Environmental Data Science (2022), 1: e30, 1–23
doi:10.1017/eds.2022.34

CAMBRIDGE
UNIVERSITY PRESS

METHODS PAPER

Variable ranking and selection with random forest for unbalanced data

Ute Bradter[1,2,*], John D. Altringham[1], William E. Kunin[1], Tim J. Thom[3], Jerome O'Connell[1,4] and Tim G. Benton[1]

[1]School of Biology, University of Leeds, Leeds, United Kingdom
[2]Department of Terrestrial Ecology, Norwegian Institute for Nature Research, Trondheim, Norway
[3]Yorkshire Wildlife Trust, Skipton, United Kingdom
[4]ProvEye, Kerry, Ireland
*Corresponding author. E-mail: Ute.bradter@nina.no

Received: 17 May 2021; **Revised:** 09 September 2022; **Accepted:** 09 November 2022

Keywords: Multi-scale; random forest; spatial scale; species distribution model; unbalanced data

Abstract

When one or several classes are much less prevalent than another class (unbalanced data), class error rates and variable importances of the machine learning algorithm random forest can be biased, particularly when sample sizes are smaller, imbalance levels higher, and effect sizes of important variables smaller. Using simulated data varying in size, imbalance level, number of true variables, their effect sizes, and the strength of multicollinearity between covariates, we evaluated how eight versions of random forest ranked and selected true variables out of a large number of covariates despite class imbalance. The version that calculated variable importance based on the area under the curve (AUC) was least adversely affected by class imbalance. For the same number of true variables, effect sizes, and multicollinearity between covariates, the AUC variable importance ranked true variables still highly at the lower sample sizes and higher imbalance levels at which the other seven versions no longer achieved high ranks for true variables. Conversely, using the Hellinger distance to split trees or downsampling the majority class already ranked true variables lower and more variably at the larger sample sizes and lower imbalance levels at which the other algorithms still ranked true variables highly. In variable selection, a higher proportion of true variables were identified when covariates were ranked by AUC importances and the proportion increased further when the AUC was used as the criterion in forward variable selection. In three case studies, known species–habitat relationships and their spatial scales were identified despite unbalanced data.

Impact Statement

Environmental data often contain many candidate covariates and unbalanced response variables. For example, when modeling species distributions of rare species, nondetections are often considerably more common than detections. While random forest is robust to situations with many covariates compared to data points, inference and prediction can be negatively affected by data imbalance, especially at the medium to small sample sizes common in environmental data. Using area under the curve (AUC) to calculate variable importances and as a criterion in variable selection improved the identification of true predictor variables when data were unbalanced. Therefore, at least when sample sizes are relatively small, we recommend the use of AUC in variable ranking and selection with random forest to improve inference and prediction.

This research article was awarded an Open Data badge for transparent practices. See the Data Availability Statement for details.

© The Author(s), 2022. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

1. Introduction

Random forest is a machine learning algorithm (Breiman, 2001) used for a variety of purposes, for example in genetics (Díaz-Uriarte and de Andrés, 2006), to predict the occurrence of storms (Ruiz-Gazen and Villa, 2007), biomass (Adam et al., 2014), wetland inundation (Karimi et al., 2019), the distribution of species (Garzón et al., 2006; Bradter et al., 2013; Robinson et al., 2018; Ryo et al., 2021), or to map vegetation from remotely sensed data (Sesnie et al., 2008; Bradter et al., 2011; O'Connell et al., 2015; Barrett et al., 2016; Bradter et al., 2020). Advantages of random forest include its often high classification accuracies (Garzón et al., 2006; Prasad et al., 2006; Cutler et al., 2007; Strobl et al., 2009; Sluiter and Pebesma, 2010) and its robustness to situations with many covariates relative to data points (Grömping, 2009). However, when one or several classes are much less prevalent than another class (unbalanced data), ranking and selection of covariates and classification with random forest can be problematic (Chen et al., 2004; Lin and Chen, 2012; Janitza et al., 2013).

Random forests use bagging and combine many regression or classification trees in an ensemble (Breiman, 2001). Each of the many trees in the forest is grown on a random selection, typically two-third, of the data. Additional randomization is introduced by restricting the available predictors at each node split to a random selection. This randomness produces a diverse ensemble of trees and can result in more accurate results (Breiman, 2001; Liaw and Wiener, 2002; Strobl et al., 2009). Each tree is used to predict the data not used in the construction of the tree, the out-of-bag (OOB) data. For each datapoint, the majority vote of all predictions is used to produce an average prediction in classification. The OOB error is the proportion of datapoints for which this average prediction is not the same as the true class (Liaw and Wiener, 2002). For inference with random forest, variable importance measures can be calculated. The permutation importance is calculated by randomly permuting each covariate in turn to destroy a potential association with the response variable. It is calculated as the difference in OOB error from the model with the permuted covariate compared to the OOB error from the model without permutations (Strobl et al., 2007). For a detailed description of random forest see, for example, Breiman (2001), Liaw and Wiener (2002), Grömping (2009), and Strobl et al. (2009).

Unbalanced data are widespread in ecology and environmental science and can affect variable importance measures and the accuracy of classification results. For example, in remote sensing of vegetation, the rarer vegetation classes tend to be much less prevalent in the data (Bradter et al., 2020). Species distributions are often mapped using reported detections and non-detections of the species at sample locations collected by survey schemes (Franklin, 2009). For rarer species, it is common to have considerably more non-detections than detections. The majority class (e.g., non-detections) tends to be predicted with a higher accuracy than the minority class (e.g., detections) in random forest (Chen et al., 2004; Lin and Chen, 2012) and permutation importances are biased at higher imbalance levels (Janitza et al., 2013). This can be problematic, for example for conservation and land management, because targeted conservation measures depend on predicting the occurrence of the rare species or vegetation types accurately and on accurate inference of the processes that lead to the observed pattern.

Several adaptations have been proposed to improve the performance of random forest with unbalanced data. However, the adaptations have frequently been evaluated regarding their ability to discriminate between classes and it remains unclear how variable importance measures and therefore the ranking and selection of covariates are affected. Adaptations to improve the performance with unbalanced data include (a) resampling of data to improve class balance, (b) applying weights, (c) applying a tree-splitting criterion that is robust to data imbalance, and (d) changing the threshold for the assignment of predicted classes. Resampling strategies to improve class balance include to sample only a proportion of the data randomly from the majority class for each classification tree in the ensemble (downsampling) or to oversample the minority class (balanced random forest; Chen et al., 2004; Liu et al., 2006; Xu-Ying et al., 2009; Lin and Chen, 2012). Both downsampling and upsampling improved the discrimination between unbalanced classes compared to not applying a resampling strategy (Chen et al., 2004). Downsampling outperformed upsampling, but may result in loss of information (Chen et al., 2004; Liu et al., 2006). Applying weights at node splits and in majority voting has improved the discrimination of classes when data were imbalanced (Chen et al., 2004; Min et al., 2018). In contrast to the tree splitting criteria Gini index and Gain Ratio, the Hellinger distance is not biased toward the majority class and the discrimination between classes can improve when the Hellinger distance is used (Cieslak et al., 2012; Aler et al., 2020). Discrimination between unbalanced classes was also improved when the threshold used to distinguish between classes was optimized (Dahinden, 2011; Freeman et al., 2012). Studies that evaluated several of the adaptations against each other, reported contrasting results. Robinson et al. (2018) used spatially biased and imbalanced data consisting of detection and nondetection data of a bird species. After first reducing the spatial bias in the majority class by filtering data based on location, they found that random forest without adjustment for unbalanced data (default random forest henceforth), weighted random forest, balanced random forest, and a more sophisticated resampling strategy, oversampling using synthetic samples (SMOTE), discriminated accurately between detections and non-detections. However, their sample size was large and despite class imbalance, the number of cases of the minority class was more than 1,000. With smaller sample sizes, Chen et al. (2004) found little difference between balanced and weighted random forests, but both discriminated better between classes than the default random forest. On presence-background data for 225 species with varying sample size, downsampled random forests and versions that combined the Hellinger splitting criterium with shallower trees were better able to discriminate between classes compared to the default random forest or weighted random forest (Valavi et al., 2021a). The same study also evaluated random forest based on regression trees applied to binary data. We do not consider this version further as it was among the poorer performers, and because binary response variables are not a good fit for a method designed for continuous response variables. Downsampled random forests were also among the algorithms with the highest discrimination for unbalanced classes when compared to methods other than random forest (Valavi et al., 2021b).

Adaptations to improve variable ranking or selection with unbalanced data are based on the AUC (area under the receiver operating curve), which is insensitive to unbalanced data (Fawcett, 2006). AUC has been used for the calculation of permutation importances (Janitza et al., 2013), thus affecting the order in which covariates will be ranked for variable selection. AUC has also been used as a criterion in a backward variable selection, however, the ranking of covariates was based on the default version of random forest (Urrea and Calle, 2012).

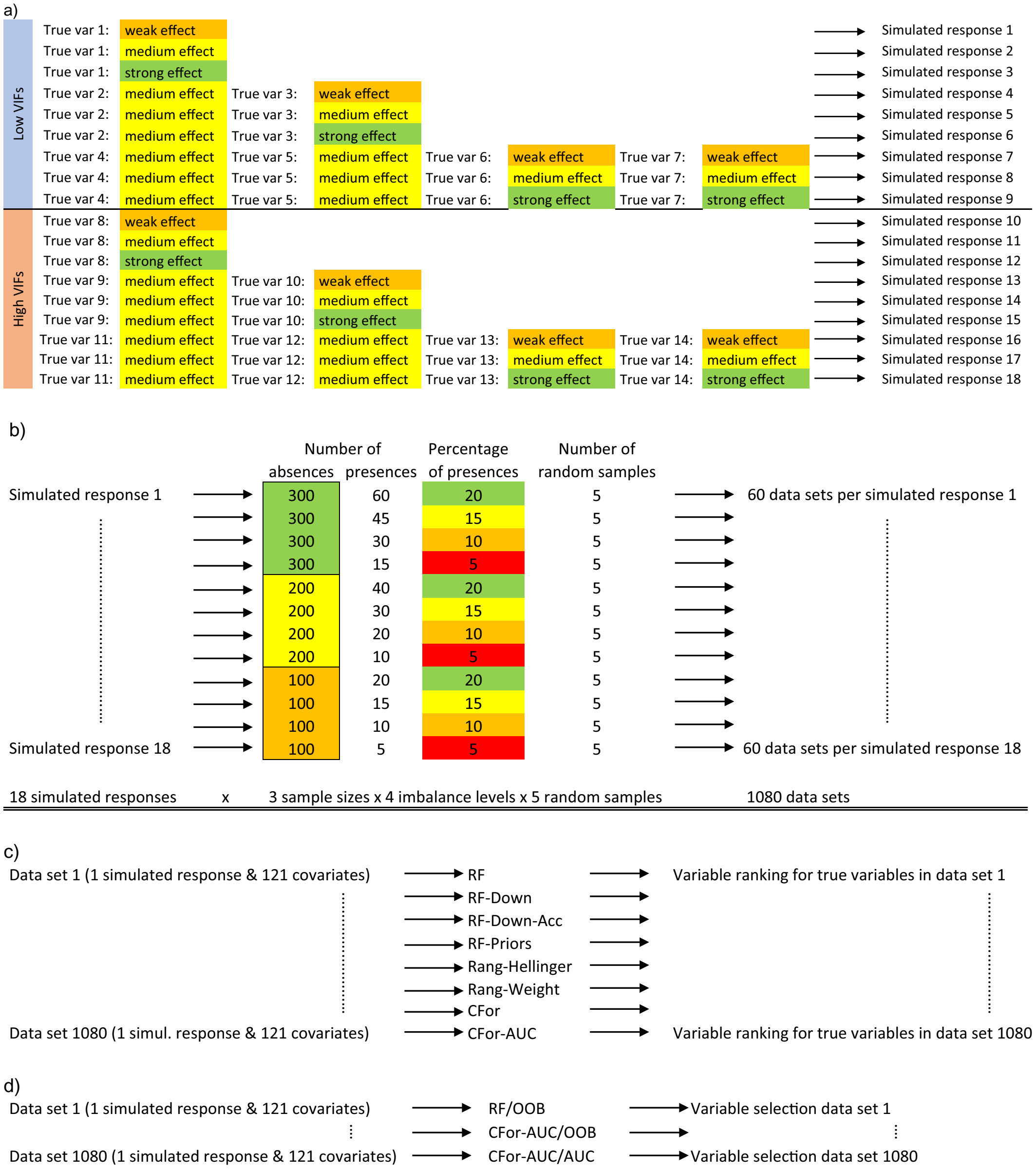
The aim of our study was to evaluate the influence of alternative versions of random forest for unbalanced data on variable ranking and variable selection. Using simulated data, we evaluated (a) how highly true covariates were ranked by alternative random forest versions based on variable importances and (b) how well alternative random forest versions were able to identify true covariates out of a large number of covariates despite data imbalance. Subsequently, we present three case studies from the field of species distribution modeling using data with a varying degree of imbalance. We based our study on small to medium sample sizes with imbalance levels of 20% and higher and with weak, medium, and strong associations between the response variables and covariates. Under these conditions, the biases due to unbalanced data are most evident. The performance of variable importances and class discrimination typically declines with increasing class overlap, increasing imbalance level, and decreasing sample size (Lin and Chen, 2012; Janitza et al., 2013). The default random forest produces good discrimination between classes when class distributions are well separated regardless of the class imbalance (Lin and Chen, 2012; Valavi et al., 2021a).

2. Methods

Our simulations and case studies were based on data sets with many covariates relative to data points. An advantage of random forest is its robustness to situations with many covariates to data points. Consequently, it is particularly valuable to include such data sets in evaluations.

2.1. *Simulation study*

Using simulated data, we compared how alternative random forest approaches ranked and were able to identify true covariates despite data imbalance. First, we generated 1,080 simulated data sets varying in size, imbalance level, number of true variables, their effect sizes, and the strength of multicollinearity between true and noise variables (Figure 1a,b). Based on 121 real-world environmental covariates, we



a)

Low VIFs	True var 1:	weak effect							→ Simulated response 1
	True var 1:	medium effect							→ Simulated response 2
	True var 1:	strong effect							→ Simulated response 3
	True var 2:	medium effect	True var 3:	weak effect					→ Simulated response 4
	True var 2:	medium effect	True var 3:	medium effect					→ Simulated response 5
	True var 2:	medium effect	True var 3:	strong effect					→ Simulated response 6
	True var 4:	medium effect	True var 5:	medium effect	True var 6:	weak effect	True var 7:	weak effect	→ Simulated response 7
	True var 4:	medium effect	True var 5:	medium effect	True var 6:	medium effect	True var 7:	medium effect	→ Simulated response 8
	True var 4:	medium effect	True var 5:	medium effect	True var 6:	strong effect	True var 7:	strong effect	→ Simulated response 9
High VIFs	True var 8:	weak effect							→ Simulated response 10
	True var 8:	medium effect							→ Simulated response 11
	True var 8:	strong effect							→ Simulated response 12
	True var 9:	medium effect	True var 10:	weak effect					→ Simulated response 13
	True var 9:	medium effect	True var 10:	medium effect					→ Simulated response 14
	True var 9:	medium effect	True var 10:	strong effect					→ Simulated response 15
	True var 11:	medium effect	True var 12:	medium effect	True var 13:	weak effect	True var 14:	weak effect	→ Simulated response 16
	True var 11:	medium effect	True var 12:	medium effect	True var 13:	medium effect	True var 14:	medium effect	→ Simulated response 17
	True var 11:	medium effect	True var 12:	medium effect	True var 13:	strong effect	True var 14:	strong effect	→ Simulated response 18

b)

	Number of		Percentage	Number of	
	absences	presences	of presences	random samples	
Simulated response 1 →	300	60	20	5	→ 60 data sets per simulated response 1
→	300	45	15	5	→
→	300	30	10	5	→
→	300	15	5	5	→
→	200	40	20	5	→
→	200	30	15	5	→
→	200	20	10	5	→
→	200	10	5	5	→
→	100	20	20	5	→
→	100	15	15	5	→
→	100	10	10	5	→
Simulated response 18 →	100	5	5	5	→ 60 data sets per simulated response 18

18 simulated responses x 3 sample sizes x 4 imbalance levels x 5 random samples 1080 data sets

c)

Data set 1 (1 simulated response & 121 covariates) → RF → Variable ranking for true variables in data set 1

→ RF-Down →

→ RF-Down-Acc →

→ RF-Priors →

→ Rang-Hellinger →

→ Rang-Weight →

→ CFor →

Data set 1080 (1 simul. response & 121 covariates) → CFor-AUC → Variable ranking for true variables in data set 1080

d)

Data set 1 (1 simulated response & 121 covariates) → RF/OOB → Variable selection data set 1

→ CFor-AUC/OOB →

Data set 1080 (1 simulated response & 121 covariates) → CFor-AUC/AUC → Variable selection data set 1080

Figure 1. *Conceptual diagram of the simulation study: (a) Out of a set of 121 environmental covariates, 14 variables were selected as true variables (true var). Seven of the 14 variables had low variance inflation factors (VIFs) indicating low multicollinearity with other covariates and the other seven variables had high variance inflation factors. True variables were standardized to a mean of 0 and a standard deviation of 1. Eighteen binary, simulated response variables were created based on weak (ß = 0.4), medium (ß = 0.9), and strong (ß = 1.4) effect sizes for true variables and either one, two, or four true variables with either low or high VIFs. (b) From each simulated response variable consisting of presences and absences, 60 datasets were created differing in the sample size (300, 200, or 100 absences) and the imbalance level (20, 15, 10, or 5% presences). Presences and absences were randomly drawn five times for each of the 12 combinations of sample size with imbalance level. In total, 1,080 data sets were created (18 simulated responses × 3 sample sizes × 4 imbalance levels × 5 random samples). (c) For each of the 1,080 datasets consisting of simulated* *(Continued)*

presences and absences and the corresponding 121 environmental covariates, variable importances were calculated for each of the 121 covariates using eight alternative random forest versions: four random forest versions were implemented with R package randomForest (RF, RF-Down, RF-Down-Acc, RF-Priors), two versions were implemented with R package ranger (Rang-Hellinger, Rang-Weight) and two versions with R package party (CFor, CFor-AUC): (1) default RF (RF), (2) the majority class downsampled to the size of the minority class (RF-Down), (3) the majority class downsampled to 64% of the size of the minority class (RF-Down-Acc), (4) weights as class priors (RF-Priors), (5) Hellinger distance as splitting criterion (Rang-Hellinger), (6) weights applied to the splitting rule and majority vote (Rang-Weight), (7) random forest based on conditional inference trees (CFor), (8) random forest based on conditional inference trees with AUC permutation importance (CFor-AUC). After the calculation of variable importances, the covariates were ranked from the covariate with the highest importance to the covariate with the lowest importance and the ranks for true variables were extracted. (d) For each of the 1,080 datasets we carried out variable selection using three alternative versions: (1) Variable ranking based on permutation importances from the default RF, followed by a forward selection with the Out-of-Bag error as selection criterion (RF/OOB), (2) Variable ranking based on AUC permutation importance and forests with conditional inference trees, followed by a forward selection with the Out-of-Bag error as selection criterion (CFor-AUC/OOB), and (3) Variable ranking based on AUC permutation importance and forests with conditional inference trees, followed by a forward selection with the AUC as selection criterion (CFor-AUC/AUC).

generated simulated binary data. We named the minority class presences and the majority class absences, thus simulating data frequently used in the field of species distribution modeling (Franklin, 2009). From the 121 covariates, we chose covariates with either low or high multicollinearity to other covariates as true variables. To estimate multicollinearity, we calculated variance inflation factors (VIFs, Zuur et al., 2009). The simulated occurrences were generated by drawing from the Bernoulli distribution with probability p, the probability of the simulated occurrence being a presence. p was calculated as the inverse logit of the linear predictor $\alpha + \beta_1 \times x_1 + \dots \beta_n \times x_n$, where $x_{1\text{-}n}$ were either one, two, or four true variables with low VIFs or one, two, or four true variables with high VIFs. For each of the six combinations of true variables, we chose regression coefficients $\beta_{1\text{-}n}$, that represented weak ($\beta = 0.4$), medium ($\beta = 0.9$), or strong ($\beta = 1.4$) effects, generating 18 simulated responses in total (Figure 1a). The true variables were standardized to a mean of 0 and a standard deviation of 1. Next, we randomly sampled to create 60 data sets from each of the simulated responses (Figure 1b). The data sets varied in sample size (300, 200, 100 absences) and the level of imbalance (20, 15, 10, 5% presences). For each of the 12 combinations of sample size with imbalance level, we randomly sampled five times, generating 1,080 simulated data sets (Figure 1b: 18 simulated responses × 3 sample sizes × 4 imbalance levels × 5 random samples = 1,080).

2.2. Variable ranking

For each of the 1,080 data sets, we evaluated how highly the true predictors were ranked in comparison to the other covariates by eight alternative algorithm versions (Figure 1c). Four random forest versions were implemented with the R package “randomForest” (Liaw and Wiener, 2002). This function implements the original programme by Breiman and Cutler (Liaw and Wiener, 2002) and forests are based on CART trees (Breiman, 2001; Strobl et al., 2009). CART trees use the Gini split criterion, which measures the decrease in node impurity, to select the predictor for a node split. Subset sampling for individual trees is with replacement by default (Liaw and Wiener, 2002). Two random forest versions were implemented with the R package “ranger,” a fast implementation of random forests (Wright and Ziegler, 2017). Two versions were implemented with function cforest from R package “party” (Strobl et al., 2007; Strobl et al. 2008). Random forest with cforest are based on conditional inference trees with predictors for node split selected based on a conditional inference independence test (Strobl et al., 2007). Variable importances based on cforest optionally allow for the calculation of a conditional version of variable importances (Strobl et al., 2008). Although this may be suitable for our data, which contained correlated predictor variables, we did not use it to avoid confounding our comparisons with the previous six algorithm versions. The following eight versions were evaluated:

1. RF: Random forest with no adjustment for unbalanced data: function randomForest in package randomForest.
2. RF-Down: For each tree in the ensemble, we randomly deselected data from the majority class to match the size of the minority class using option sampsize in function randomForest.
3. RF-Down-Acc: Downsampling the majority class to the size of the minority class does not necessarily result in the best performance (Ruiz-Gazen and Villa, 2007). We downsampled the majority class to 64% of the size of the minority class, which had produced good error rates for the minority class on preliminary simulated data.
4. RF-Priors: We provided priors for the classes using the argument classwt of function randomForest (package randomForest). We used the proportion of the absences in the whole data set as the prior for the absence class, and the proportion of presences as the prior for the presence class. To explore the variability in output due to variation in priors, we additionally reversed the order and supplied the proportion of presences as the prior for the absence class and the proportion of absences as the prior for the presence class.
5. Rang-Hellinger: We used the Hellinger distance as the splitting rule in function ranger of package ranger.
6. Rang-Weight: We used the argument class.weights of function ranger to apply weights for the classes in the splitting rule and in the majority voting. We used the proportion of the absences in the whole data set as the weight for the presence class, and the proportion of presences as the weight for the absence class. To explore the variability in output due to variation in weights, we additionally reversed the order and supplied the proportion of presences as the weight for the presence class and the proportion of absences as the weight for the absence class.
7. CFor: We applied random forest based on conditional inference trees with function cforest in package party.
8. CFor-AUC: We calculated AUC permutation importance using function varImpAUC in package “varImp” (Probst, 2020) in conjunction with function cforest from package party. Function varImpAUC uses AUC instead of OOB in the calculation of permutation importance (Janitza et al., 2013). The permutation importance is calculated as the difference in AUC before and after permutation of each covariate in turn. Permutation importances are therefore expected to be more insensitive to the imbalance level.

For the tuning parameters of random forest, ntree (the number of trees in a forest) and mtry (the number of variables tried at each node split), we used ntree = 2,000 and mtry = $p/2$ following Genuer et al. (2010), where p is the number of predictors. Higher ntree settings result in less variability and more stable variable importances and OOB error (Liaw and Wiener, 2002; Díaz-Uriarte and de Andrés, 2006; Genuer et al., 2010). The default value in the R package random forest is ntree = 500 (Liaw and Wiener, 2002). With higher mtry values, the permutation importance becomes more conditional; with lower mtry values the permutation importance becomes more marginal (Strobl et al., 2008; Grömping, 2009). The default value for classification in the R package random forest is mtry = $\sqrt{p}$ (Liaw and Wiener, 2002). For each of the 1,080 test sets and each of the eight algorithms, we repeated the calculation of variable importance measures 10 times. We ranked all 121 covariates based on the average variable importances from the 10 repetitions.

2.3. Variable selection

We evaluated how well variable selection identified true predictors out of all 121 covariates for three algorithm versions. We implemented the variable selection approach of Genuer et al. (2010), which consists of a forward selection of variables ranked by their variable importances resulting in a subset of important variables inclusive of some redundant variables. Optionally, the redundant variables can be reduced in an additional step, which we did not implement. The variable selection strategy has been described as a prediction-oriented strategy accepting a higher risk of false negatives, or important covariates that are not in the final selection set (Genuer et al., 2015). The variable selection consisted of the following steps for the default version without adjustments to account for unbalanced data: First, all covariates were ranked by their permutation importance, averaged over 50 repetitions. Second, we calculated OOB error for the model with the highest ranked covariate, averaged over 25 repetitions for each set. We added the second highest ranked covariate, recalculated OOB error and repeated this procedure until all covariates were included. Third, we identified the model with the lowest mean OOB error and added its standard deviation. Then we selected the smallest (fewest covariates) model with an OOB error less than this value. We used the tuning parameters suggested by Genuer et al. (2010): for the calculation of variable importances, we used ntree = 2,000 and mtry = $p/2$, where p equals the number of covariates and for the calculation of OOB error we used the default ntree and mtry values of function randomForest. To increase computational efficiency we also followed their recommendation for initial removal of the lowest ranked covariates, based on high standard deviations of permutation importances relative to small mean permutation importance per covariate. For full details of the variable selection strategy, see Genuer et al. (2010).

We evaluated the following three approaches (Figure 1d):

1. RF/OOB: Permutation importances were calculated with the default RF.
2. CFor-AUC/OOB: AUC permutation importances were calculated based on the cforest implementation.
3. CFor-AUC/AUC: AUC permutation importances were calculated based on the cforest implementation of random forest. Instead of using OOB as the selection criterium in the forward selection, AUC was used. In other words, instead of calculating OOB for each nested model, we calculated AUC. Then, analogue to the selection method based on OOB, we identified the model with the highest mean AUC, subtracted its standard deviation from the mean AUC value, then selected the smallest model with a mean AUC larger than this value as the final model (Figure 2).

For each selected model, we assessed how well true predictors were identified out of all 121 covariates using sensitivity and specificity. Sensitivity is higher the more true predictors are correctly identified (true positives) and is calculated as:

$$\frac{\text{true positives}}{\text{true positives} + \text{false negatives}},$$

where false negatives are true predictors that were not selected. Specificity is higher the more covariates which are not the true predictors are identified as such (true negatives) and was calculated as:

$$\frac{\text{true negatives}}{\text{false positives} + \text{true negatives}},$$

where false positives are covariates falsely identified as true predictors (Lin and Chen, 2012).

2.4. Case studies

We present studies from the field of species distribution modeling as case studies. Species occurrence data from sample sites are modeled as a function of environmental information in species distribution models. Subsequently, maps of the occurrence of the species can be produced based on the species-environment associations established in the models (Franklin, 2009). Species distribution modelers are often encouraged to carefully select suitable covariates based on ecological knowledge of the species to avoid spurious associations. However, the ecology of many species is poorly studied and understood (Dormann et al., 2012, Urban et al., 2016), which carries the risk that important processes are inadvertently omitted or

https://doi.org/10.1017/eds.2022.34 Published online by Cambridge University Press

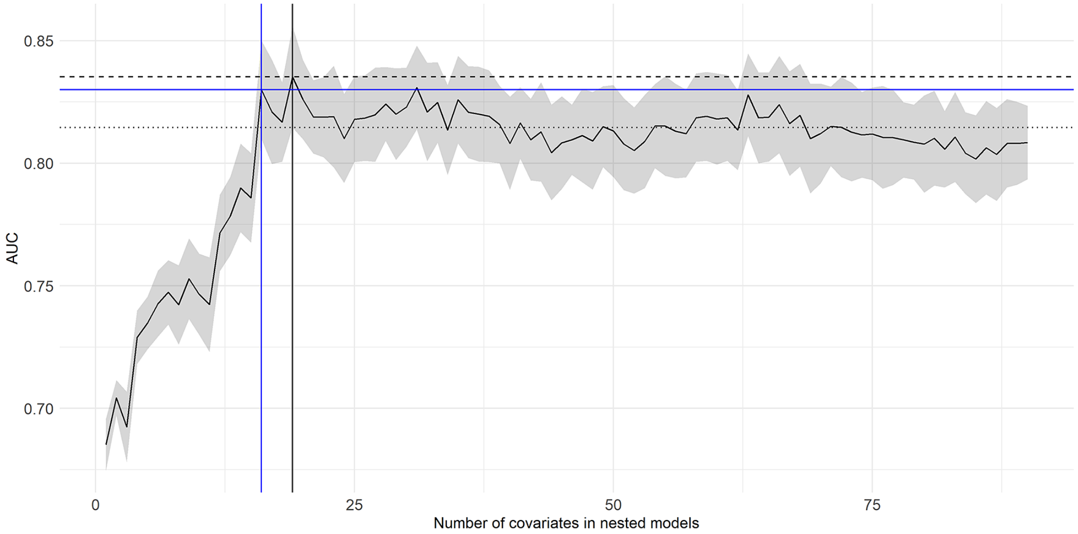


Figure 2. *Example of model selection using AUC. First, AUC is calculated for all nested models using several repetitions for each nested model. The black curve shows the mean AUC over all repetitions with the gray polygon showing the mean ± standard deviation. Then, the model (vertical black line) with the highest mean AUC (dashed horizontal black line) is identified. The AUC value corresponding to mean–standard deviation for this model (dotted horizontal black line) serves as a threshold accounting for the variation between repetitions. The selected model is the smallest (lowest number of covariates) model with a mean AUC at or above the threshold value (blue lines).*

misspecified in models. A key consideration in ecology is that each ecological process has an appropriate spatial scale and that appropriate spatial scales are dependent on the organism under consideration (Wiens, 1976; Addicott et al., 1987). This scale of effect is the scale at which the association between a species occurrence and an environmental covariate is strongest (Jackson and Fahrig, 2012). Another key consideration is that ecological processes can act at multiple spatial scales and influence species at individual, group, or population level (Johnson, 1980; Wiens, 1989; Levin, 1992; Cunningham and Johnson, 2006; McGarigal et al., 2016). A covariate influencing a species' distribution at one scale (e.g., surface water availability within 500 m of a species' nest site) may exert a different or no influence at a different scale (e.g., within 10 km; Wiens, 1989, Hamer and Hill, 2000). Including covariates at appropriate spatial scales in ecological models is therefore important. However, if the spatial process scales are unknown, appropriate spatial scales of covariates have to be selected empirically out of a potentially large pool of covariate × scale combinations. Random forest is a candidate algorithm for such problems due to its ability to assess species-environment associations with all covariate × scale combinations simultaneously in one model (Bradter et al., 2013). Such models may generate new hypotheses about the ecological processes determining the distribution of species.

2.5. Study species and data collection

As case studies, we used the wader species (shorebirds in North America) Northern lapwing (*Vannellus vanellus*), common snipe (*Gallinago gallinago*), and common redshank (*Tringa totanus*). We analyzed and mapped their distribution in a part of the Yorkshire Dales, an upland area of the UK. Habitat selection of the three species is well-studied, particularly at finer spatial scales (e.g., Baines, 1990; Green et al., 1990; O'Brien, 2002; Taylor and Grant, 2004; Smart et al., 2006; Sharps et al., 2016). Therefore, species-environment relationships suggested by our models can be assessed against existing knowledge on the species ecology, while at larger spatial scales, there is potential for the generation of new hypotheses.

https://doi.org/10.1017/eds.2022.34 Published online by Cambridge University Press

Specifically, at the territory level, all three species are known to be associated with wet conditions (Green et al., 1990; O’Brien, 2002; Smart et al., 2006). Northern lapwing is also known to preferred gentle over steep slopes and to avoid improved pastures (Henderson et al., 2002; Taylor and Grant, 2004). Northern lapwing is listed as globally near threatened (IUCN, 2022). In the UK, Northern lapwing is listed as red, and common snipe and common redshank as amber conservation status (Woodward et al., 2020). Managing these and other declining species is a major concern to many land managers.

We surveyed Northern lapwing, common redshank and common snipe in 244 observation units (each ca. 300 × 500 m) in the Yorkshire Dales, UK (Supplementary Figure S1). With observation units of this size we aimed to approximate the reported sizes of core areas used during the breeding season (home ranges henceforth): common redshank chicks were found up to 180 ± 68 m from their nests in salt marshes in Germany (Thyen et al., 2008); Northern lapwing chicks up to 202 m (61–386 m) from nests in Swedish farmland (Johansson and Blomqvist, 1996) and incubating female common snipe foraged 17–390 m from their nests in lowland sites of the UK (Green et al., 1990). We surveyed the observation units from 61 transects (each 2 km long) to minimize traveling time between observation units, where each transect dissected four contiguous observation units. Surveys targeted elevations below 500 m and were carried out during the 2008 breeding season. To reduce the risk of recording false absences each transect was surveyed three times between April 2 and July 1. This covered much of the incubation and chick-rearing periods of the species (Robinson, 2017). Imperfect detection of species can for example be accounted for in occupancy models (MacKenzie et al., 2003). Modeling approaches that retain the flexibility of machine learning approaches while accounting for imperfect detection are an emerging field (see Mohankumar and Hefley, 2022 for guidance) and we do not apply these methods here. Each observation unit was recorded as having the focal species present when at least one individual (excluding sightings of flocks) was recorded during at least one survey (for Northern lapwing, due to its earlier laying date relative to common snipe and common redshank, exclusive of the last repeat survey). The data were slightly unbalanced for Northern lapwing and highly unbalanced for common snipe and common redshank: with 37% (90) presences in 244 observation units for Northern lapwing, 18% (44) for common snipe and 8% (19) for common redshank. For a full description of the bird surveys, see Supplementary Appendix S1.

2.6. *Environmental covariates*

We studied environmental variation at the scale of observation units and coarser scales. Bird species distributions are frequently influenced by food availability (Green et al., 1990; Pearce-Higgins and Yalden, 2004), microclimate (Wiebe and Martin, 1998), habitat structure and composition (Pearce-Higgins and Yalden, 2004; Chalfoun and Martin, 2009), disturbance (Gill et al., 1996), or perceived predation risk (Fontaine and Martin, 2006; van der Wal and Palmer, 2008). Distribution data of these variables are rarely available for large areas, so we used more widely available Geographical Information System (GIS) data with credible possible relationships to these possible drivers as proxies. For food availability, we used soil characteristics, elevation, slope, aspect, and rainfall; for microclimatic conditions, we used elevation, aspect, and rainfall; for habitat structure and composition we used soil characteristics, livestock numbers, and satellite data; for wet areas, we used satellite data; for disturbance, we used human settlements, roads, paths, and so forth, and the number of walking groups recorded during surveys; for perceived predation risk we used human settlements, field walls, and viewshed. For a description of the covariates and the possible links between our proxies and food availability, microclimate, habitat structure and composition, disturbance and perceived predation risk, see Supplementary Appendix S1.

Several variables were categorical but we did not always have good prior knowledge of the appropriate grouping for categories. For example, we expected that west-facing areas exposed to the stronger winds (Met Office, 2015) would have a more unfavorable microclimate than south-facing areas receiving more solar irradiation, but did not know if south-west-facing areas should be grouped with the former or the latter. Therefore, we initially created a fine division of categories and checked for possible grouping of neighboring categories after variable selection with random forest (see analysis below).

https://doi.org/10.1017/eds.2022.34 Published online by Cambridge University Press

Boyce et al. (2017) argued that covariates should be measured in a way that is relevant to how the study species perceives the environment. We calculated covariates as the amount of a certain habitat (area or length) within circular buffers of varying size. These values correspond, for example, to the area of wet soil within 0.5 km or the length of paths and roads within 1 km. Our rationale was that birds are highly mobile and therefore the area of, for example, wet soil within a certain distance will be more relevant than the number of patches of wet soil or their size or shape. We chose circular buffers with radii of 0.25, 1, 2.5, 5, 7, and 10 km. Some covariates were used at fewer scales or at a single scale only, such as covariates extracted from satellite imagery. The values of each covariate at each spatial scale were extracted for the centroid of the part of each observation unit visible from the transect. The total number of covariates was 288: 216 multi-scale and 72 single-scale.

2.7. Analysis

For each species, we simultaneously evaluated all single and multi-scale covariates with random forest and the cFor-AUC/AUC variable selection approach outlined above. In other words, we used AUC both in the calculation of permutation importances and as the threshold criterion in variable selection. We used the conditional version of permutation importance where permutations of a covariate are carried out such as to preserve the correlation structure with other covariates, which allows a better distinction between true covariates and covariates correlated to true variables (Strobl et al., 2008). Additionally, for the selected covariates, we identified clusters of covariates with Spearman rho > |0.7|, a commonly used threshold (Dormann et al., 2013), and reduced the covariates to the highest ranked of the cluster. We compared the OOB error with this reduced set to the OOB error before omission of the correlated covariates. As the OOB error remained similar, or even improved slightly (common snipe) we simplified the models by accepting the reduced sets to aid model interpretation.

For the covariates thus selected, we reviewed category groupings. We inspected partial dependence plots of retained covariates and additively grouped neighboring categories at the same spatial scale if the shape of the plots suggested a similar relationship with presence–absence of the species. For Northern lapwing, we additively grouped north-east, east, south-east, and south-facing aspects at the 10 km scale. Partial dependence plots were also used to describe the direction of species–habitat relationships in the final models. To generate predicted maps of the species distribution based on the covariates thus selected, we downsampled the majority class to the number of samples in the minority class to better balance prediction error between the classes.

2.8. Software used

The analyses were carried out in R 4.0.3 and R 4.1.2 (R Core Team, 2020). AUC was calculated using package ROCR (Sing et al., 2005). Processing of GIS data was carried out in ArcGIS 9.2 and 10.3 (ESRI, 2010).

3. Results

3.1. Simulation study

For medium effect sizes of 0.9 and a single true variable with low VIFs, all algorithm versions ranked the true variable highly for the data set with the largest sample size and lowest imbalance level (sample size: 300 absences; 20% presences; Figure 3a, left). As sample size decreased and imbalance level increased, all algorithm versions produced low ranks for the true variable eventually. The poorest performers were the version based on the Hellinger distance and downsampling: the ranks of the true variable became highly variable already at an imbalance level of 10% even for the largest sample size (300 absences). For the smallest sample size (100 absences), the true variable was less frequently ranked highly already at an imbalance level of 20%. The default RF, the versions with priors or weights and cforest ranked the true variable still highly in some situations in which the Hellinger and downsampling versions did not, but with

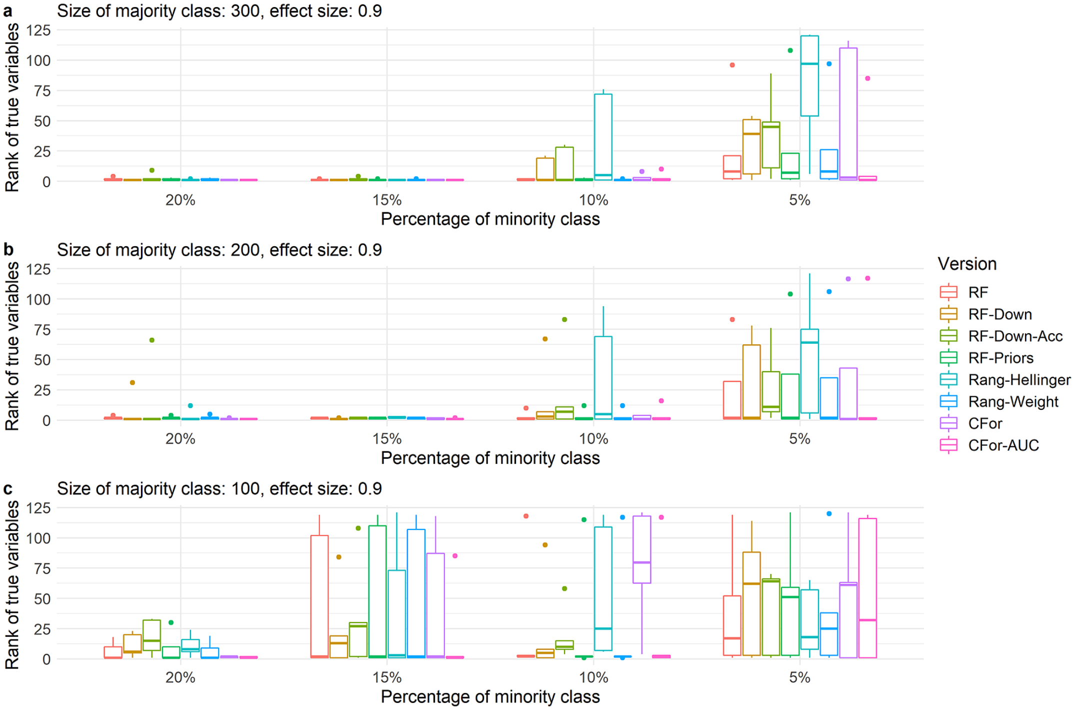


Figure 3. *Boxplots showing ranks of permutation importance for the single true predictor with low VIF and a medium effect size of 0.9. For each of the simulated 60 test sets with a total of 121 covariates, permutation importance was calculated 10 times for each of the eight algorithm versions. The true predictors were ranked by the mean permutation importance averaged over the 10 repetitions. Boxes showing interquartile range and median; whiskers show the maximum of 1.5 × interquartile range. Boxes represent the eight algorithm versions: default RF (RF), two downsampled versions (RF-Down, RF-Down-Acc), priors representing the proportion of class sizes (RF-Priors), Hellinger distance as splitting criterion (Rang-Hellinger), weighted random forest (Rang-Weight), random forest based on conditional inference trees (CFor), AUC permutation importance (CFor-AUC; from left to right). Sample sizes decrease from row a to row c. Imbalance levels increase from left to right.*

a decrease in sample size and an increase in imbalance level, their performance declined also. The only version to consistently rank the true variable highly apart from when imbalance level was highest and sample size lowest (100 absences, 5% presences) was the AUC permutation importance (CFor-AUC; Figure 3).

As the number of true variables increased or the VIFs of true variables increased and correct ranking of true variables became more difficult, the relative order of the eight algorithm versions was preserved (Figure 4 and Supplementary Figure S2). However, for all algorithm versions, the decline in performance occurred at larger sample sizes and lower imbalance levels compared to simulated data with fewer true variables and lower VIFs. For example for four true variables and low VIFs, the versions based on the Hellinger distance and downsampling performed poorly already for the largest data set (300 absences) and lowest imbalance level (20%, Figure 4a). As sample size decreased or imbalance level increased, the AUC permutation importance ranked true predictors still highly when other versions no longer did. (Figure 4 and Supplementary Figure S2).

Similarly, for high effect sizes of 1.4, the AUC permutation importation importance ranked true variables still highly for smaller sample sizes and higher imbalance levels when the other seven algorithm versions no longer did (Figure 5 and Supplementary Figure S3). The versions based on Hellinger distance and downsampling produced lower ranks for true variables sooner as sample sizes decreased and

https://doi.org/10.1017/eds.2022.34 Published online by Cambridge University Press

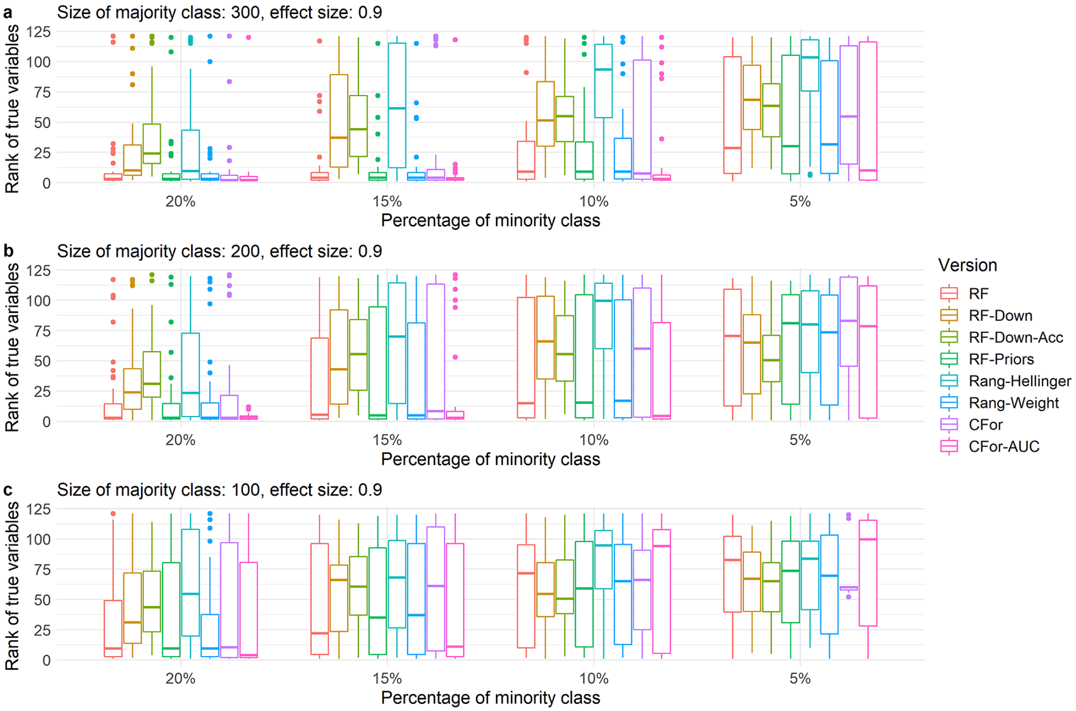


Figure 4. *Boxplots showing ranks of permutation importance for four true variables with low VIFs and a medium effect size of 0.9. For each of the 60 simulated test sets with a total of 121 covariates, permutation importance was calculated 10 times for each of the eight algorithm versions. The true predictors were ranked by the mean permutation importance averaged over the 10 repetitions. Boxes showing interquartile range and median; whiskers show the maximum of 1.5 × interquartile range. Boxes represent the eight algorithm versions: default RF (RF), two downsampled versions (RF-Down, RF-Down-Acc), priors representing the proportion of class sizes (RF-Priors), Hellinger distance as splitting criterion (Rang-Hellinger), weighted random forest (Rang-Weight), random forest based on conditional inference trees (CFor), AUC permutation importance (CFor-AUC; from left to right). Sample sizes decrease from row a to row c. Imbalance levels increase from left to right.*

imbalance levels increased compared to the other five algorithm versions. For all algorithm versions, the performance decreased as the number of true variables increased or the VIFs of true variables increased. Compared to true variables with medium effects sizes, true variables with high effect sizes were often still ranked highly at sample sizes and imbalance levels at which true variables with medium effect sizes no longer were ranked highly (Supplementary Figures S2 and S3). For low effect sizes of 0.4, all algorithm versions produced low ranks for true variables in most situations and only for the highest sample sizes and lowest imbalance levels and with few true variables did the four algorithm versions AUC permutation importance, default RF and the versions based on priors and weights rank true variables relatively highly (Supplementary Figure S4). Reversing the class weights in Rang-Weight had little effect on the ranking of true variables, but ranks of true variables were substantially lower when reversing the class priors in RF-Priors (Supplementary Figure S5).

At the sample sizes, imbalance levels, number of true variables, VIFs of true variables, and effect sizes at which the AUC permutation importance produced higher ranks than the default RF permutation importance, sensitivity in variable selection increased when the AUC permutation importance was used to rank all covariates. Using AUC instead of OOB as the criterion in variable selection increased sensitivity in variable selection even further at higher sample sizes and imbalance levels. Specificity showed a slight

https://doi.org/10.1017/eds.2022.34 Published online by Cambridge University Press

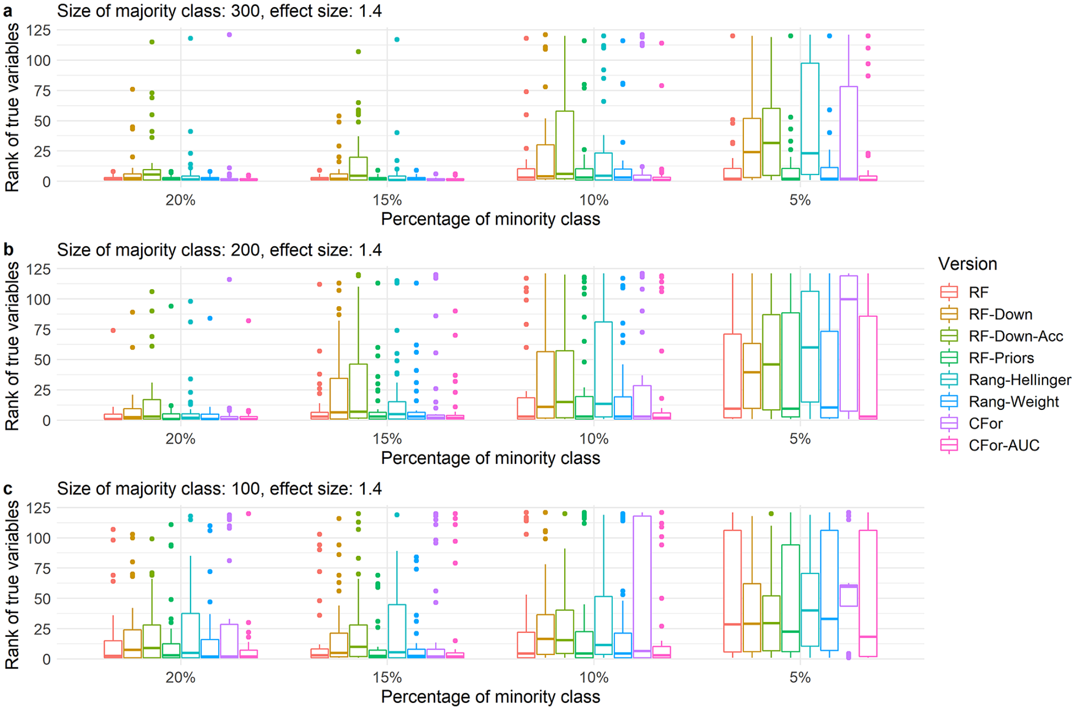


Figure 5. *Boxplots showing ranks of permutation importance for true variables with a high effect size of 1.4. For each of the 360 simulated test sets with a total of 121 covariates, permutation importance was calculated 10 times for each of the eight algorithm versions. The true predictors were ranked by the mean permutation importance averaged over the 10 repetitions. The boxplots summarize the ranking across all true variables with high effect sizes of 1.4 across simulated data with one, two, and four true variables and with low and with high VIFs of the true variables. Boxes showing interquartile range and median; whiskers show the maximum of 1.5 × interquartile range. Boxes represent the eight algorithm versions: default RF (RF), two downsampled versions (RF-Down, RF-Down-Acc), priors representing the proportion of class sizes (RF-Priors), Hellinger distance as splitting criterion (Rang-Hellinger), weighted random forest (Rang-Weight), random forest based on conditional inference trees (CFor), AUC permutation importance (CFor-AUC; from left to right). Sample sizes decrease from row a to row c. Imbalance levels increase from left to right.*

reverse trend among the two versions based on AUC permutation importance, but was relatively high overall (Figure 6).

3.2. Case studies

The final models for the distributions of the three breeding waders (Figure 7) contained covariates at scales between 0.25 and 10 km (Table 1). At finer spatial scales the models identified species–habitat relationships we expected to find based on previous knowledge about the species. At fine scales similar to the home ranges of the species (250 m), all species had an association to wetter conditions: for Northern lapwing to moist soils, for common snipe to soils with impeded drainage and for common redshank we found an association to a satellite data vegetation index (IVR) which has been linked to marshes (Table 1). Northern lapwings avoided elevations of 200–300 m, which in the study area are dominated by bottoms of major valleys and intensive spring grazing, hence our model indicated an avoidance of improved pastures. As expected for Northern lapwings, our model also showed that Northern lapwings avoided steep slopes and preferred gentle slopes (Table 1).

https://doi.org/10.1017/eds.2022.34 Published online by Cambridge University Press

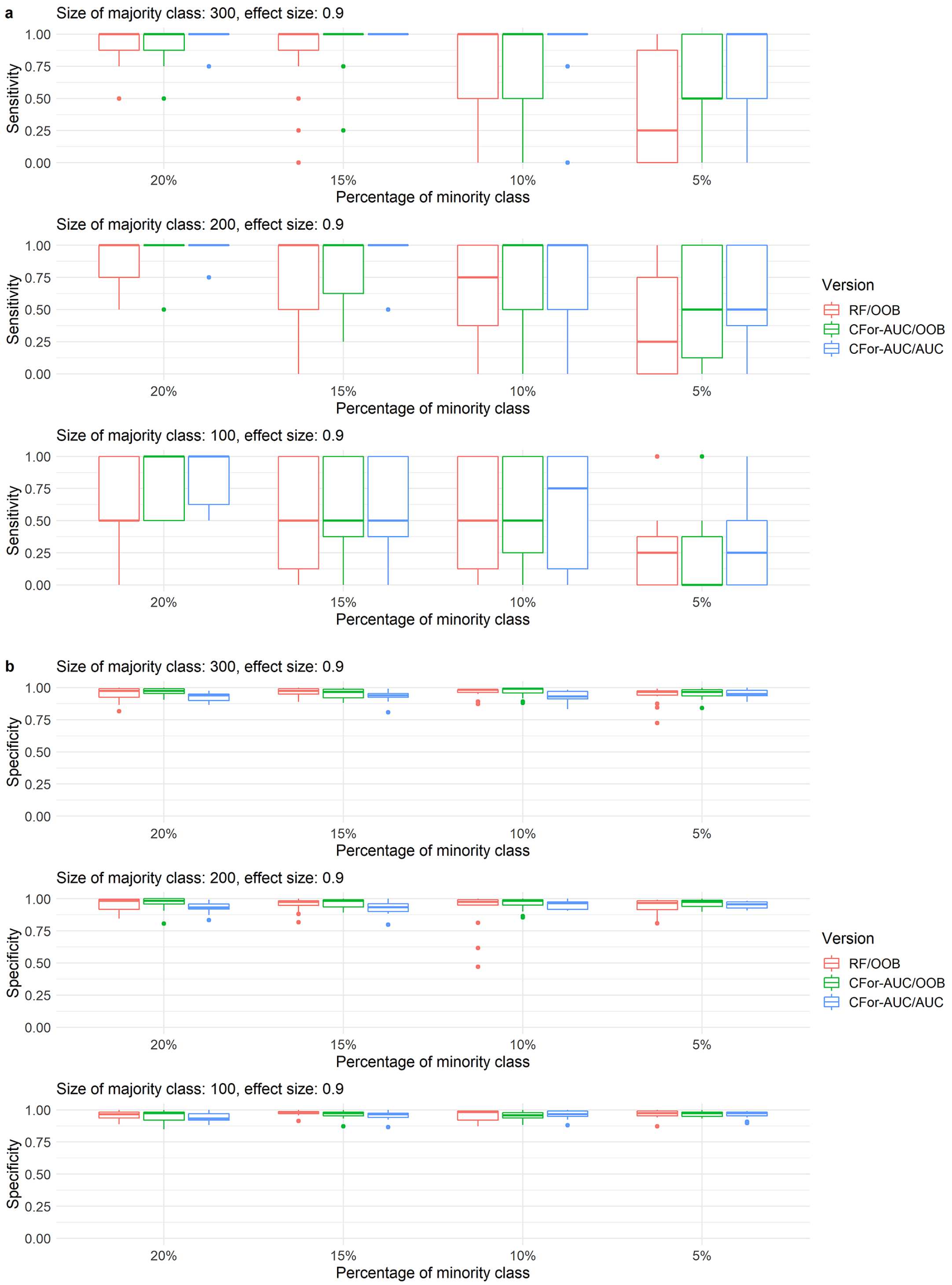


Figure 6. *Boxplots of (a) sensitivity and (b) specificity of variable selection for three alternative variable selection approaches with random forest. Variable selections were implemented for simulated data with 121 covariates including one, two, or four covariates with a medium effect size of 0.9 and with low (Continued)*

https://doi.org/10.1017/eds.2022.34 Published online by Cambridge University Press

variable inflation factors. Sensitivity represents the rate at which true covariates are identified. Specificity represents the rate at which noise covariates are rejected. Boxes represent from left to right (1) RF/OOB: the default variable selection approach with no adjustments for unbalanced data, (2) CFor-AUC/OOB: covariates are ranked by the AUC permutation importance instead of the default permutation importance, and (3) CFor-AUC/AUC: covariates are ranked by the AUC permutation importance and the threshold criterion in the forward selection is AUC instead of OOB. Boxes showing interquartile range and median, whiskers show the maximum of 1.5 × interquartile range, dots: outliers.

The models suggested several species–habitat relationships for the less well-studied coarser scales. Aspect was frequently selected at scales of 5–10 km (Table 1). Partial dependence plots suggested that relationships of breeding common snipe with west-facing aspect at a scale of 7 km were negative. Relationships of all three species with either north, north-east, east, south-east, or south-facing aspects were positive at scales of 5–10 km. At scales of 5–10 km, all species had a positive relationship to the area of wet soils. Northern lapwing and common redshank had a positive relationship with gently sloping (2–5°) land at the 7–10 km scale.

4. Discussion

With unbalanced data and random forest, the majority class may be predicted with a higher accuracy than the minority class and permutation importances may be biased (Chen et al., 2004; Lin and Chen, 2012; Janitza et al., 2013). For variable rankings based on simulated data we found that the AUC permutation importance was least affected by data imbalance out of eight alternative algorithm versions. All eight algorithm versions ranked true variables lower and more variably as sample size decreased, the imbalance level increased, the number of true variables increased, their effect sizes decreased or their VIFs increased. For the same number of true variables, effect sizes, and VIFs, the AUC permutation importance ranked true variables still highly at sample sizes and imbalance levels at which the other seven algorithms did not. Conversely, the version with the Hellinger distance and the two downsampling versions already ranked true variables lower and more variably at sample sizes and imbalance levels at which the other five algorithms still ranked true variables highly. Higher ranking of true variables with the AUC permutation importance lead to a higher proportion of true variables identified during variable selection compared to the default permutation importance. Sensitivity increased further when the AUC instead of the OOB error was used as the criterion in the forward variable selection. The sample sizes and imbalance levels in our simulated data are typical for some national monitoring programmes of species (Schmid et al., 2004; Snäll et al., 2011). Within the field of species distribution modeling, our results can therefore be relevant for data from national or regional studies (see our case studies) and outside of species distribution modeling for data with similar sample sizes and imbalance levels as in our simulated data.

Even the highest performing algorithm version, the AUC permutation importance failed to rank true variables highly at the lowest sample sizes and highest imbalance levels and consequently failed to achieve a high sensitivity in the identification of true variables during variable selection. When the absolute number of cases of the minority class is very low, the corresponding lack of information cannot be overcome with improvements to the calculation of variable importances (Janitza et al., 2013). Nonetheless, our results for the AUC permutation importance suggest that when VIFs of true variables were low and the classification problems relatively simple (1–2 true variables, medium effect sizes), as few as 10–20 cases for the minority class were sufficient to achieve high ranks of true variables despite data imbalance and despite a large number of covariates relative to sample size. With high effect sizes and 1–2 true variables, even as few as five cases for the minority class were at times sufficient to rank true variables highly. When VIFs of true variables are high, analysts may be able to increase their success of correctly ranking covariates by manually eliminating covariates with high VIFs before a ranking of covariates with random forest. However, in the subsequent interpretation of results, it may be important to

https://doi.org/10.1017/eds.2022.34 Published online by Cambridge University Press

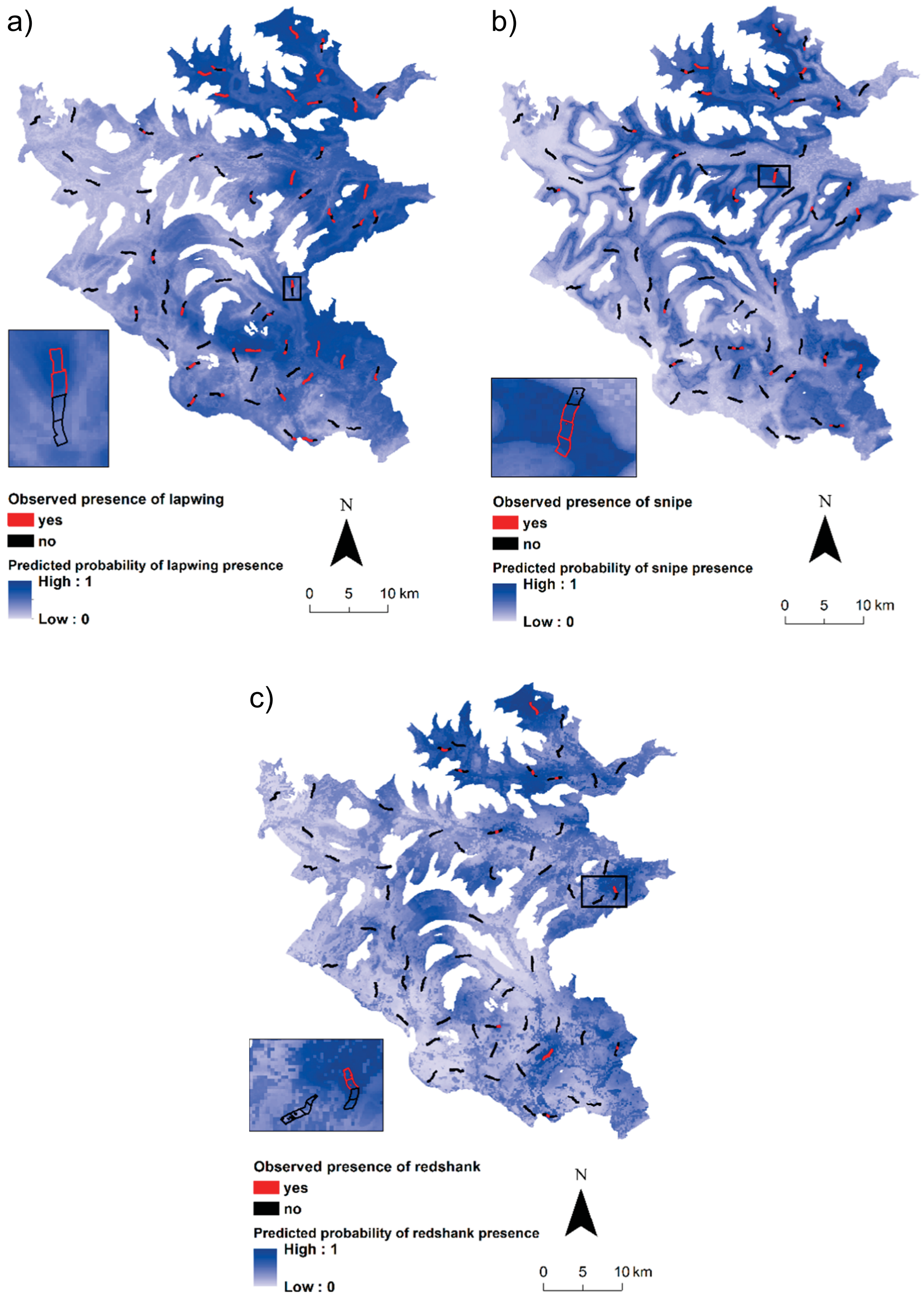

a)
Observed presence of lapwing
🟥 yes
■ no
Predicted probability of lapwing presence
High : 1
Low : 0
N
0 5 10 km
b)
Observed presence of snipe
🟥 yes
■ no
Predicted probability of snipe presence
High : 1
Low : 0
N
0 5 10 km
c)
Observed presence of redshank
🟥 yes
■ no
Predicted probability of redshank presence
High : 1
Low : 0
N
0 5 10 km

Figure 7. *Predicted probability of (a) Northern lapwing, (b) common snipe, and (c) common redshank presence in the study area. Only areas below 500 m elevation shown, which were those surveyed. Also shown is whether these species were recorded on the transects. Maps derived: using NATMAP soilscapes Cranfield University (NSRI) and for the Controller of HMSO 2009, Land-Form PANORAMA, OS MasterMap, Strategi data downloaded from the EDINA Digimap OS service. Crown Copyright/ database right 1993, 2007, and 2009. An Ordnance Survey / EDINA supplied service, Landsat Surface Reflectance products courtesy of the U.S. Geological Survey Earth Resources Observation and Science Center.*

https://doi.org/10.1017/eds.2022.34 Published online by Cambridge University Press

Table 1. Selected multi-scale covariates in models for Northern lapwing, common snipe, and common redshank in the Yorkshire Dales, UK.

Northern lapwing			
Covariate	Scale (km)	Rank	Relationship
Slope: 0°–2°	0.25 and 1	8, 14	+, +
Slope: 2°–5°	10	2	+
Slope: 15°–25°	0.25	13	–
Elevation: 200–300 m	0.25	3	–
Aspect: flat	10	9	U
Aspect: north-east, east, south-east, south-facing	10	1	+
Aspect: north-facing	5	7	+
Soil: moist	0.25	12	+
Soil: impeded drainage	5	6	(–)
Soil: wet	5	4	+
Soil: moderate and high fertility	5	16	+
Soil: low fertility	10	5	–
Soil: loam	0.25	10	(–)
Landsat: standard deviation thermal band	0.25	11	–
Landsat: NDWI	0.25	15	–
Common snipe			
Elevation: 300–400 m	0.25	1	+
Elevation: 500–600 m	5	4	+
Aspect: north-facing	5	2	+
Aspect: south-facing	10	8	+
Aspect: west-facing	7	6	–
Soil: wet	10	3	+
Soil: impeded drainage	0.25	5	(+)
Soil: loamy	0.25	7	–
Landsat: standard deviation thermal band	0.25	9	(+)
Common redshank			
Slope: 0°–2°	0.25	6	+
Slope: 2°–5°	7	9	+
Elevation: 200–300 m	5	11	(–)
Aspect: level	0.25 and 1	2, 4	+, +
Aspect: north	7	3	+

Table 1. *Continued*

Common redshank			
Aspect: south-facing	10	10	+
Aspect: north-west-facing	0.25 and 2.5	7, 8	+, +
Soil: wet	5	1	+
Soil: high fertility	1	12	+
Landsat: IVR	0.25	5	–

Note. Rank refers to the position of the covariate when ordered by mean permutation importance. Relationship: Directions of the species–habitat relationships were inferred from partial dependence plots of a random forest model. +, Probability of species presence increased with higher values of the covariate. −, Probability of species presence decreased with higher values of the covariate. (), considerable variation in the overall trend. U, Probability of a species presence increased with lower and higher values of the covariate and decreased with intermediate values.

consider that highly ranked covariates may then be only proxies for true drivers that were manually eliminated. When the aim of a study is to find many of the important variables associated with a response variable (Nicodemus et al., 2010; Genuer et al., 2015), a preselection of covariates with low VIFs may not be desirable.

Downsampling and use of the Hellinger distance in node splits resulted not only in lower average ranks of true variables at sample sizes and imbalance levels at which the five other algorithm versions still produced high ranks, but also in highly variable ranking of true predictors, which impedes unbiased variable selection and inference. In our simulations, we averaged permutation importances over 10 repetitions of a random forest model using a relatively high number of trees (2,000). In studies in which permutation importances from a single random forest model are used, the variability in importance measures will likely be higher. In contrast to our results based on variable ranking, (Valavi et al., 2021a) found that downsampling and the Hellinger distance combined with shallow trees produced the highest discrimination between classes. Their sample sizes were higher compared to our study, which may limit the risk of information loss in downsampling. They suggested that further research is needed to investigate the contributions of the Hellinger distance as splitting criterion versus tree depth to the improvement in discriminative ability. We found that reversing the weights in Rang-Weight had little effect on the ranking of true variables, suggesting not only that no fine-tuning of weights is necessary for variable ranking, but also that the weighing may have little effect on variable ranking. Conversely, reversing the priors resulted in substantially lower rankings of true variables.

Nonlinear covariate effects and interactions among covariates are common in environmental sciences, including species distribution modeling (Franklin, 2009). Random forest learns nonlinearities and interactions from the data without the need to explicitly specify these (Grömping, 2009). Our simulations were based on covariate effects that were linear (at the scale of the linear predictor) and additive. Further research is necessary to investigate if the AUC permutation importance performs better compared to other approaches also in more complex situations that involve covariate interactions and nonlinear effects.

Our simulation and case studies considered random forest for classification. Data imbalance may also be an issue in random forest for regression. Few datapoints in a region of interest in environmental space may result in poor predictive performance for the region of interest (Branco et al., 2017). Strategies to improve predictive performance in the area of interest with random forest for regression are less well-researched compared to the problem of unbalanced data in random forest for classification (Branco et al., 2017).

4.1. Case studies

The results from our case studies suggest that our models correctly identified a number of known species–habitat relationships despite the problems of unbalanced data. Habitat selection at finer spatial scales such as home ranges is well-studied for all three species (e.g., Baines, 1990; Green et al., 1990; O'Brien, 2002; Taylor and Grant, 2004; Smart et al., 2006; Sharps et al., 2016). At this scale, our models identified species–habitat relationships consistent with previous studies. We found preferences for wet conditions in all species, which has previously been found (Green et al., 1990; O'Brien, 2002; Smart et al., 2006). We also found that Northern lapwing preferred gentle over steep slopes and that they avoided improved pastures, consistent with previous studies (Henderson et al., 2002; Taylor and Grant, 2004).

At coarser spatial scales, less knowledge about species–habitat relationships is available. Based on the case studies, we suggest the following hypotheses about processes at larger spatial scales: First, all species showed a positive relationship to south-facing aspects and common snipe had a negative relationship to a west-facing aspect. The birds may be selecting for warmer microclimates with higher solar irradiation (south-facing aspects) and avoiding the strong westerly wind (Met Office, 2015). Whinchat (*Saxicola rubetra*) also favored south and east-facing aspects (for home ranges) (Calladine and Bray, 2012). This was hypothesized to be linked to higher food availability in warmer aspects. In the hilly landscape of the Yorkshire Dales, small patches can be sheltered from wind or sun by local topography, and the relationship may only be revealed at coarser spatial scales. Second, at the 10 km scale, we found that Northern lapwing was associated with gentle slopes (2°–5°). In the prebreeding period, female-dominated Northern lapwing flocks forage in fields with high earthworm densities (often improved fields) and later nest in surrounding fields (Baines, 1990) preferring gentle slopes (Taylor and Grant, 2004) and avoiding improved pastures (Henderson et al., 2002). The positive relationship with gentle slopes at the 10 km scale may thus reflect a preference for the wider valleys of the Yorkshire Dales, which offer more areas with a gentle slope outside of intensively grazed valley bottoms in combination with many potential prebreeding foraging fields in the vicinity. Third, there was a positive relationship with wet soil at coarser spatial scales for all species. A positive relationship with wet conditions at the scale of home ranges is well known for all three species (Green et al., 1990; O'Brien, 2002; Smart et al., 2006). In landscapes with more abundant wet areas, habitat unoccupied in other landscapes may be used, because pairs are less dependent on the quality of a single small area if the surroundings offer opportunities to supplement resources. Alternatively, in landscapes with abundant wet areas and consequently a higher density of breeding pairs, conspecific attraction may lead to increased settlement. There is evidence from some bird species, that areas with higher breeding densities attract more immigrants (Doligez et al., 2004). The modeled species–habitat relationships only reveal patterns, not causation and the new hypotheses will need to be tested in additional studies (Dormann et al., 2012).

The importance of considering multiple spatial scales of environmental covariates to derive accurate species distribution maps is well known and reflects that multiple-scale processes are shaping species distributions (Wiens, 1976; Johnson, 1980; Addicott et al., 1987; Wiens et al., 1987; Wiens, 1989; Levin, 1992; Thogmartin and Knutson, 2007; Mateo-Tomas and Olea, 2009; Jackson and Fahrig, 2012; Bellamy et al., 2013; Bradter et al., 2013; Bellamy and Altringham, 2015; Galitsky and Lawler, 2015; McGarigal et al., 2016). Our multi-scale models revealed that, at least in the case of the three wader species, conservation policy and delivery may need to consider approaches that work at the scale of farm clusters or even whole valleys bringing together groups of landowners to support wader populations over wider areas.

Abbreviations

AUC	area under the curve
CFor	random forest based on conditional inference trees
CFor-AUC	AUC permutation importance in random forest based on conditional inference trees
CFor-AUC/AUC	variable selection in which variables are ranked based on the AUC permutation importance and AUC is used as the threshold criterion in forward variable selection
CFor-AUC/OOB	variable selection in which variables are ranked based on the AUC permutation importance and OOB error is used as the threshold criterion in forward variable selection
mtry	the number of variables tried at each node split
ntree	the number of trees in a forest

OOB	out-of-bag
Rang-Hellinger	random forest with the Hellinger distance as splitting criterion
Rang-Weight	random forest with weights to account for differences in class size
RF	random forest with no adjustment for unbalanced data
RF-Down	random forest in which the majority class is downsampled to the size of the minority class
RF-Down-Acc	random forest in which the majority class is downsampled to 64% of the size of the minority class
RF/OOB	variable selection with no adjustment for unbalanced data. Variables are ranked based on the permutation importance and OOB error is used as the threshold criterion in forward variable selection
RF-Priors	random forest with priors to account for differences in class size

Author Contributions. U.B. conceived the idea, collected the survey data, carried out the analysis, and wrote the first draft. T.G.B., J.D.A., W.E.K., and T.J.T. advised on study design and analysis. J.O. advised on remotely sensed data and their acquisition. All authors contributed to the final version.

Competing Interests. The authors declare no competing interests exist.

Data Availability Statement. The simulated data and the bird survey data are available on Zenodo: http://doi.org/10.5281/zenodo.7337856. The GIS data described in Supplementary Appendix S1 are available from Digimap: https://digimap.edina.ac.uk, the Met office: www.metoffice.gov.uk, the U.S. Geological Survey Earth Resources Observation and Science Center: http://landsat.usgs.gov and the National Soil Resources Institute: www.landis.org.uk. R code is provided in the Github repository: https://github.com/UteBradter/Unbalanced_RandomForest.

Funding Statement. The study was funded by the Yorkshire Dales National Park Authority, the Biotechnology and Biological Sciences Research Council (Reference number: BB/J005851/1), and the Nigel Bertram Charitable Trust.

Supplementary Materials. To view supplementary material for this article, please visit http://doi.org/10.1017/eds.2022.34.

References

Adam E, **Mutanga O**, **Abdel-Rahman EM and Ismail R** (2014) Estimating standing biomass in papyrus (*Cyperus papyrus* L.) swamp: Exploratory of *in situ* hyperspectral indices and random forest regression. *International Journal of Remote Sensing 35*, 693–714.

Addicott JF, **Aho JM**, **Antolin MF**, **Padilla DK**, **Richardson JS and Soluk DA** (1987) Ecological neighborhoods: Scaling environmental patterns. *Oikos 49*, 340–346.

Aler R, **Valls JM and Boström H** (2020) Study of Hellinger distance as a splitting metric for random forests in balanced and imbalanced classification datasets. *Expert Systems with Applications 149*, 113264.

Baines D (1990) The roles of predation, food and agricultural practice in determining the breeding success of the lapwing (*Vanellus vanellus*) on upland grasslands. *Journal of Animal Ecology 59*(3), 915–929.

Barrett B, **Raab C**, **Cawkwell F and Green S** (2016) Upland vegetation mapping using random forests with optical and radar satellite data. *Remote Sensing in Ecology and Conservation 2*, 212–231.

Bellamy C and Altringham J (2015) Predicting species distributions using record Centre data: Multi-scale modelling of habitat suitability for bat roosts. *PLoS One 10*, e0128440.

Bellamy C, **Scott C and Altringham J** (2013) Multiscale, presence-only habitat suitability models: Fine-resolution maps for eight bat species. *Journal of Applied Ecology 50*, 892–901.

Boyce MS, **Mallory CD**, **Morehouse AT**, **Prokopenko CM**, **Scrafford MA and Warbington CH** (2017) Defining landscapes and scales to model landscape-organism interactions. *Current Landscape Ecology Reports 2*, 89–95.

Bradter U, **Kunin WE**, **Altringham JD**, **Thom TJ and Benton TG** (2013) Identifying appropriate spatial scales of predictors in species distribution models with the random forest algorithm. *Methods in Ecology and Evolution 4*, 167–174.

Bradter U, **O'Connell J**, **Kunin WE**, **Boffey CWH**, **Ellis RJ and Benton TG** (2020) Classifying grass-dominated habitats from remotely sensed data: The influence of spectral resolution, acquisition time and the vegetation classification system on accuracy and thematic resolution. *Science of the Total Environment 711*, 134584.

Bradter U, **Thom TJ**, **Altringham JD**, **Kunin WE and Benton TG** (2011) Prediction of National Vegetation Classification communities in the British uplands using environmental data at multiple spatial scales, aerial images and the classifier random forest. *Journal of Applied Ecology 48*, 1057–1065.

Branco P, **Torgo L and Ribeiro RP** (2017) SMOGN: A pre-processing approach for imbalanced regression. *Proceedings of Machine Learning Research 74*, 36–50.

Breiman L (2001) Random forests. *Machine Learning 45*(1), 5–32.

Calladine J and Bray J (2012) The importance of altitude and aspect for breeding whinchat *Saxicola rubetra* in the uplands: Limitations of the uplands as a refuge for a declining, formerly widespread species? *Bird Study 59*, 43–51.

Chalfoun AD and Martin TE (2009) Habitat structure mediates predation risk for sedentary prey: Experimental tests of alternative hypotheses. *Journal of Animal Ecology 78*(3), 497–503.

Chen C, **Liaw A and Breiman L** (2004) *Using Random Forest to Learn Imbalanced Data.* Technical Report 666, Statistics Department, University of California at Berkeley.

Cieslak DA, **Hoens TR**, **Chawla NV and Kegelmeyer WP** (2012) Hellinger distance decision trees are robust and skew-insensitive. *Data Mining and Knowledge Discovery 24*, 136–158.

Cunningham MA and Johnson DH (2006) Proximate and landscape factors influence grassland bird distributions. *Ecological Applications 16*(3), 1062–1075.

Cutler DR, **Edwards TC**, **Beard KH**, **Cutler A**, **Hess KT**, **Gibson J and Lawler JJ** (2007) Random forests for classification in ecology. *Ecology 88*, 2783–2792.

Dahinden C (2011) An improved random forest approach with application to the performance prediction challenge datasets. In Guyon I, Cawley G, Dror G and Saffari A (eds), *Hands-on Pattern Recognition, Challenges in Machine Learning*, Vol. 1. Brookline, MA: Microtone, pp. 223–230.

Díaz-Uriarte R and de Andrés SA (2006) Gene selection and classification of microarray data using random forest. *BMC Bioinformatics 7*, 3.

Doligez B, **Pärt T**, **Danchin E**, **Clobert J and Gustafsson L** (2004) Availability and use of public information and conspecific density for settlement decisions in the collared flycatcher. *Journal of Animal Ecology 73*, 75–87.

Dormann CF, **Elith J**, **Bacher S**, **Buchmann C**, **Carl G**, **Carré G**, **García Marquéz JR**, **Gruber B**, **Lafourcade B**, **Leitão PJ**, **Münkemüller T**, **McClean C**, **Osborne PE**, **Reineking B**, **Schröder B**, **Skidmore AK**, **Zurell D and Lautenbach S** (2013) Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography 36*, 27–46.

Dormann CF, **Schymanski SJ**, **Cabral J**, **Chuine I**, **Graham C**, **Hartig F**, **Kearney M**, **Morin X**, **Römermann C**, **Schröder B and Singer A** (2012) Correlation and process in species distribution models: Bridging a dichotomy. *Journal of Biogeography 39*, 2119–2131.

ESRI (2010) ArcGIS 10.7.1. ESRI, Redlands, CA.

Fawcett T (2006) An introduction to ROC analysis. *Pattern Recognition Letters 27*, 861–874.

Fontaine JJ and Martin TE (2006) Parent birds assess nest predation risk and adjust their reproductive strategies. *Ecology Letters 9* (4), 428–434.

Franklin J (2009) *Mapping Species Distributions - Spatial Inference and Prediction*. Cambridge: Cambridge University Press.

Freeman EA, **Moisen GG and Frescino TS** (2012) Evaluating effectiveness of down-sampling for stratified designs and unbalanced prevalence in random Forest models of tree species distributions in Nevada. *Ecological Modelling 233*, 1–10.

Galitsky C and Lawler JJ (2015) Relative influence of local and landscape factors on bird communities vary by species and functional group. *Landscape Ecology 30*, 287–299.

Garzón MB, **Blazek R**, **Neteler M**, **de Dios RS**, **Ollero HS and Furlanello C** (2006) Predicting habitat suitability with machine learning models: The potential area of *Pinus sylvestris* L. in the Iberian Peninsula. *Ecological Modelling 197*(3–4), 383–393.

Genuer R, **Poggi J-M and Tuleau-Malot C** (2010) Variable selection using random forests. *Pattern Recognition Letters 31*, 2225–2236.

Genuer R, **Poggi J-M and Tuleau-Malot C** (2015) VSURF: An R package for variable selectoin using random forests. *The R Journal 7*, 19–33.

Gill JA, **Sutherland WJ and Watkinson AR** (1996) A method to quantify the effects of human disturbance on animal populations. *Journal of Applied Ecology 33*, 786–792.

Green RE, **Hirons GJM and Cresswell BH** (1990) Foraging habitats of female common snipe *Gallinago gallinago* during the incubation period. *Journal of Applied Ecology 27*(1), 325–335.

Grömping U (2009) Variable importance assessment in regression: Linear regression versus random Forest. *The American Statistician 63*(4), 308–319.

Hamer KC and Hill JK (2000) Scale-dependent effects of habitat disturbance on species richness in tropical forests. *Conservation Biology 14*(5), 1435–1440.

Henderson IG, **Wilson AM**, **Steele D and Vickery JA** (2002) Population estimates, trends and habitat associations of breeding lapwing *Vanellus vanellus*, curlew *Numenius arquata* and snipe *Gallinago gallinago* in Northern Ireland in 1999. *Bird Study 49*, 17–25.

IUCN (2022) The IUCN Red List of Threatened Species. Version 2021-3. Available at https://www.iucnredlist.org. (accessed 14 March 2022).

Jackson HB and Fahrig L (2012) What size is a biologically relevant landscape? *Landscape Ecology 27*, 929–941.

Janitza S, **Strobl C and Boulesteix AL** (2013) An AUC-based permutation variable importance measure for random forests. *BMC Bioinformatics 14*, 119.

Johansson OC and Blomqvist D (1996) Habitat selection and diet of lapwing *Vanellus vanellus* chicks on coastal farmland in S.W. Sweden. *Journal of Applied Ecology 33*(5), 1030–1040.

https://doi.org/10.1017/eds.2022.34 Published online by Cambridge University Press

Johnson DH (1980) The comparison of usage and availability measurements for evaluating resource preference. *Ecology 6*, 65–71.

Karimi SS, **Saintilan N**, **Wen L and Valavi R** (2019) Application of machine learning to model wetland inundation patterns across a large semiarid floodplain. *Water Resources Research 55*(11), 8765–8778.

Levin SA (1992) The problem of pattern and scale in ecology. *Ecology 73*(6), 1943–1967.

Liaw A and Wiener M (2002) Classification and regression by random forest. *R News 2/3*, 18–22.

Lin W-J and Chen JJ (2012) Class-imbalanced classifiers for high-dimensional data. *Briefings in Bioinformatics 14*, 13–26.

Liu Y, **Chawla NV**, **Harper MP**, **Shriberg E and Stolcke A** (2006) A study in machine learning from imbalanced data for sentence boundary detection in speech. *Computer Speech and Language 20*, 468–494.

MacKenzie DI, **Nichols JD**, **Hines JE**, **Knutson MG and Franklin AB** (2003) Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly. *Ecology 84*, 2200–2207.

Mateo-Tomas P and Olea PP (2009) Combining scales in habitat models to improve conservation planning in an endangered vulture. *Acta Oecologica 35*(4), 489–498.

McGarigal K, **Wan HY**, **Zeller KA**, **Timm BC and Cushman SA** (2016) Multi-scale habitat selection modelling: A review and outlook. *Landscape Ecology 31*, 1161–1175.

Met Office (2015) North East England: Climate. Available at http://www.metoffice.gov.uk/climate/uk/regional-climates/ne (accessed 4 August 2015).

Min Z, **Jing X**, **Xiaoqing J**, **Molei Y**, **Guolong C**, **Jing Y and Gangmin N** (2018) Class weights random forest algorithm for processing class imbalanced medical data. *IEEE Access 6*, 4641–4652.

Mohankumar NM and Hefley TJ (2022) Using machine learning to model nontraditional spatial dependence in occupancy data. *Ecology 103*(2), e03563.

Nicodemus KK, **Malley JD**, **Strobl C and Ziegler A** (2010) The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC Bioinformatics 11*, 110.

O'Brien M (2002) The relationship between field occupancy rates by breeding lapwing and habitat management on upland farmland in northern Britain. *Aspects of Applied Biology 67*, 85–92.

O'Connell J, **Bradter U and Benton TG** (2015) Wide-area mapping of small-scale features in agricultural landscapes using airborne remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing 109*, 165–177.

Pearce-Higgins JW and Yalden DW (2004) Habitat selection, diet, arthropod availability and growth of a moorland wader: The ecology of European golden plover *Pluvialis apricaria* chicks. *Ibis 146*(2), 335–346.

Prasad AM, **Iverson LR and Liaw A** (2006) Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. *Ecosystems 9*(2), 181–199.

Probst P (2020) varImp: RF Variable Importance for Arbitrary Measures. R Package Version 0.4.

R Core Team (2020) *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Robinson RA (2017) *BirdFacts: Profiles of Birds Occurring in Britain & Ireland.* BTO Research Report 407. BTO, Thetford. Available at http://bto.org/birdfacts (accessed 23 January 2017).

Robinson OJ, **Ruiz-Gutierrez V and Fink D** (2018) Correcting for bias in distribution modelling for rare species using citizen science data. *Diversity and Distributions 24*, 460–472.

Robinson OJ, **Ruiz-Gutierrez V**, **Fink D**, **Meese RJ**, **Holyoak M and Cooch EG** (2018) Using citizen science data in integrated population models to inform conservation. *Biological Conservation 227*, 361–368.

Ruiz-Gazen A and Villa N (2007) Storms prediction: Logistic regression vs random forest for unbalanced data. *Case Studies in Business, Industry and Government Statistics (CSBIGS) 1*, 91–101.

Ryo M, **Angelov B**, **Mammola S**, **Kass JM**, **Benito BM and Hartig F** (2021) Explainable artificial intelligence enhances the ecological interpretability of black-box species distribution models. *Ecography 44*(2), 199–205.

Schmid H, **Zbinden N and Keller V** (2004) Überwachung der Bestandsentwicklung häufiger Brutvögel in der Schweiz, Schweizerische Vogelwarte, Sempach.

Sesnie SE, **Gessler PE**, **Finegan B and Thessler S** (2008) Integrating Landsat TM and SRTM-DEM derived variables with decision trees for habitat classification and change detection in complex neotropical environments. *Remote Sensing of Environment 112*(5), 2145–2159.

Sharps E, **Garbutt A**, **Hiddink JG**, **Smart MA and Skov MW** (2016) Light grazing of saltmarshes increases the availability of nest sites for common redshank *Tringa totanus*, but reduces their quality. *Agriculture, Ecosystems and Environment 221*, 71–78.

Sing T, **Sander O**, **Beerenwinkel N and Lengauer T** (2005) ROCR: Visualizing classifier performance in R. *Bioinformatics 21*, 3940–3941.

Sluiter R and Pebesma EJ (2010) Comparing techniques for vegetation classification using multy- and hyperspectral images and ancillary environmental data. *International Journal of Remote Sensing 31*, 6143–6161.

Smart J, **Gill JA**, **Sutherland WJ and Watkinson AR** (2006) Grassland-breeding waders: Identifying key habitat requirements for management. *Journal of Applied Ecology 43*(3), 454–463.

Snäll T, **Kindvall O**, **Nilsson J and Pärt T** (2011) Evaluating citizen-based presence data for bird monitoring. *Biological Conservation 144*, 804–810.

Strobl C, **Boulesteix A-L**, **Kneib T**, **Augustin T and Zeileis A** (2008) Conditional variable importance for random forests. *BMC Bioinformatics 9*, 307.

https://doi.org/10.1017/eds.2022.34 Published online by Cambridge University Press

Strobl C, **Boulesteix A-L**, **Zeileis A and Hothorn T** (2007) Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics 8*, 25.

Strobl C, **Malley J and Tutz G** (2009) An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods 14*, 323–348.

Taylor IR and Grant MC (2004) Long-term trends in the abundance of breeding lapwing *Vanellus vanellus* in relation to land-use change on upland farmland in southern Scotland. *Bird Study 51*, 133–142.

Thogmartin WE and Knutson MG (2007) Scaling local species–habitat relations to the larger landscape with a hierarchical spatial count model. *Landscape Ecology 22*, 61–75.

Thyen S, **Exo KM**, **Cervencl A**, **Esser W and Oberdiek N** (2008) Salzwiesen im niedersächsischen Wattenmeer als Brutgebiet für Rotschenkel *Tringa totanus*: Wertvolle Rückzugsgebiete oder ökologische fallen? *Vogelwarte 46*, 121–130.

Urban MC, **Bocedi G**, **Hendry AP**, **Mihoub J-B**, **Pe'er G**, **Singer A**, **Bridle JR**, **Crozier LG**, **De Meester L**, **Godsoe W**, **Gonzalez A**, **Hellmann JJ**, **Holt RD**, **Huth A**, **Johst K**, **Krug CB**, **Leadley PW**, **Palmer SCF**, **Pantel JH**, **Schmitz A**, **Zollner PA and Travis JMJ** (2016) Improving the forecast for biodiversity under climate change. *Science 353*, aad8466.

Urrea V and Calle ML (2012) AUCRF: Variable Selection with Random Forest and the Area under the Curve. R package version 1.1. Available at http://CRAN.R-project.org/package=AUCRF (accessed 25 July 2017).

Valavi R, **Elith J**, **Lahoz-Monfort JJ and Guillera-Arroita G** (2021a) Modelling species presence-only data with random forests. *Ecography 44*, 1731–1742.

Valavi R, **Guillera-Arroita G**, **Lahoz-Monfort J and Elith J** (2021b) Predictive performance of presence-only species distribution models: A benchmark study with reproducible code. *Ecological Monographs 92*, e01486.

van der Wal R and Palmer SCF (2008) Is breeding of farmland wading birds depressed by a combination of predator abundance and grazing? *Biology Letters 4*, 256–258.

Wiebe KL and Martin K (1998) Costs and benefits of nest cover for ptarmigan: Changes within and between years. *Animal Behaviour 56*, 1137–1144.

Wiens JA (1976) Population responses to patchy environments. *Annual Review of Ecology and Systematics 7*, 81–120.

Wiens JA (1989) Spatial scaling in ecology. *Functional Ecology 3*(4), 385–397.

Wiens JA, **Rotenberry JT and Van Horne B** (1987) Habitat occupancy patterns of north American shrubsteppe birds - the effects of spatial scale. *Oikos 48*(2), 132–147.

Woodward ID, **Massimino D**, **Hammond MJ**, **Barber L**, **Barimore C**, **Harris SJ**, **Leech DI**, **Noble DG**, **Walker RH**, **Baillie SR and Robinson RA** (2020) *BirdTrends 2020: Trends in Numbers, Breeding Success and Survival for UK Breeding Birds.* BTO Research Report 732, BTO, Thetford.

Wright MN and Ziegler A (2017) Ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software 77*, 1–17.

Xu-Ying L, **Jianxin W and Zhi-Hua Z** (2009) Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics 39*, 539–550.

Zuur AF, **Ieno EN**, **Walker NJ**, **Saveliev AA and Smith GM** (2009) *Mixed Effects Models and Extensions in Ecology with R.* New York: Springer.

Cite this article: Bradter U. Altringham JD Kunin WE Thom TJ O'Connell J. and Benton TG (2022). Variable ranking and selection with random forest for unbalanced data. *Environmental Data Science*, 1: e30. doi:10.1017/eds.2022.34

https://doi.org/10.1017/eds.2022.34 Published online by Cambridge University Press