude condition that is not found in nature. As stated in Fortin & Dale, 2005, negative autocorrelation is also often observed in cyclic patterns of nature and can nullify the effect of autocorrelation itself.

Further regarding the use of percentage autocorrelation as an indicator for spatial autocorrelation in the covariates might be flawed, as the shape of the modelled autocorrelation can vary (spherical, exponential etc.) and thus estimated relationships between relative SAC and other parameters investigated can be inconsistent. Also, the range value used here (range/total extent *100 is the percentage SAC) is the input to a covariance matrix and not the explicitly calculated range of the variogram, though the terms can be theoretically interchangeable since the exponential distance function in the covariance matrix implies an exponential variogram of the same range (Nychka et al. 2017). The use of Global Moran's I is a better alternative, but it would have been better to design the experiment based on it and not just add it as an afterthought.

Another important aspect controlled in this study is the simplistic simulations of species response geometries, i.e. the monotonic and linear, against which many studies have stated the improbability of finding such 'perfect' shapes in nature. Species response curves are a result of many complicated processes that are not independent from each other (e.g. competition amongst species populations, non-environmental factors that obstruct dispersal)(Austin 2006); therefore most real species responses follow highly variable shapes. Interactions between variables also play an appreciable role in defining species responses, yet the study only considers additive habitat suitability for the species. The controlled precision of the unimodal curves is also not often replicable in nature. Precise unimodal curves for autocorrelated data can be common if data at high resolution for a global dataset is used. However, at smaller scales the width (standard deviation) or environmental tolerance of the species will also add further complexity to the model (Jamil et al. 2014, Rydgren et al. 2003). Also, the maximum suitability of the response curves was set to the highest value of one (min-0, max-1) for all the covariates. This need not be the case as covariates can vary in their explanatory power for the species niche, i.e. not all will affect the suitability for the species to the same magnitude (Ververk 2011), which brings us to the next controlled setting in the study, that of a constant initial variable importance. The effect of unequal variable importance for the different covariates must also be investigated to estimate a better guide for the effect of species responses in combination with spatial autocorrelation. And finally, it is also important to note that the sampling density for all such above-mentioned conditions can further confuse the results, with patterns in datasets induced by the survey (purposive) methods used (Veloz 2009). The use of purely random sampling like the ones used in this simulation study is not realistically possible as external conditions of extreme topographies etc can affect the sampling scheme.

However, beyond the limitations stated, the inferences from simplistic simulations are still valid as a starting point for testing statistical accuracies/inaccuracies that can get confounded when using real datasets with complex patterns and relationships (Bini et al 2009). Nonetheless, future studies can move forward by adding more parameters to assess complex scenarios that are more replicable and applicable to real datasets.

5. CONCLUSION

Many studies had found inconclusive results regarding the effect of spatial autocorrelation in variable importance, though the problem of inflation of significances is a widely accepted issue. This study clears out the basic functionalities of species distribution models when faced with exogenous autocorrelation and clarifies few other determinant factors, mainly species response curve geometry, that either exaggerate or compensate for this inherent spatial structure. The study concludes that the red-shift coined by Lennon is an evident flaw when assessing model independent variable importance in the presence of exogenous autocorrelation, with species response characteristics and the relative levels of spatial autocorrelation of the covariates being the most determinant factors. Consequences of such red-shifted variables though robust in a single spatiotemporal frame, can affect the predictive accuracy of the models immensely when transferring the same to different regions or different time periods when the spatial correlation structure of the covariates now vary. It is misguided to consider the inflated importance as an actual ecological mechanism, and thus proper methods to account for the spatial structure must be incorporated while computing model independent variable importance estimates from species distribution models.

LIST OF REFERENCES

Allouche O, Tsoar A and Kadmon R (2006) Assessing the accuracy of species distribution models: Prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology* 43(6): 1223–1232.

Araújo MB and Guisan A (2006) Five (or so) challenges for species distribution modelling. *Journal of Biogeography* 33(10): 1677–1688. Available at: www.blackwellpublishing.com/jbi (accessed 27/01/19).

Atkinson PM (1993) The effect of spatial resolution on the experimental variogram of airborne MSS imagery. *International Journal of Remote Sensing*. Taylor & Francis Group 14(5): 1005–1011. Available at: https://www.tandfonline.com/doi/full/10.1080/01431169308904391 (accessed 05/02/19).

Austin M. (2002) Spatial prediction of species distribution: An interface between ecological theory and statistical modelling. *Ecological Modelling*. Elsevier 157(2–3): 101–118. Available at: https://www.sciencedirect.com/science/article/pii/S0304380002002053 (accessed 16/02/19).

Austin M (2006) Species distribution models and ecological theory: A critical assessment and some possible new approaches. . Available at: http://www.atmosp.physics.utoronto.ca/people/lev/ESSgc2/speciesdistmodelseval.pdf (accessed 27/01/19).

Austin MP and Nicholls AO (1997) To fix or not to fix the species limits, that is the ecological question: Response to Jari Oksanen. *Journal of Vegetation Science*. John Wiley & Sons, Ltd (10.1111) 8(5): 743–748. Available at: http://doi.wiley.com/10.2307/3237380 (accessed 24/02/19).

Austin MP, Nicholls AO and Margules CR (1990) Measurement of the Realized Qualitative Niche: Environmental Niches of Five Eucalyptus Species. *Ecological Monographs*. John Wiley & Sons, Ltd 60(2): 161–177. Available at: http://doi.wiley.com/10.2307/1943043 (accessed 24/02/19).

Austin MP, Patraw K and Niel V (2010) Improving species distribution models for climate change studies: Variable selection and scale. *Article in Journal of Biogeography*. Available at: https://www.researchgate.net/publication/229689687 (accessed 05/02/19).

Bahn V, J. O'Connor R and B. Krohn W (2006) Importance of spatial autocorrelation in modeling bird distributions at a continental scale. *Ecography* 29(6): 835–844.

Beale CM, Lennon JJ, Elston DA, Brewer MJ and Yearsley JM (2007) Red herrings remain in geographical ecology: A reply to Hawkins et al. (2007). *Ecography* 30(6): 845–847.

Beguería S and Pueyo Y (2009) A comparison of simultaneous autoregressive and generalized least squares models for dealing with spatial autocorrelation. *Global Ecology and Biogeography* 18(3): 273–279. Available at: www.blackwellpublishing.com/geb (accessed 21/01/19).

Bell DM and Schlaepfer DR (2016) On the dangers of model complexity without ecological justification in species distribution modeling. *Ecological Modelling* 330: 50–59. Available at: http://dx.doi.org/10.1016/j.ecolmodel.2016.03.012 (accessed 13/01/19).

Berdugo M, Maestre FT, Kéfi S, Gross N, Le Bagousse-Pinguet Y and Soliveres S (2019) Aridity preferences alter the relative importance of abiotic and biotic drivers on plant species abundance in global drylands. *Journal of Ecology*. John Wiley & Sons, Ltd (10.1111) 107(1): 190–202. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1111/1365-2745.13006 (accessed 24/02/19).

Betts MG, Ganio LM, Huso MMP, Som NA, Huettmann F, Bowman J and A.Wintle B (2007) Comment on Methods to account for spatial autocorrelation in the analysis of species distributional data: A review. *Ecography* 30(5): 609–628.

Bini LM, Diniz-Filho JAF, Rangel TFLVB, Akre TSB, Albaladejo RG, Albuquerque FS, Aparicio A, Araújo MB, Baselga A, Beck J, Bellocq MI, Böhning-Gaese K, Borges PAV, Castro-Parga I, Chey VK, Chown SL, De Marco P, Dobkin DS, Ferrer-Castán D, Field R, Filloy J, Fleishman E, Gómez JF, Hortal J, Iverson JB, Kerr JT, Kissling WD, Kitching IJ, León-Cortés JL, Lobo JM, Montoya D, Morales-Castilla I, Moreno JC, Oberdorff T, Olalla-Tárraga MÁ, Pausas JG, Qian H, Rahbek C, Rodríguez MÁ, Rueda M, Ruggiero A, Sackmann P, Sanders NJ, Terribile LC, Vetaas OR and Hawkins BA (2009) Coefficient shifts in geographical ecology: An empirical evaluation of spatial and non-spatial regression. *Ecography* 32(2): 193–204.

Breiman L (2001) *Random Forests*. . Available at: https://link.springer.com/content/pdf/10.1023%2FA%3A1010933404324.pdf (accessed 23/01/19).

Chen Y (2015) A New Methodology of Spatial Cross-Correlation Analysis. *PLOS ONE*. Public Library of Science 10(5): e0126158. Available at: https://dx.plos.org/10.1371/journal.pone.0126158 (accessed 24/02/19).

Diniz-Filho JAF, Bini LM and Hawkins BA (2003) Spatial autocorrelation and red herrings in geographical ecology. *Global Ecology and Biogeography* 12(1): 53–64.

Dormann CF (2007) Effects of incorporating spatial autocorrelation into the analysis of species distribution data. *Global Ecology and Biogeography* 16(2): 129–138.

Dormann CF, M. McPherson J, B. Araújo M, Bivand R, Bolliger J, Carl G, G. Davies R, Hirzel A, Jetz W, Daniel Kissling W, Kühn I, Ohlemüller R, R. Peres-Neto P, Reineking B, Schröder B, M. Schurr F and Wilson R (2007) Methods to account for spatial autocorrelation in the analysis of species distributional data: A review. *Ecography* 30(5): 609–628.

Drake JM, Randin C and Guisan A (2006) Modelling ecological niches with support vector machines. *Journal of Applied Ecology*. John Wiley & Sons, Ltd (10.1111) 43(3): 424–432. Available at: http://doi.wiley.com/10.1111/j.1365-2664.2006.01141.x (accessed 23/01/19).

Duque-Lazo J (2013) Transferability of species distribution models . A case study of the fungus Phytophthora cinnamomi in Andalusia and Southwest Australia. Elsevier B.V. 320(JUNE 2013): 88.

E. P and K. S (2018) ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. R Foundation for Statistical Computing.

Elith J and Leathwick JR (2009) Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annual Review of Ecology, Evolution, and Systematics*. Annual Reviews 40(1): 677–697. Available at: http://www.annualreviews.org/doi/10.1146/annurev.ecolsys.110308.120159 (accessed 21/02/19).

Elith J, Leathwick JR and Hastie T (2008) A working guide to boosted regression trees. *Journal of Animal Ecology* 77(4): 802–813.

F. Dormann C, M. McPherson J, B. Araújo M, Bivand R, Bolliger J, Carl G, G. Davies R, Hirzel A, Jetz W, Daniel Kissling W, Kühn I, Ohlemüller R, R. Peres-Neto P, Reineking B, Schröder B, M. Schurr F and Wilson R (2007) Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography* 30(5): 609–628. Available at: http://doi.wiley.com/10.1111/j.2007.0906-7590.05171.x (accessed 09/01/19).

Fortin M-JM-J and Dale MRT (2005) Chapter 5: Dealing with spatial autocorrelation. *Spatial Analysis: a guid for ecologists*, 212–255. Available at: https://doi.org/10.1017/CBO9780511542039.006 (accessed 22/01/19).

Fotheringham AS, Brunsdon C and Charlton M (2000) *Quantitative geography : perspectives on spatial data analysis* (First.). Great Britain: Sage Publications.

Franklin J and Miller JA (2010) *Mapping Species Distributions:* . Cambridge University Press. Available at: https://www.dawsonera.com:443/abstract/9780511765605.

Freeman EA and Moisen G (2008) PresenceAbsence: An R Package for Presence-Absence Model Analysis. Journal of Statistical Software, 23(11):1-31. Available at: http://www.jstatsoft.org/v23/i11.

Graham JH and Duda JJ (2011) The Humpbacked Species Richness-Curve: A Contingent Rule for Community Ecology. *International Journal of Ecology*. Hindawi 2011: 1–15. Available at: http://www.hindawi.com/journals/ijecol/2011/868426/ (accessed 27/01/19).

Guisan A, Edwards TC and Hastie T (2002) Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological Modelling* 157(157): 89–100. Available at: www.elsevier.com/locate/ecolmodel (accessed 29/01/19).

Guisan A, Lehmann A, Ferrier S, Austin M, Overton JMC, Aspinall R and Hastie T (2006) Making better biogeographical predictions of species' distributions. *Journal of Applied Ecology* 43: 386–392. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.515.8215&rep=rep1&type=pdf (accessed 27/01/19).

Guisan A and Zimmermann NE (2000) Predictive habitat distribution models in ecology. *Ecological Modelling* 135(2–3): 147–186. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0304380000003549.

Guo J (2014) Modelling Loggerhead Sea Turtle (Caretta Caretta) Nesting Habitat. . Available at: https://library.itc.utwente.nl/papers_2014/msc/nrm/guo.pdf (accessed 27/01/19).

Hanberry BB and He HS (2013) Prevalence, statistical thresholds, and accuracy assessment for species distribution models. *Web Ecol* 13: 13–19. Available at: www.web-ecol.net/13/13/2013/ (accessed 27/01/19).

Hastie T, Elith J, Dudík M, Chee YE, Yates CJ and Phillips SJ (2010) A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions* 17(1): 43–57. Available at: http://www.prbo.org/ (accessed 24/01/19).

Hawkins BA, Diniz-Filho JAF, Mauricio Bini L, De Marco P and Blackburn TM (2007) Red herrings revisited: Spatial autocorrelation and parameter estimation in geographical ecology. *Ecography* 30(3):

375–384.

Heikkinen RK, Luoto M, Kuussaari M and Pöyry J (2005) New insights into butterfly–environment relationships using partitioning methods. *Proceedings of the Royal Society B: Biological Sciences* 272(1577): 2203–2210. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16191631 (accessed 05/02/19).

Hirzel AH and Guisan A (2002) Which is the optimal sampling strategy for habitat suitability modelling. *Ecological Modelling* 157(2–3): 331–341.

Hirzel AH, Helfer V and Metral F (2001) Assessing habitat-suitability models with a virtual species. *Ecological Modelling* 145(2–3): 111–121.

Hothorn T, Müller J, Schröder B, Kneib T and Brandl R (2011) Decomposing environmental, spatial, and spatiotemporal components of species distributions. *Ecological Monographs*. John Wiley & Sons, Ltd 81(2): 329–347. Available at: http://doi.wiley.com/10.1890/10-0602.1 (accessed 04/02/19).

Huston MA (2002) Introductory essay: Critical issues for improving predictions. *Predicting species occurrences: Issues of Accuracy and Scale*. Covelo, California: Island press.

Hutchinson GE (1957) Concluding Remarks. *Cold Spring Harbor Symposia on Quantitative Biology*. Cold Spring Harbor Laboratory Press 22(0): 415–427. Available at: http://symposium.cshlp.org/cgi/doi/10.1101/SQB.1957.022.01.039 (accessed 27/01/19).

Jamil T, Kruk C and ter Braak CJF (2014) A Unimodal Species Response Model Relating Traits to Environment with Application to Phytoplankton Communities. *PLoS ONE*. Public Library of Science 9(5): e97583. Available at: https://dx.plos.org/10.1371/journal.pone.0097583 (accessed 27/01/19).

Keitt TH, Bjørnstad ON, Dixon PM and Citron-Pousty S (2002) Accounting for spatial pattern when modeling organism-environment interactions. *Ecography* 25(5): 616–625.

Kissling WD and Carl G (2008) Spatial autocorrelation and the selection of simultaneous autoregressive models. *Global Ecology and Biogeography* 17(1): 59–71.

De Knegt HJ, Van Langevelde F, Coughenour MB, Skidmore AK, De Boer WF, Heitkönig IMA, Knox NM, Slotow R, Van Der Waal C and Prins HHT (2010) Spatial autocorrelation and the scaling of species-environment relationships. *Ecology* 91(8): 2455–2465.

Kühn I (2007) Incorporating spatial autocorrelation may invert observed patterns. *Diversity and Distributions* 13(1): 66–69.

Legendre P (1993) Spatial Autocorrelation : Trouble or New Paradigm. 74(6): 1659–1673.

Legendre P and Fortin MJ (1989) Spatial pattern and ecological analysis. *Vegetatio* 80(2): 107–138.

Lennon JJ (2000) Red-shifts and red herrings in geographical ecology. *Ecography* 23(1): 101–113. Available at: http://doi.wiley.com/10.1111/j.1600-0587.2000.tb00265.x.

Leroy B, Meynard CN, Bellard C and Courchamp F (2016) virtualspecies, an R package to generate virtual species distributions. *Ecography*. John Wiley & Sons, Ltd (10.1111) 39(6): 599–607. Available at: http://doi.wiley.com/10.1111/ecog.01388 (accessed 14/01/19).

Levy MB (1992) the Problem of Pattern and Scale in Ecology. *Ecology* 73(6): 1943–1967.

Luoto M, Pöyry J, Heikkinen RK and Saarinen K (2005) Uncertainty of bioclimate envelope models based on the geographical distribution of species. *Global Ecology and Biogeography*. John Wiley & Sons, Ltd (10.1111) 14(6): 575–584. Available at: http://doi.wiley.com/10.1111/j.1466-822X.2005.00186.x (accessed 27/01/19).

Meineri E, Skarpaas O and Vandvik V (2012) Modeling alpine plant distributions at the landscape scale: Do biotic interactions matter? *Ecological Modelling*. Elsevier B.V. 231(April): 1–10. Available at: http://dx.doi.org/10.1016/j.ecolmodel.2012.01.021.

Meynard CN and Quinn JF (2007) Predicting species distributions: A critical comparison of the most common statistical models using artificial species. *Journal of Biogeography*. John Wiley & Sons, Ltd (10.1111) 34(8): 1455–1469. Available at: http://doi.wiley.com/10.1111/j.1365-2699.2007.01720.x (accessed 27/01/19).

Miller HJ (2004) Tobler's first law and spatial analysis. *Annals of the Association of American Geographers* 94(2): 284–289. Available at: http://www.tandfonline.com/doi/full/10.1111/j.1467-8306.2004.09402005.x.

Miller JA (2014) Virtual species distribution models: Using simulated data to evaluate aspects of model performance. *Progress in Physical Geography* 38(1): 117–128.

Murray K and Conner MM (2009) *Methods to quantify variable importance: implications for the analysis of noisy ecological data. Ecology*. Available at: https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.1890/07-1929.1 (accessed 24/02/19).

Naimi B (2015) *On uncertainty in species distribution modelling* (PhD thesis.). Enschede, Netherlands: University of Twente. Available at: http://purl.org/utwente/doi/10.3990/1.9789036538404.

Naimi B and Araújo MB (2016) Sdm: A reproducible and extensible R platform for species distribution modelling. *Ecography* 39(4): 368–375.

Nychka D, Furrer R, Paige J and Sain S (2017) fields: Tools for spatial data. Boulder, CO, USA: University Corporation for Atmospheric Research. Available at: www.image.ucar.edu/~nychka/Fields.

Oksanen J and Michin P (2002) Continuum theory revisted: what shape are species responses along ecological gradients? *Ecological Modelling* 157: 119–129. Available at: https://ac.els-cdn.com/S0304380002001904/1-s2.0-S0304380002001904-main.pdf?_tid=e8a68e6f-33c1-4f10-b89a-6aac8537db52&acdnat=1548074312_48c7454493a93bf23d33dd60ed852b41 (accessed 21/01/19).

Peers MJL, Thornton DH and Murray DL (2012) Reconsidering the specialist-generalist paradigm in niche breadth dynamics: resource gradient selection by Canada lynx and bobcat. *PloS one*. Public Library of Science 7(12): e51488. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23236508 (accessed 03/02/19).

Phillips SJ, Anderson RP and Schapire RE (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling*. Elsevier 190(3–4): 231–259. Available at: https://www.sciencedirect.com/science/article/pii/S030438000500267X (accessed 24/01/19).

Phillips SJ and Schapire RE (2004) *A Maximum Entropy Approach to Species Distribution Modeling*. . Available at: https://www.cs.princeton.edu/~schapire/papers/maxent_icml.pdf (accessed 24/01/19).

R Core team (2017) R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Ripley B and Venables W (2002) *Modern Applied Statistics with S*. Springer. Available at: http://www.insightful.com. (accessed 20/02/19).

Rocha AD, Groen TA, Skidmore AK, Darvishzadeh R and Willemen L (2017) The Naïve Overfitting Index Selection (NOIS): A new method to optimize model complexity for hyperspectral data. *ISPRS Journal of Photogrammetry and Remote Sensing*. International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS) 133: 61–74. Available at: https://doi.org/10.1016/j.isprsjprs.2017.09.012.

Rydgren K, Økland RH and Økland T (2003) Species response curves along environmental gradients. A case study from SE Norwegian swamp forests. *Journal of Vegetation Science*. John Wiley & Sons, Ltd (10.1111) 14(6): 869–880. Available at: http://doi.wiley.com/10.1111/j.1654-1103.2003.tb02220.x (accessed 27/01/19).

Santika T and Hutchinson MF (2009) The effect of species response form on species distribution model prediction and inference. *Ecological Modelling*. Elsevier 220(19): 2365–2379. Available at: https://www.sciencedirect.com/science/article/pii/S0304380009004049?via%3Dihub (accessed 28/01/19).

Segurado P and Araújo MB (2004) An evaluation of methods for modelling species distributions. *Journal of Biogeography*. John Wiley & Sons, Ltd (10.1111) 31(10): 1555–1568. Available at: http://doi.wiley.com/10.1111/j.1365-2699.2004.01076.x (accessed 28/01/19).

Segurado P, Araújo MB and Kunin WE (2006) Consequences of spatial autocorrelation for niche-based models. *Journal of Applied Ecology* 43(3): 433–444.

Sor R, Park Y-S, Boets P, Goethals PLM and Lek S (2017a) Effects of species prevalence on the performance of predictive models. *Ecological Modelling*. Elsevier 354: 11–19. Available at: https://www.sciencedirect.com/science/article/pii/S0304380016304367 (accessed 27/01/19).

Sor R, Park Y-S, Boets P, Goethals PLM and Lek S (2017b) Effects of species prevalence on the performance of predictive models. *Ecological Modelling* 354: 11–19. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0304380016304367 (accessed 19/02/19).

Thuiller W (2003) BIOMOD - Optimizing predictions of species distributions and projecting potential future shifts under global change. *Global Change Biology* 9(10): 1353–1362.

Thuiller W, Lafourcade B, Engler R and Araújo MB (2009) BIOMOD - A platform for ensemble forecasting of species distributions. *Ecography* 32(3): 369–373.

Veloz SD (2009) Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models. *Journal of Biogeography*. John Wiley & Sons, Ltd (10.1111) 36(12): 2290–2299. Available at: http://doi.wiley.com/10.1111/j.1365-2699.2009.02174.x (accessed 21/02/19).

Ververk W (2011) Explaining General Patterns in Species Abundance and Distributions. *Nature Education Knowledge* 3(10):38. Available at: https://www.nature.com/scitable/knowledge/library/explaining-general-patterns-in-species-abundance-and-23162842.

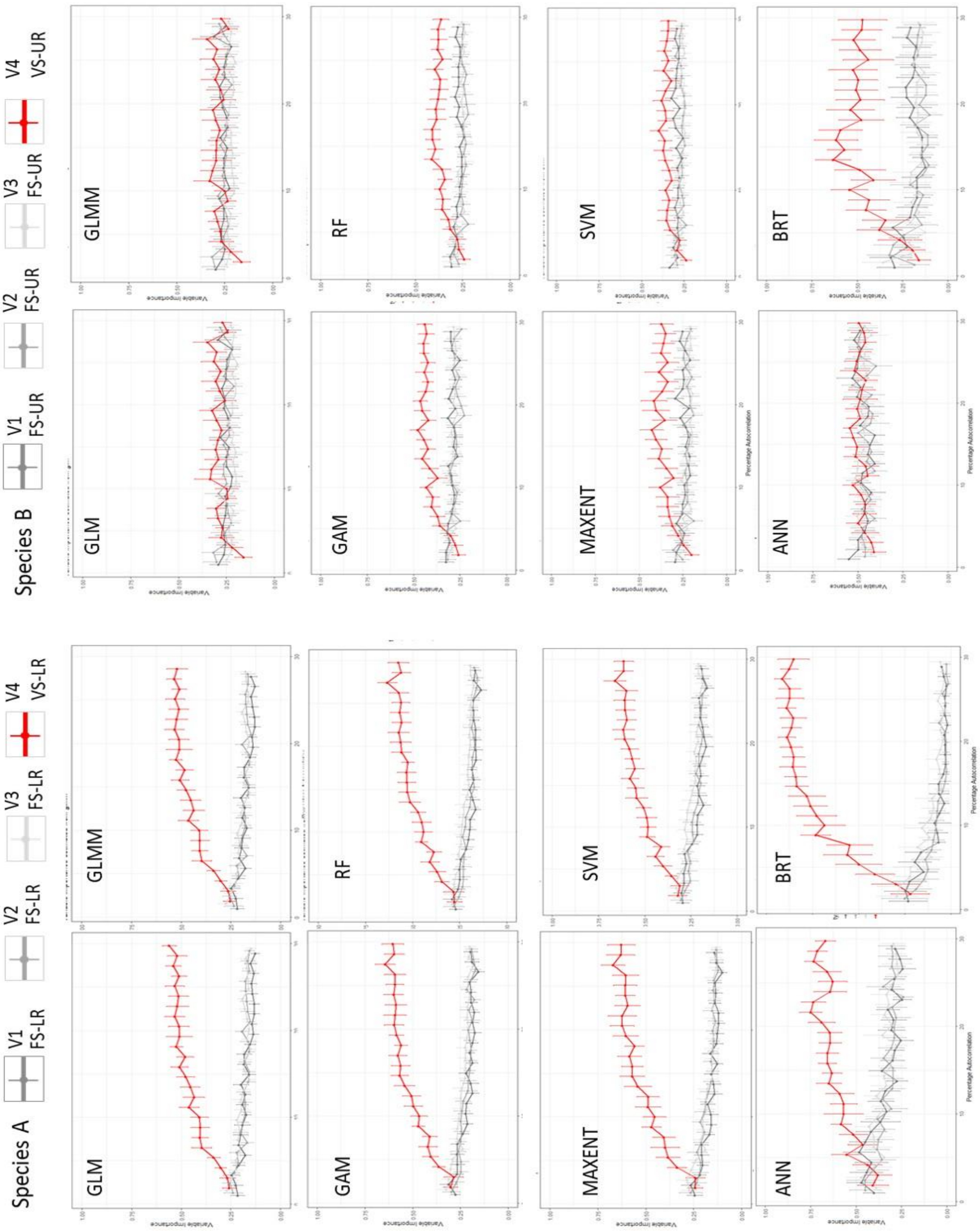
Wei P, Lu Z and Song J (2015) Variable importance analysis: A comprehensive review. *Reliability Engineering and System Safety* 142: 399–432.

Wenger SJ and Olden JD (2012) Assessing transferability of ecological models: an underappreciated aspect of statistical validation. *Methods in Ecology and Evolution*. John Wiley & Sons, Ltd (10.1111) 3(2): 260–267. Available at: http://doi.wiley.com/10.1111/j.2041-210X.2011.00170.x (accessed 04/02/19).

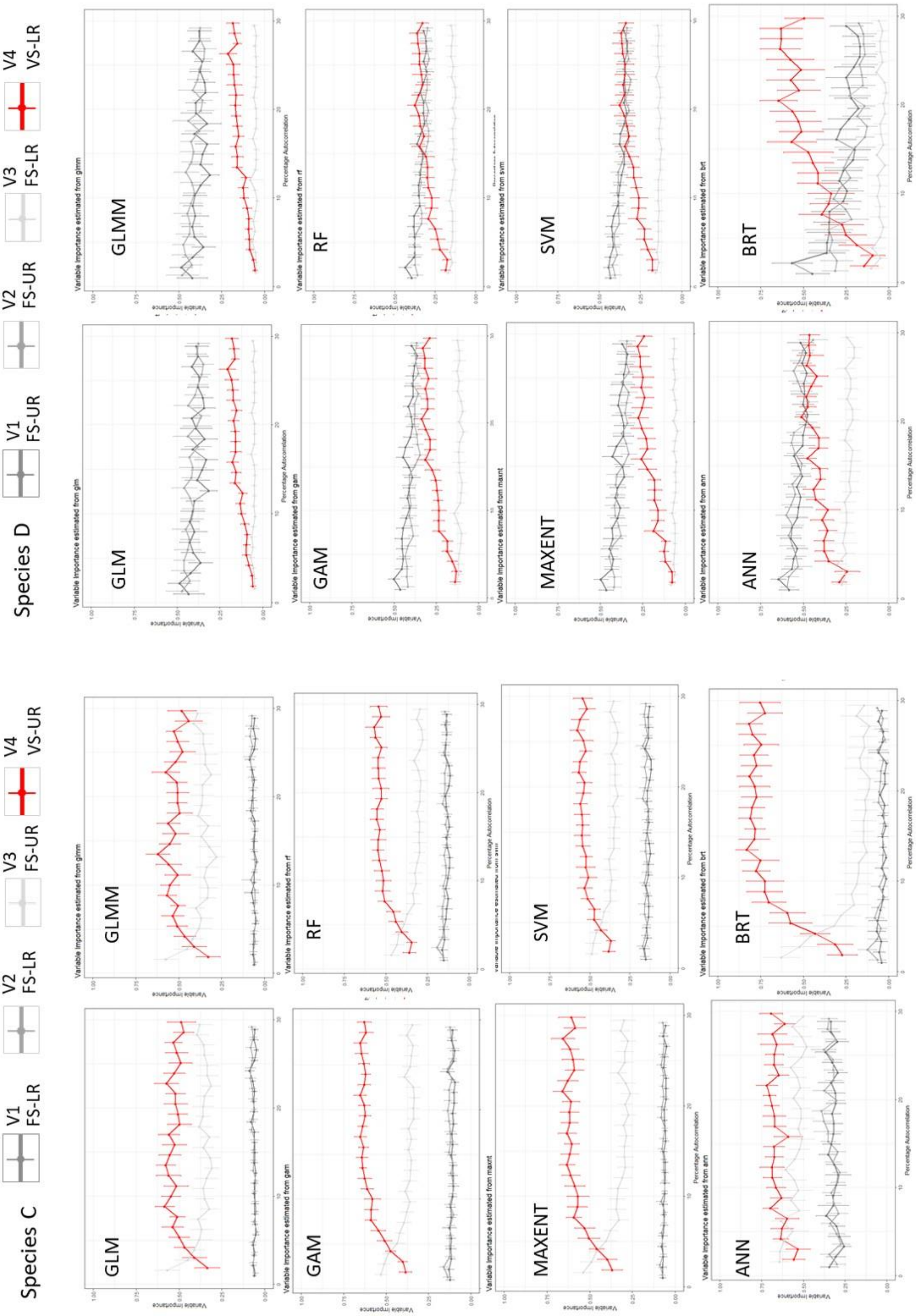
6. APPENDIX

A- Variable importance curves from all models across 4 species & 2 baseline SAC (0%,12.5%)

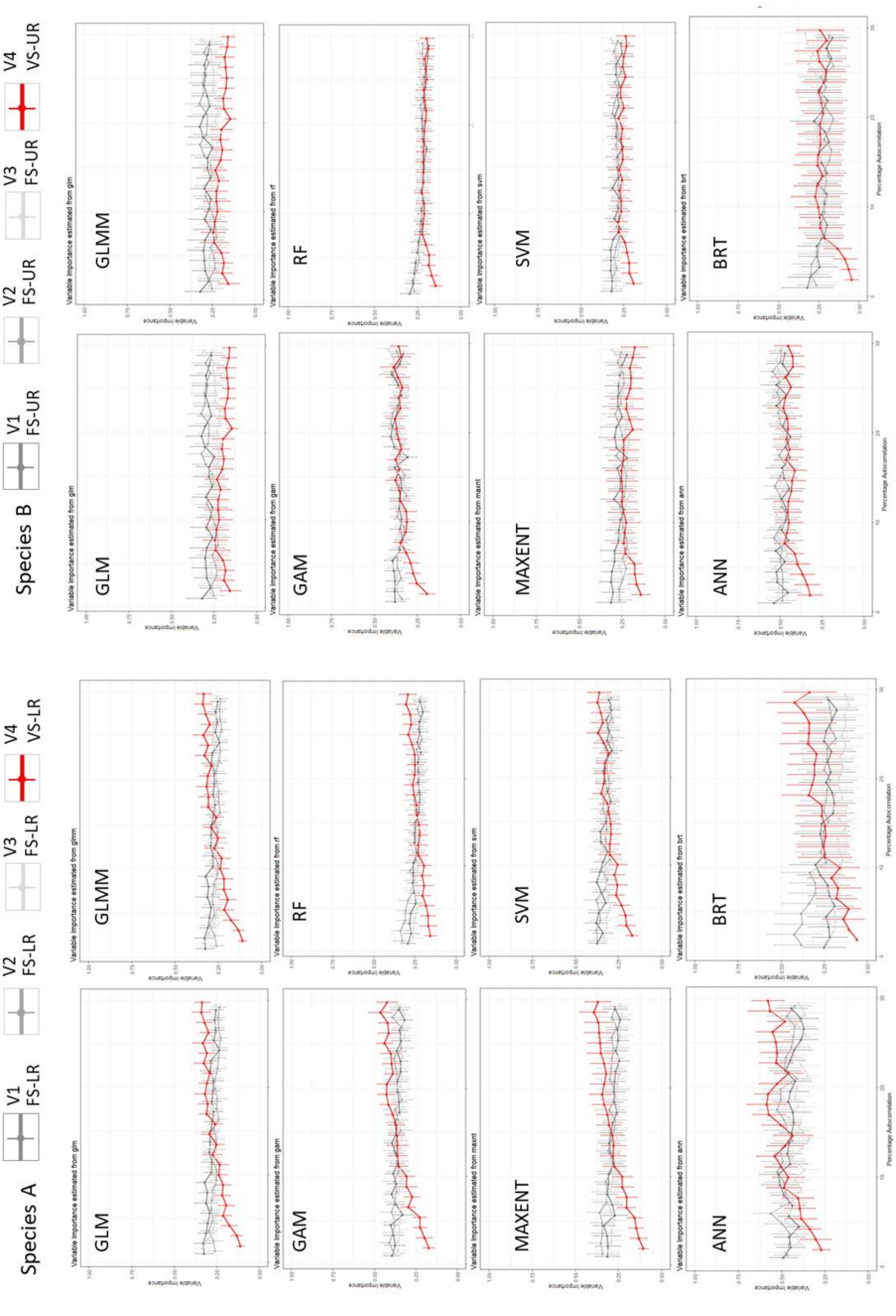
A1) SPECIES A, B (FS- Fixed SAC; VS- Varying SAC; LR- Linear response; US- Unimodal response)
Baseline autocorrelation for fixed SAC covariates= 0%



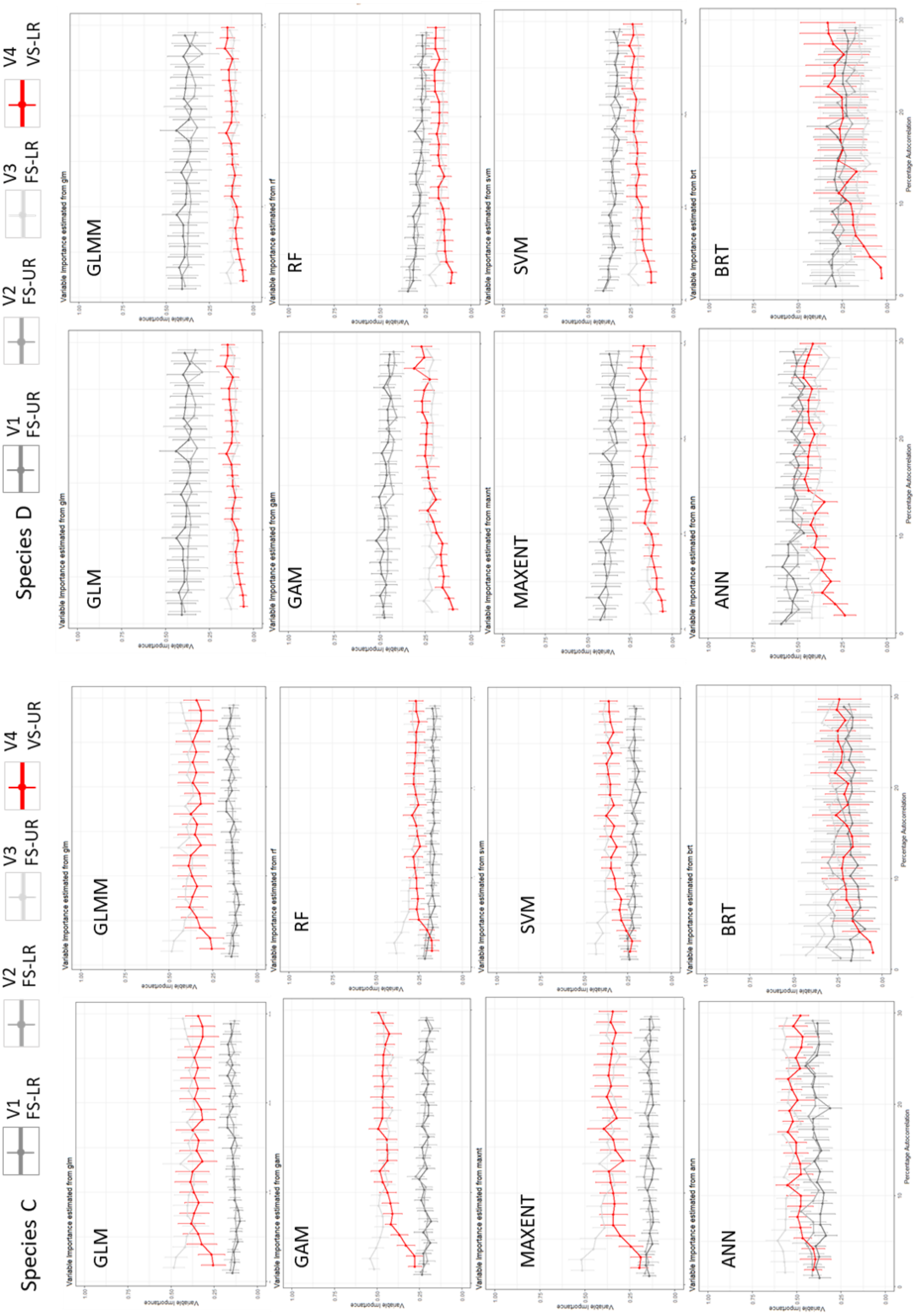
A2) SPECIES C, D (FS- Fixed SAC; VS- Varying SAC; LR- Linear response; US- Unimodal response)
Baseline autocorrelation for fixed SAC covariates= 0%



A3) SPECIES A, B (FS- Fixed SAC; VS- Varying SAC; LR- Linear response; US- Unimodal response)
Baseline autocorrelation for fixed SAC covariates=50 %



A4) SPECIES C, D (FS- Fixed SAC; VS- Varying SAC; LR- Linear response; US- Unimodal response)
Baseline autocorrelation for fixed SAC covariates=50 %



Appendix B – ANOVA values for different species and baseline SAC (0%,12.5%)

(green boxes represent insignificant ANOVA at 95% i.e. similar variable importance and grey otherwise)

B1) Anova values for SPECIES A and B; Baseline SAC = 0%

	SAC range	% Relative Morans I	SPECIES A								SPECIES B							
			GLM	GLMM	GAM	BRT	RF	MAXENT	SVM	ANN	GLM	GLMM	GAM	BRT	RF	MAXENT	SVM	ANN
Percentage Autocorrelation	0	0.0	0.18	0.16	0.29	0.77	0.89	0.81	0.87	0.51	0.00	0.00	0.06	0.71	0.55	0.03	0.25	0.12
	1.25	3.7	0.37	0.42	0.09	0.14	0.18	0.25	0.22	0.35	0.02	0.03	0.01	0.12	0.06	0.05	0.29	0.13
	2.5	9.0	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.24	0.57	0.79	0.06	0.09	0.00	0.17	0.14	0.00
	3.75	13.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.44	0.80	0.81	0.38	0.37	0.95	0.63	0.84	0.75
	5	16.7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.12	0.26	0.00	0.02	0.00	0.08	0.06	0.83
	6.25	19.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.09	0.11	0.00	0.00	0.00	0.00	0.00	0.49
	7.5	21.8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.97	0.89	0.00	0.00	0.00	0.00	0.00	0.34
	8.75	23.7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.27	0.15	0.00	0.00	0.00	0.00	0.18	0.34
	10	25.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.00	0.00	0.00	0.01	0.05	0.01
	11.25	26.6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.11	0.00	0.00	0.00	0.00	0.09	0.66
	12.5	27.7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.33	0.38	0.00	0.00	0.00	0.00	0.18	0.89
	13.75	28.7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.14	0.30	0.00	0.00	0.00	0.00	0.00	0.26
	15	29.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.17	0.11	0.00	0.00	0.00	0.00	0.01	0.05
	16.25	30.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.15	0.21	0.00	0.01	0.00	0.04	0.16	0.72
	17.5	30.8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.31	0.38	0.00	0.00	0.00	0.00	0.01	0.20
	18.75	31.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.14	0.00	0.00	0.00	0.00	0.00	0.25
	20	31.8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.73	0.74	0.00	0.00	0.00	0.00	0.00	0.04
	21.25	32.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.18	0.26	0.00	0.00	0.00	0.00	0.01	0.68
	22.5	32.7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.00	0.00	0.00	0.01	0.01	0.75
	23.75	33.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.46	0.63	0.00	0.01	0.00	0.00	0.03	0.32
	25	33.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.04	0.00	0.01	0.00	0.06	0.09	0.94
	26.25	33.6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.09	0.31	0.00	0.00	0.00	0.00	0.00	0.14
	27.5	33.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.33
	28.75	34.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.17	0.11	0.00	0.00	0.00	0.00	0.00	0.89
	30	34.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.49	0.60	0.00	0.00	0.00	0.00	0.01	0.53

B2) Anova values for SPECIES C and D; Baseline SAC = 0%

	SAC range	% Relative Morans I	SPECIES C								SPECIES D							
			GLM	GLMM	GAM	BRT	RF	MAXENT	SVM	ANN	GLM	GLMM	GAM	BRT	RF	MAXENT	SVM	ANN
Percentage Autocorrelation	0	0.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	1.25	3.7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	2.5	9.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	3.75	13.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	5	16.7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	6.25	19.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	7.5	21.8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	8.75	23.7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

	10	25.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	11.25	26.6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	12.5	27.7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	13.75	28.7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	15	29.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	16.25	30.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	17.5	30.8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	18.75	31.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	20	31.8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	21.25	32.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	22.5	32.7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	23.75	33.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	25	33.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	26.25	33.6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	27.5	33.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	28.75	34.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	30	34.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

B3) Anova values for SPECIES A and B; Baseline SAC = 12.5%

	SAC range	% Relative Morans I	SPECIES A								SPECIES B							
			GLM	GLMM	GAM	BRT	RF	MAXENT	SVM	ANN	GLM	GLMM	GAM	BRT	RF	MAXENT	SVM	ANN
Percentage Autocorrelation	0	-25.6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00
	1.25	-21.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01
	2.5	-16.6	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.19	0.00	0.00	0.01	0.03	0.01	0.01	0.01	0.01
	3.75	-12.4	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.26	0.00	0.00	0.01	0.20	0.01	0.01	0.01	0.01
	5	-8.9	0.12	0.12	0.09	0.23	0.05	0.02	0.04	0.02	0.12	0.12	0.11	0.83	0.16	0.37	0.04	0.25
	6.25	-6.1	0.02	0.02	0.01	0.50	0.09	0.06	0.12	0.87	0.02	0.02	0.96	0.99	0.86	0.43	0.60	0.69
	7.5	-3.8	0.10	0.10	0.14	0.52	0.34	0.25	0.09	0.28	0.10	0.10	0.67	0.91	0.85	0.70	0.91	0.19
	8.75	-1.9	0.10	0.10	0.18	0.12	0.05	0.12	0.03	0.69	0.10	0.10	0.66	0.96	0.89	0.79	0.57	0.71
	10	-0.4	0.26	0.26	0.66	0.40	0.78	0.86	0.68	0.06	0.26	0.26	0.45	0.82	0.87	0.95	0.91	0.43
	11.25	1.0	0.48	0.48	1.00	0.69	0.94	0.96	0.76	0.52	0.48	0.48	0.99	0.90	0.99	0.89	0.96	0.83
	12.5	2.1	0.59	0.54	0.82	0.76	0.78	0.91	0.51	0.19	0.59	0.54	0.44	0.93	0.98	0.23	0.28	0.50
	13.75	3.0	0.10	0.11	0.43	0.39	0.75	0.76	0.54	0.31	0.10	0.11	0.59	0.89	0.94	0.91	1.00	0.21
	15	3.8	0.35	0.35	0.88	0.66	0.98	0.81	0.97	0.47	0.35	0.35	0.70	0.86	0.84	0.91	0.72	0.21
	16.25	4.6	0.56	0.54	0.76	0.81	0.88	0.74	0.86	0.08	0.56	0.54	0.82	0.72	0.96	0.71	0.87	0.95
	17.5	5.2	0.23	0.29	0.13	0.18	0.40	0.22	0.31	0.06	0.23	0.29	0.50	0.81	0.65	0.45	0.57	0.37
	18.75	5.7	0.03	0.02	0.04	0.37	0.53	0.28	0.86	0.10	0.03	0.02	0.31	0.80	0.90	0.52	0.64	1.00
	20	6.2	0.38	0.37	0.24	0.58	0.89	0.45	0.85	0.35	0.38	0.37	0.53	0.50	0.61	0.13	0.95	0.84
	21.25	6.6	0.73	0.77	0.58	0.34	0.75	0.48	0.64	0.27	0.73	0.77	0.60	0.98	0.75	0.43	0.26	0.57
	22.5	7.0	0.67	0.71	0.30	0.58	0.80	0.31	0.85	0.14	0.67	0.71	0.57	0.92	0.77	0.50	0.69	0.82
	23.75	7.4	0.53	0.61	0.23	0.14	0.33	0.10	0.43	0.46	0.53	0.61	0.98	0.95	0.95	0.75	0.94	0.76
	25	7.7	0.31	0.25	0.03	0.23	0.06	0.15	0.21	0.23	0.31	0.25	0.36	0.86	0.62	0.45	0.25	0.19
	26.25	8.0	0.01	0.01	0.10	0.21	0.22	0.11	0.70	0.02	0.01	0.01	0.96	0.31	0.39	0.39	0.38	0.79

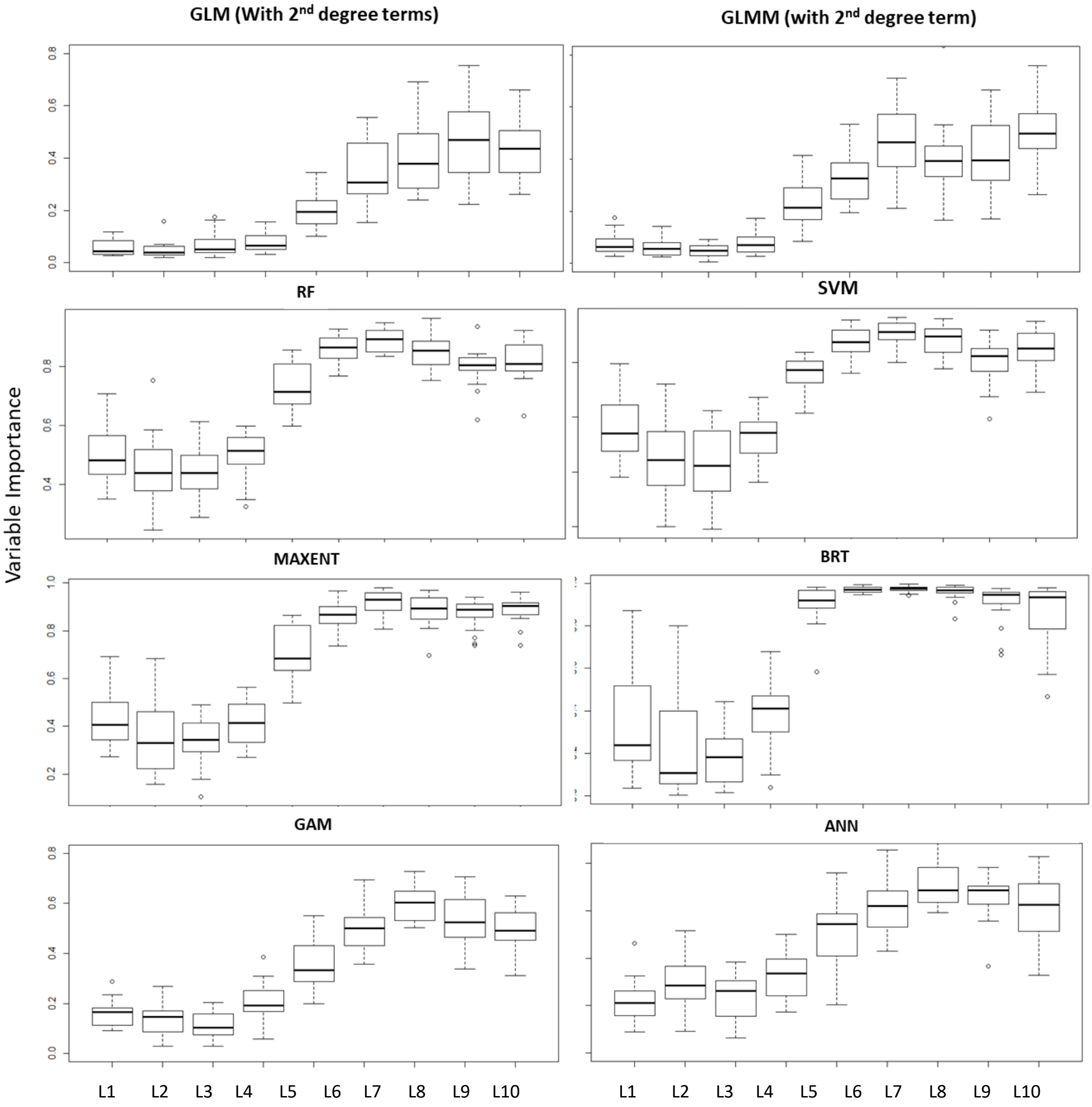
	27.5	8.2	0.24	0.22	0.04	0.08	0.06	0.03	0.22	0.39	0.24	0.22	0.99	0.50	0.61	0.62	0.55	0.04
	28.75	8.4	0.04	0.04	0.00	0.01	0.00	0.00	0.08	0.05	0.04	0.04	0.65	0.34	0.11	0.38	0.14	0.29
	30	8.7	0.02	0.01	0.15	0.21	0.04	0.03	0.33	0.05	0.02	0.01	0.14	0.20	0.04	0.05	0.07	0.80

B4) Anova values for SPECIES C and D; Baseline SAC = 12.5%

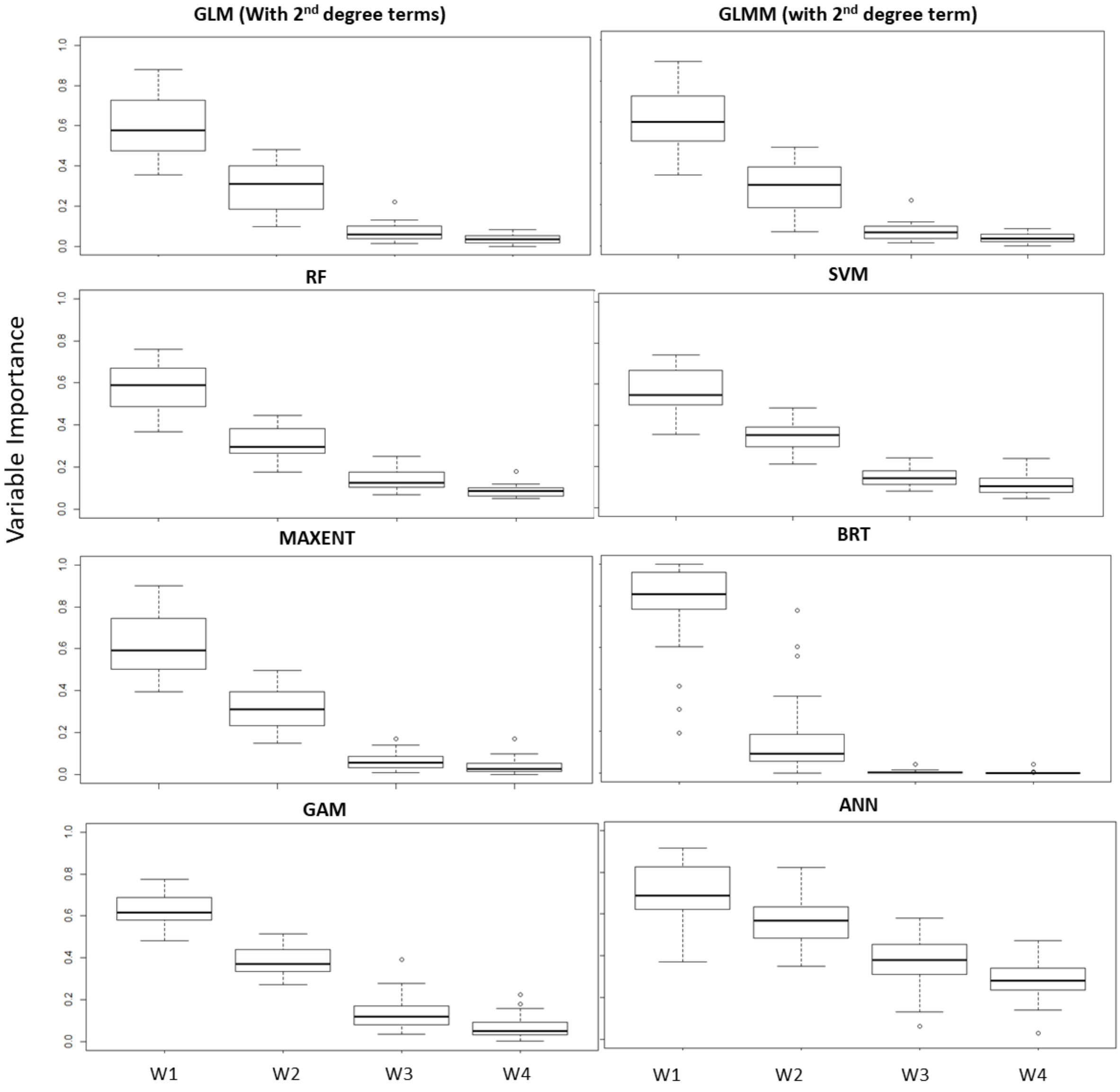
	SAC range	% Relative Morans I	SPECIES C								SPECIES D							
			GLM	GLMM	GAM	BRT	RF	MAXENT	SVM	ANN	GLM	GLMM	GAM	BRT	RF	MAXENT	SVM	ANN
Percentage Autocorrelation	0	-25.6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	1.25	-21.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	2.5	-16.6	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00
	3.75	-12.4	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.19	0.00	0.00	0.00	0.00
	5	-8.9	0.00	0.00	0.00	0.35	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00
	6.25	-6.1	0.00	0.00	0.00	0.12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.34	0.00	0.00	0.00	0.00
	7.5	-3.8	0.00	0.00	0.00	0.26	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.36	0.00	0.00	0.00	0.04
	8.75	-1.9	0.00	0.00	0.00	0.12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00
	10	-0.4	0.00	0.00	0.00	0.68	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.61	0.00	0.00	0.00	0.29
	11.25	1.0	0.00	0.00	0.00	0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.55	0.00	0.00	0.00	0.08
	12.5	2.1	0.00	0.00	0.00	0.39	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.21	0.00	0.00	0.00	0.00
	13.75	3.0	0.00	0.00	0.00	0.40	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.11
	15	3.8	0.00	0.00	0.00	0.79	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.47	0.00	0.00	0.00	0.04
	16.25	4.6	0.00	0.00	0.00	0.52	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.38	0.00	0.00	0.00	0.10
	17.5	5.2	0.00	0.00	0.00	0.52	0.00	0.00	0.00	0.13	0.00	0.00	0.00	0.35	0.00	0.00	0.00	0.02
	18.75	5.7	0.00	0.00	0.00	0.88	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.09
	20	6.2	0.00	0.00	0.00	0.52	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.68	0.00	0.00	0.00	0.02
	21.25	6.6	0.00	0.00	0.00	0.35	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.46	0.00	0.00	0.00	0.01
	22.5	7.0	0.00	0.00	0.00	0.51	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.19	0.00	0.00	0.01	0.39
	23.75	7.4	0.00	0.00	0.00	0.78	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.38	0.00	0.00	0.00	0.03
	25	7.7	0.00	0.00	0.00	0.73	0.00	0.00	0.00	0.15	0.00	0.00	0.00	0.52	0.00	0.00	0.00	0.17
	26.25	8.0	0.00	0.00	0.00	0.82	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.39	0.00	0.00	0.00	0.12
	27.5	8.2	0.00	0.00	0.00	0.26	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.44	0.00	0.00	0.00	0.01
	28.75	8.4	0.00	0.00	0.00	0.24	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.19	0.00	0.00	0.00	0.00
	30	8.7	0.00	0.00	0.00	0.80	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.12	0.00	0.00	0.00	0.03

Appendix C: Results from sub-experiments from all the models

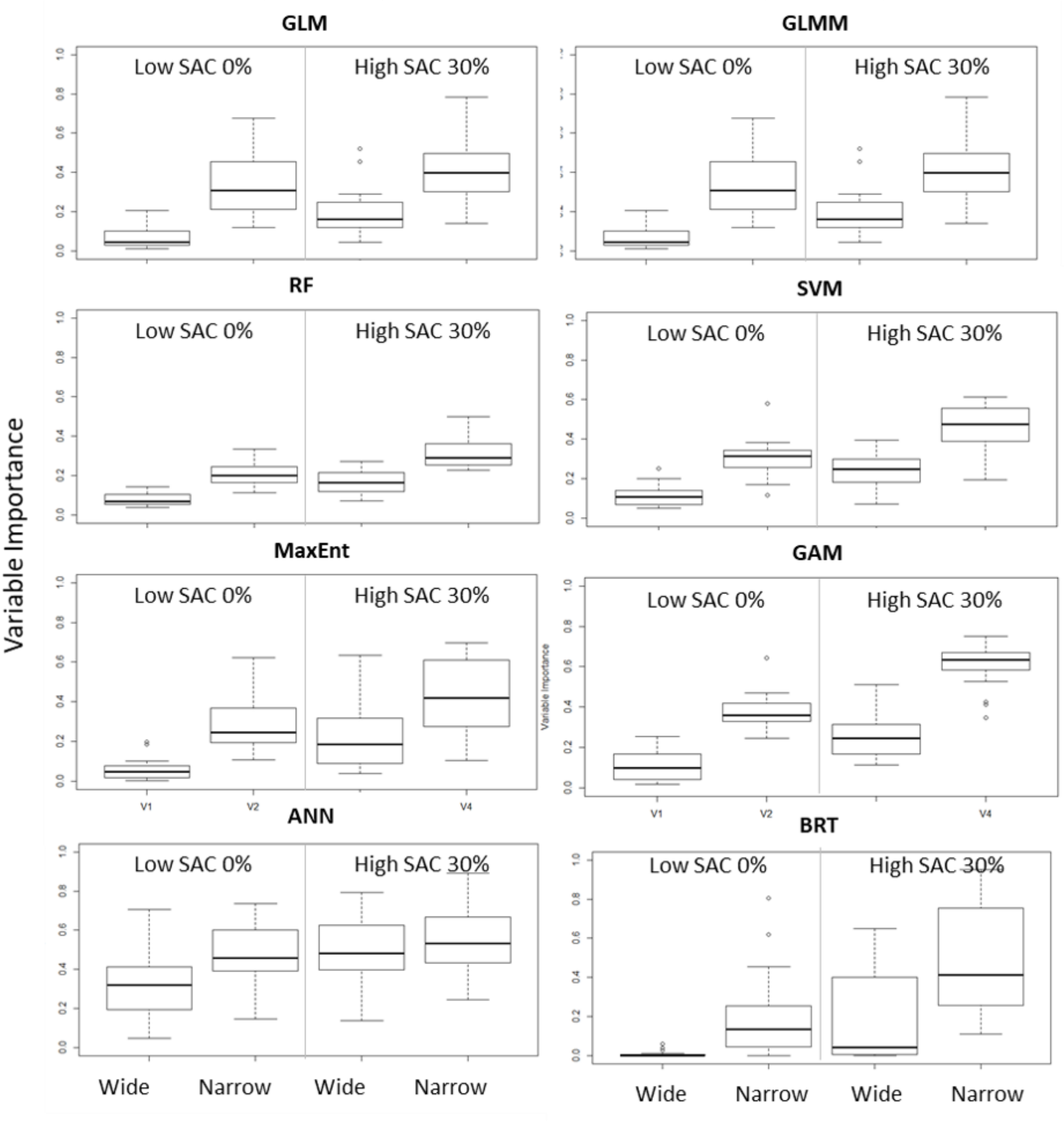
C1: Sub-experiment: Linear vs Unimodal



C2: Sub-experiment: Effect of changing widths



C3: Sub-experiment: Effect of the combination of changing widths and SAC extremes

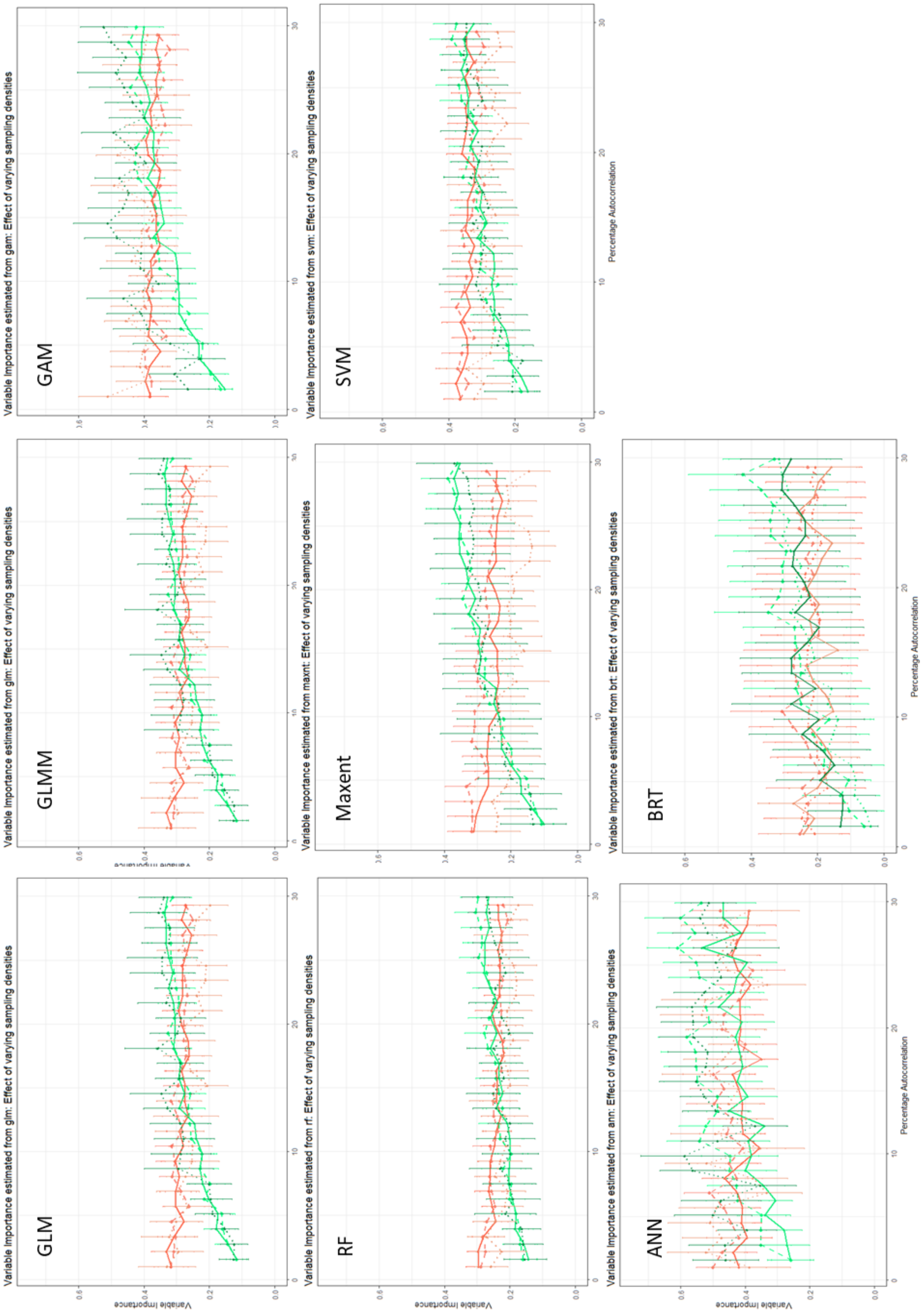


Appendix D – Variable importance plots across 3 sampling densities for all models for species A and B; Baseline SAC = 12.5%

D1) Species A

Green lines- Varying SAC variable; Red lines- Fixed SAC variable

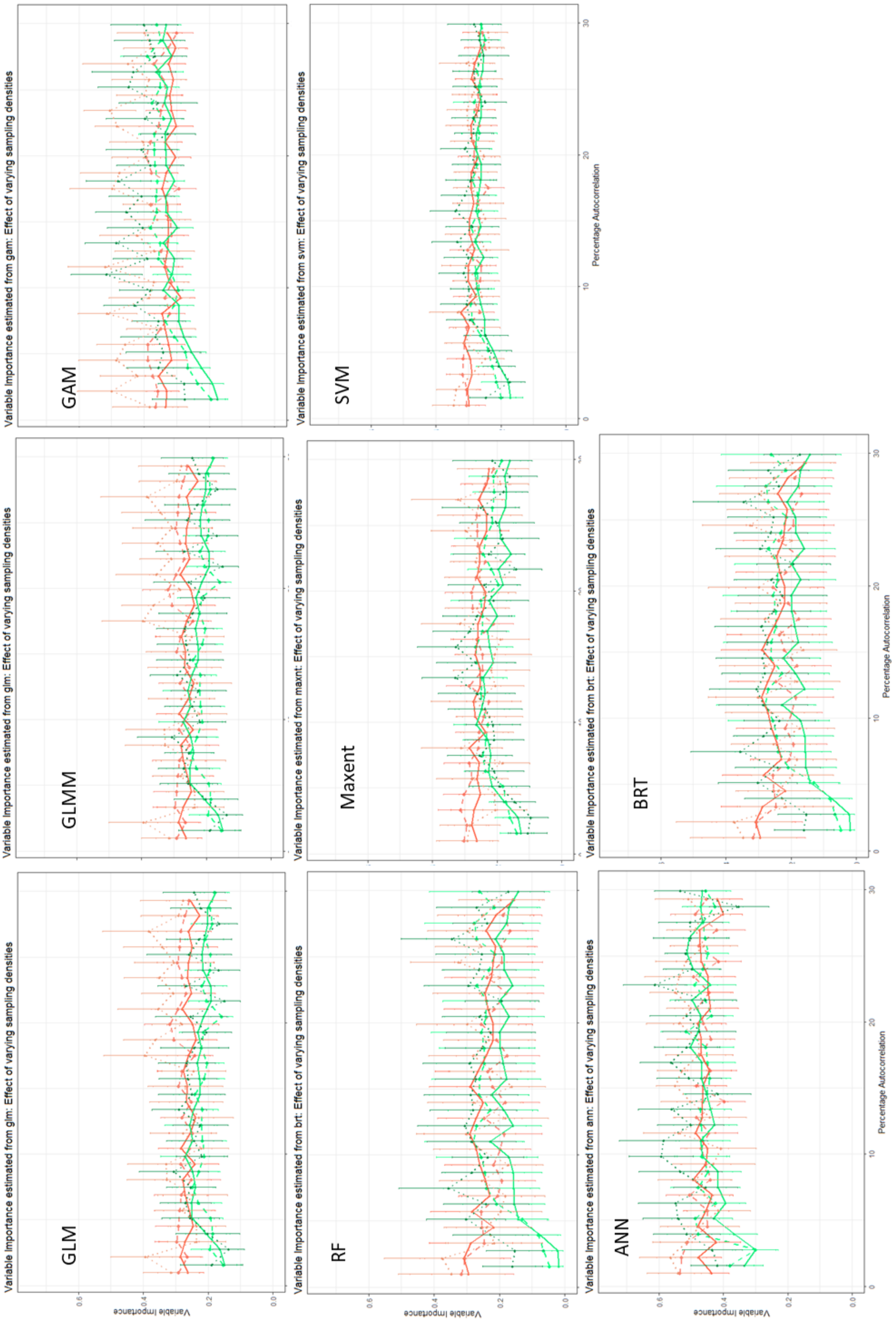
Dotted lines- 0.5% Sampling Density; Dashed lines- 2%; Normal lines- 3.5% Sampling Density



D2) Species B

Green lines- Varying SAC variable; Red lines- Fixed SAC variable

Dotted lines- 0.5% Sampling Density; Dashed lines- 2%; Normal lines- 3.5% Sampling Density



Appendix E: Tables showing accuracy metrics (AUC, Kappa) for different species and models.

E1: Table showing mean and standard deviation of AUC values, for three levels of autocorrelation, from all the eight models for Species A and Species B.

	Species A						*Species B*					
	Low SAC (0%)		Equal SAC (12.5%)		High SAC (30%)		Low SAC (0%)		Equal SAC (12.5%)		High SAC (30%)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
RF	0.89	0.04	0.92	0.02	0.91	0.02	0.90	0.02	0.91	0.03	0.92	0.03
SVM	0.91	0.04	0.93	0.02	0.93	0.02	0.89	0.02	0.90	0.03	0.91	0.03
ANN	0.85	NA	0.90	NA	0.89	NA	0.87	0.05	0.91	0.04	0.92	0.04
BRT	0.85	0.05	0.86	0.05	0.86	0.04	0.84	0.03	0.85	0.04	0.86	0.04
MAXENT	0.91	0.03	0.93	0.02	0.93	0.02	0.90	0.03	0.90	0.03	0.91	0.04
GAM	0.94	0.02	0.96	0.02	0.96	0.02	0.96	0.02	0.97	0.02	0.98	0.02
GLM	0.93	0.02	0.94	0.02	0.95	0.02	0.91	0.02	0.91	0.03	0.91	0.03
GLMM	0.93	0.02	0.94	0.02	0.95	NA	NA	NA	0.90	0.03	0.91	0.03

E2: Table showing mean and standard deviation of AUC values, for three levels of autocorrelation, from all the eight models for Species C and Species D.

	Species C						*Species D*					
	Low SAC (0%)		Equal SAC (12.5%)		High SAC (30%)		Low SAC (0%)		Equal SAC (12.5%)		High SAC (30%)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
RF	0.90	0.03	0.91	0.03	0.92	0.03	0.91	0.03	0.91	0.03	0.92	0.03
SVM	0.91	0.03	0.91	0.03	0.91	0.03	0.90	0.03	0.91	0.03	0.92	0.03
ANN	0.89	0.03	0.91	0.05	0.92	0.04	0.85	0.05	0.92	0.05	0.93	0.03
BRT	0.84	0.04	0.85	0.04	0.86	0.05	0.85	0.04	0.86	0.03	0.86	0.03
MAXENT	0.91	0.02	0.92	0.02	0.92	0.04	0.92	0.03	0.92	0.02	0.92	0.03
GAM	0.96	0.02	0.96	0.02	0.97	0.02	0.96	0.01	0.97	0.02	0.97	0.02
GLM	0.91	0.03	0.92	0.03	0.92	0.04	0.92	0.03	0.92	0.01	0.93	0.03
GLMM	0.90	0.06	0.92	0.03	0.92	0.03	0.92	0.03	0.92	0.01	0.91	0.04

E3: Table showing mean and standard deviation of Kappa values, for three levels of autocorrelation, from all the eight models for Species A and Species B.

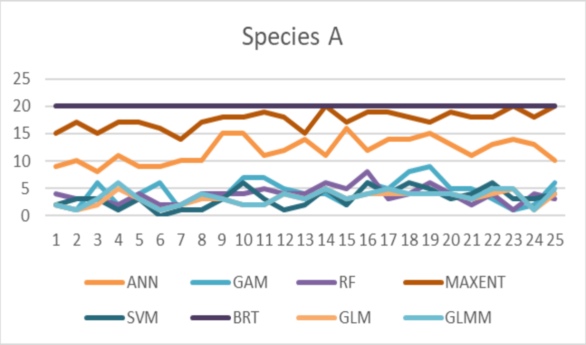
	Species A			*Species B*		
	Low SAC (0%)	Equal SAC (12.5%)	High SAC (30%)	Low SAC (0%)	Equal SAC (12.5%)	High SAC (30%)

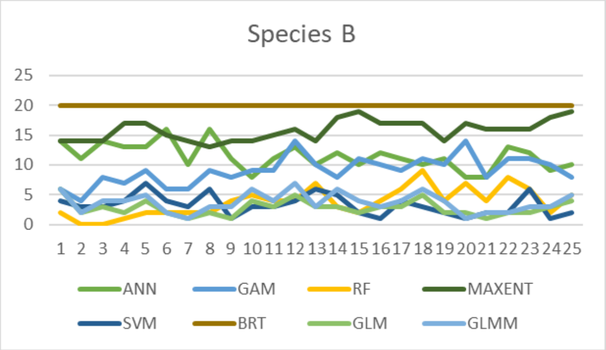
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
RF	0.99	0.02	0.99	0.02	0.98	0.03	0.98	0.02	0.99	0.01	0.99	0.01
SVM	0.75	0.06	0.76	0.05	0.78	0.06	0.72	0.05	0.74	0.07	0.74	0.04
ANN	0.60	NA	0.68	NA	0.69	NA	0.64	0.09	0.70	0.09	0.74	0.09
BRT	0.55	0.09	0.52	0.19	0.53	0.15	0.47	0.24	0.56	0.10	0.54	0.10
MAXENT	0.68	0.05	0.69	0.08	0.71	0.09	0.69	0.03	0.71	0.07	0.71	0.04
GAM	0.74	0.07	0.77	0.06	0.79	0.06	0.79	0.07	0.82	0.09	0.86	0.07
GLM	0.67	0.07	0.70	0.06	0.74	0.06	0.64	0.06	0.65	0.06	0.65	0.06
GLMM	0.67	0.08	0.70	0.06	0.74	0.06	0.63	0.05	0.64	0.06	0.65	0.06

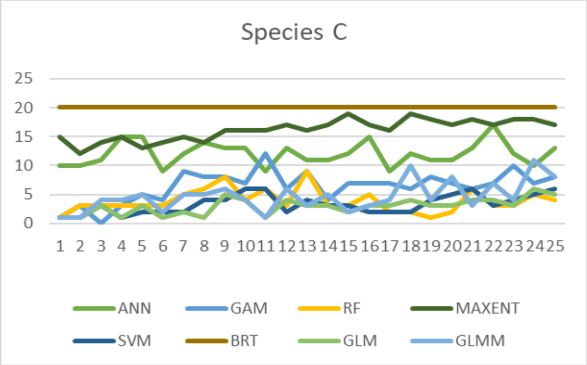
E4: Table showing mean and standard deviation of Kappa values, for three levels of autocorrelation, from all the eight models for Species C and Species D.

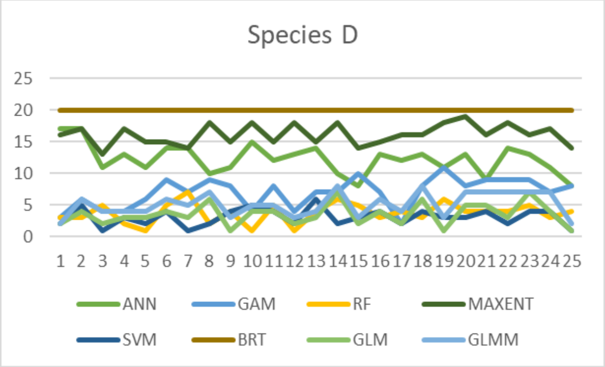
	Species C						*Species D*					
	Low SAC (0%)		Equal SAC (12.5%)		High SAC (30%)		Low SAC (0%)		Equal SAC (12.5%)		High SAC (30%)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
RF	0.98	0.02	0.99	0.01	0.99	0.01	1.00	0.01	0.98	0.02	0.98	0.02
SVM	0.75	0.05	0.74	0.07	0.75	0.05	0.74	0.06	0.74	0.06	0.74	0.04
ANN	0.67	0.08	0.71	0.12	0.74	0.10	0.61	0.10	0.72	0.10	0.76	0.07
BRT	0.47	0.19	0.57	0.12	0.54	0.10	0.54	0.12	0.56	0.10	0.50	0.16
MAXENT	0.67	0.06	0.71	0.08	0.75	0.06	0.72	0.05	0.72	0.05	0.72	0.06
GAM	0.79	0.06	0.80	0.10	0.82	0.07	0.79	0.05	0.81	0.08	0.81	0.07
GLM	0.66	0.06	0.68	0.08	0.67	0.09	0.66	0.07	0.67	0.04	0.68	0.08
GLMM	0.58	0.29	0.68	0.08	0.67	0.09	0.66	0.07	0.67	0.05	0.62	0.24

Appendix F – Plots for the number of times each model produces a significant autocorrelation in residuals as a function of the different SAC ranges (0-30%) at baseline SAC = 50%


Species A
ANN
GAM
RF
MAXENT
SVM
BRT
GLM
GLMM


Species B
ANN
GAM
RF
MAXENT
SVM
BRT
GLM
GLMM


Species C
ANN
GAM
RF
MAXENT
SVM
BRT
GLM
GLMM


Species D
ANN
GAM
RF
MAXENT
SVM
BRT
GLM
GLMM