Received: 22 January 2022 | Revised: 14 July 2022 | Accepted: 19 July 2022

DOI: 10.1111/ddi.13619

RESEARCH ARTICLE

Diversity and Distributions WILEY

Modelling the distribution of rare invertebrates by correcting class imbalance and spatial bias

Willson Gaul[1,2,3] | Dinara Sadykova[4,5] | Hannah J. White[1,2,6] | Lupe León-Sánchez[4] | Paul Caplat[4] | Mark C. Emmerson[4] | Jon M. Yearsley[1,2]

[1]School of Biology and Environmental Science, University College Dublin, Dublin, Ireland

[2]Earth Institute, University College Dublin, Dublin, Ireland

[3]Northern Marianas College, Saipan, Northern Mariana Islands, USA

[4]School of Biological Sciences, The Queen's University Belfast, Belfast, UK

[5]UK Centre for Ecology and Hydrology, Wallingford, UK

[6]School of Life Sciences, Anglia Ruskin University, Cambridge, UK

Correspondence
Jon M. Yearsley, School of Biology and Environmental Science, University College Dublin, Dublin, Ireland.
Email: jon.yearsley@ucd.ie

Funding information
Science Foundation Ireland, Grant/Award Number: 15/IA/2881

Editor: Stefano Mammola

Abstract

Aim: Soil arthropods are important decomposers and nutrient cyclers, but are poorly represented on national and international conservation Red Lists. Opportunistic biological records for soil invertebrates are sparse, and contain few observations of rare species but a relatively large number of non-detection observations (a problem known as class imbalance). Robinson et al. (*Diversity and Distributions*, **24**, 460) proposed a method for under-sampling non-detection data using a spatial grid to improve class balance and spatial bias in bird data. For taxa that are less intensively sampled, datasets are smaller, which poses a challenge because under-sampling data removes information. We tested whether spatially stratified under-sampling improved prediction performance of species distribution models for millipedes, for which large datasets are not available. We also tested whether using environmental predictor variables provided additional information beyond what is captured by spatial position for predicting species distributions.

Location: Island of Ireland.

Methods: We tested the spatially stratified under-sampling method of Robinson et al. (*Diversity and Distributions*, **24**, 460) by using biological records to train species distribution models of rare millipedes.

Results: Using spatially stratified under-sampled data improved species distribution model sensitivity (true positive rate) but decreased model specificity (true negative rate). The spatial pattern of under-sampling affected model performance. Training data that was under-sampled in a spatially stratified way sometimes produced worse models than did data that was under-sampled in an unstratified way. Geographic coordinates were as good as or better than environmental variables for predicting distributions of one out of six species.

Main Conclusions: Spatially stratified under-sampling improved prediction performance of species distribution models for rare millipedes. Spatially stratified under-sampling was most effective for rarer species, although unstratified under-sampling was sometimes more effective. The good prediction performance of models using geographic coordinates is promising for modelling distributions of poorly studied species for which little is known about ecological or physiological determinants of occurrence.

KEYWORDS

class imbalance, Diplopoda, millipede, rare species, spatial bias, spatial under-sampling, species distribution model

1 | INTRODUCTION

Biological records datasets contain relatively few records of rare species, because rare species generally have lower abundances and occur in fewer locations than common species. Rare species that have behavioural or physical characteristics (e.g. small size) that make them difficult to find or identify are even less well represented in datasets (Boakes et al., 2016). Two common problems for modelling rare species distributions using biological records are class imbalance (He & Garcia, 2009) and spatial bias in the data. Spatial bias is pervasive at many spatial scales in biological records data for all taxa (Amano & Sutherland, 2013; Oliveira et al., 2016). Class imbalance—usually in the form of a preponderance of non-detection observations and few detections of the focal species—is a problem more restricted to species that are rare, difficult to find or difficult to identify.

Robinson et al. (2018) proposed a method of under-sampling opportunistic species occurrence data in a spatially stratified way that improves both class balance and spatial bias in data before modelling distributions of rare bird species. Their innovation was to use a spatial grid to filter only non-detection data, of which there is plenty, while keeping all detection data, of which there is little for rare species. This approach improves both class balance and spatial bias in the data, and avoids the risk of removing too much information about the locations where rare species exist when filtering data to decrease spatial bias (El-Gabbas & Dormann, 2018; Fourcade et al., 2014). Fithian and Hastie (2014) described class imbalance as a problem for large datasets (they used a simulated dataset with a sample size of 10^6) that require large computational resources. By removing redundant observations from the majority class, the computational burden is reduced with minimal loss of information (Fithian & Hastie, 2014). Robinson et al. (2018) demonstrated spatially stratified under-sampling by modelling the distribution of a rare bird species in California, USA, using eBird data (https://ebird.org). Robinson et al.'s (2018) datasets of 302,655 observations and 108,880 observations were much smaller than the simulated datasets in Fithian and Hastie (2014). We tested spatially stratified under-sampling to model distributions of six millipede species in Ireland using a dataset even smaller than that used by Robinson et al. (2018).

Previous studies have found that spatial sampling bias is most problematic for SDMs when the spatial bias in non-detection or background data does not match spatial bias in detection data (Phillips et al., 2009). Using presence-background SDM techniques still requires determining how many background points to use (a class balance problem) (Barbet-Massin et al., 2012), and how they should be spatially arranged (a spatial bias problem) (Barbet-Massin et al., 2012; Phillips et al., 2009). One strategy to reduce the effects of spatial sampling bias in presence data when using presence-background SDMs that use artificially generated background or pseudo-absence points is to generate the background points with a spatial bias that matches the spatial bias in the detection points (e.g. target group approach for MaxEnt, Phillips et al., 2009). In contrast, spatially stratified under-sampling reduces the spatial bias in non-detection data, while leaving unchanged the spatial bias in detection data (Robinson et al., 2018). Given the minimal impact of spatial sampling bias on SDMs when the biases are similar in detection and non-detection data (Gaul et al., 2020; Johnston et al., 2020; Thibaud et al., 2014), and the potential negative impact of having spatial bias in detection data that differs from the bias in non-detection data (Phillips et al., 2009), it seems possible that manipulating the spatial bias in non-detection data but not in detection data during spatially stratified under-sampling might make SDMs worse, not better. Robinson et al. (2018) compared SDMs trained with raw and with spatially stratified under-sampled data, but did not test models trained with data that were under-sampled in a spatially unstratified way. Under-sampling in an unstratified way, in which non-detection data are chosen randomly, rather than using a spatial grid, would improve class balance while preserving the spatial bias in the data, which may result in better SDM performance because biases in detection and non-detection data would remain similar (Phillips et al., 2009).

Few invertebrates (with the possible exception of butterflies) will ever have datasets as large as those available for birds (Heberling et al., 2021). To the best of our knowledge, ours is the first test of spatially stratified under-sampling to improve species distribution model (SDM) predictions using such a small dataset, and thus provides important insight about how relevant this method is for non-charismatic, poorly recorded taxa.

Invertebrates are poorly represented in conservation research (Donaldson et al., 2016), on national and international lists of threatened and endangered species, including the IUCN Red List (Cardoso et al., 2011, 2012), and receive less conservation funding than do vertebrates (Mammola et al., 2020). Invertebrates play key roles in many ecosystem processes, including decomposition and nutrient cycling in soil (Bardgett, 2005; Bardgett & Wardle, 2010), pollination (Potts et al., 2016) and structuring ecosystems (Risch et al., 2018). Evaluations of extinction risk in invertebrates largely depend on knowledge of species distributions (e.g. criteria B and D of IUCN;

IUCN, 2012). Species distribution modelling methods that produce reliable predicted distributions can aid conservation threat assessment (Cardoso et al., 2011; Maes et al., 2015). For invertebrates, this will often require modelling distributions using datasets much smaller than the datasets available for more easily recorded taxa such as birds.

Biological records data for millipedes in Ireland are not nearly as extensive as data for some other taxa such as birds and vascular plants, but there have been two relatively intense periods of millipede recording in Ireland, culminating in a millipede distribution atlas (Lee, 2006). The data were vetted (and largely collected) by regional experts (Lee, 2006), so the dataset is of high taxonomic quality.

For millipedes in our study area, the direct environmental drivers of species occurrence (Austin, 2007) are not well known. Lee (2006) provided a comprehensive analysis of habitat affinities for millipedes in Great Britain and Ireland based on habitat data recorded as part of a British and Irish millipede recording scheme. But sampling in the recording scheme was opportunistic, and low numbers of records for some habitats made inference about habitat associations imprecise (Lee, 2006). The associations examined in Lee (2006) were largely land use, habitat and soil characteristics—climatic variables were not tested. Kime (1999, 2001, 2004) discussed the environmental determinants of millipede species distributions in the United Kingdom, Ireland and continental Europe using descriptive evaluations of locations of records and distribution patterns, but did not perform statistical analyses. Previous analyses (Kime, 1999, 2001, 2004; Lee, 2006) therefore provide only a starting point for identifying the direct drivers of Irish millipede species distributions; it seems likely that those studies did not identify all important environmental drivers, and it is possible that some drivers were incorrectly identified as important.

Even with well-studied taxa, the ability of modellers and models to identify biologically meaningful environmental predictors is limited (Beale et al., 2008; Currie et al., 2019). In a study of North American breeding birds, Bahn and McGill (2007) found that SDMs using only geographic coordinates performed better than models using environmental predictors. Fourcade et al. (2018) found that distributions of European Red Listed species were predicted nearly as well when the colour values of paintings were used as predictors as when environmental predictors were used. This suggests that environmental variables might not be capturing anything more than spatial autocorrelation from sources that are either exogenous (e.g. correlation in environmental drivers) or endogenous (e.g. dispersal) to the modelled species. The ability to accurately predict species distributions using only spatial information (Bahn & McGill, 2007) presents an opportunity for predictive modelling of taxa such as millipedes, for which physiological and ecological knowledge is poor. In contrast, the reliability of inferences about the effects of environmental drivers of species distributions are cast into doubt if models using purely spatial information predict distributions as well as models using environmental variables.

We modelled the distribution of six millipede species in Ireland using biological records and the spatially stratified under-sampling method of Robinson et al. (2018). We asked the following questions. (1) Does spatially stratified under-sampling of training data improve the predictive performance of SDMs for rare millipedes? (2) Does spatially stratified under-sampling of training data improve predictive performance of SDMs more than unstratified under-sampling? (3) Does the effectiveness of using spatially stratified under-sampled training data depend on the rarity of the species? (4) Do models using environmental predictors have better prediction performance than models using geographic coordinates as predictor variables? (5) Does spatially stratified under-sampling change the apparent relative importance of predictor variables compared to models trained with raw data?

2 | METHODS

We modelled the distribution of rare millipedes in Ireland using random forests (Breiman, 2001), which can model non-linear relationships and interactions between predictor variables, and were used by Robinson et al. (2018). To determine what types of information (environmental, spatial or seasonal) were important for predicting millipede detections, we trained four types of models: (1) a seasonal base model (SEASON + LIST LENGTH) that predicted millipede occurrence records as a function of month and a "checklist length" variable (details below) that we expected would capture variability in sampling effort among checklists; (2) a spatial model (COORDINATES + SEASON + LIST LENGTH) that used the base model plus geographic coordinates (eastings and northings of the TM75 Irish Grid Reference system); (3) an environmental model (ENVIRONMENT + SEASON + LIST LENGTH) that used the base model plus environmental covariates; and (4) an environmental and spatial model (ENVIRONMENT + COORDINATES + SEASON + LIST LENGTH) that used the base model plus environmental covariates and geographic coordinates.

We trained all models with three arrangements of the data: raw data, in which there was considerable class imbalance and spatial bias; unstratified under-sampled data, in which non-detections were randomly discarded to improve class balance; and with spatially stratified under-sampled data (Robinson et al., 2018) in which both class imbalance and spatial bias were adjusted (Figure 1, Figures S1 and S2).

2.1 | Study species

We modelled distributions of four rare millipede species detected on fewer than 10% of checklists: *Macrosternodesmus palicola* Brölemann, 1908; *Boreoiulus tenuis* (Bigler, 1913); *Ommatoiulus sabulosus* (Linnaeus, 1758); and *Blaniulus guttulatus* (Fabricius, 1798), and two more common species: *Glomeris marginata* (Villers, 1789) and *Cylindroiulus punctatus* (Leach, 1815). *Cylindroiulus punctatus* was the most commonly recorded species in our dataset

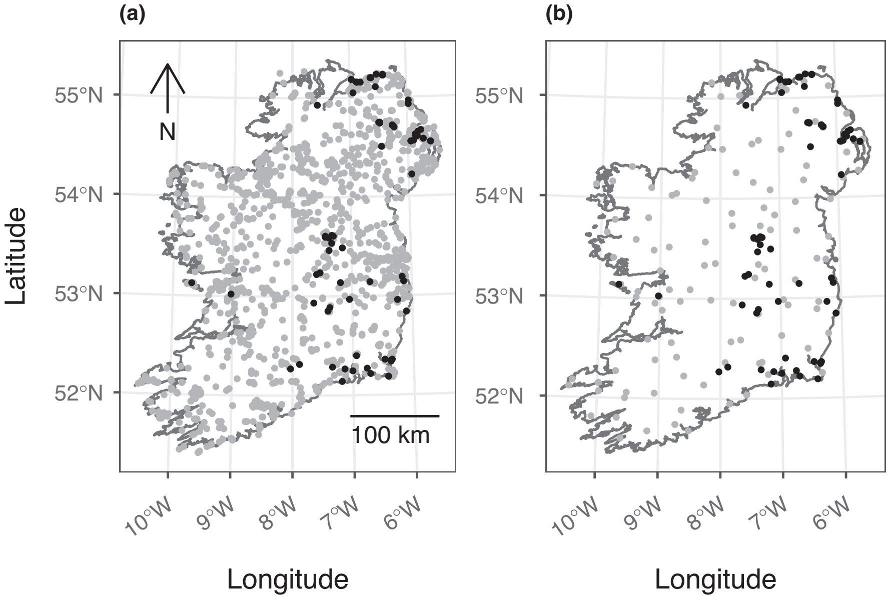

FIGURE 1 Raw (a) and spatially stratified under-sampled (b) observation data for the millipede *Ommatoiulus sabulosus* on the island of Ireland. Spatially stratified under-sampling involved keeping all checklists on which the species was detected (black points), but spatially filtering the non-detection checklists (grey points) by randomly choosing only a single non-detection checklist from each cell of a 30×30 km grid that was randomly positioned over the study extent (grid not shown on these figures). Spatially stratified under-sampling improved class balance and reduced the spatial bias of the non-detection data.

(Table 1). Three of the species (*M. palicola*, *Boreoiulus tenuis* and *Blaniulus guttulatus*) are believed to be somewhat synanthropic, while the other three (*O. sabulosus*, *G. marginata* and *C. punctatus*) are not (Lee, 2006).

All of the species we modelled take multiple years to reach maturity in our study area, so they should be present year round. However, millipedes in Ireland show seasonal patterns of activity and detectability. For example, many species move deeper into leaf litter or soil to avoid cold temperatures or dry conditions (Lee, 2006). Differences in maturation speeds, life spans and activity patterns between species (and between sexes within species) mean that the number of individuals, number of adults and relative proportions of the sexes may not be constant at all times of year. Adults are generally easier to identify than juveniles. For some species, only sexually mature males can be identified to species level based on morphology (i.e. without molecular evidence), which means both the total number of species that can be identified and recorded, and the probability of recording any particular species, are likely to change over the course of a year in any given location.

2.2 | Millipede occurrence data

We downloaded records of all millipedes (including but not limited to our six focal species) on the island of Ireland for the years 1971 to 2020 from the Global Biodiversity Information Facility (GBIF.org, 2021). The data included records from multiple sources, including the British Myriapod and Isopod Group recording scheme (Biological Records Centre, 2017; Lee, 2006). The data were presence-only records of millipede species occurrence, and did not include explicit sampling effort, sampling method or non-detection information. We grouped records into recording event "checklists," where a checklist was defined as a unique combination of date, location and observer, and each species was either detected or not detected (van Strien et al., 2010). We calculated checklist length by counting the number of species detected on each checklist. We retained for analysis only checklists with spatial precision of 1 km or less as reported in the downloaded data ($n = 1757$ checklists).

2.3 | Environmental data

The species we selected are believed to respond to a variety of land cover and habitat characteristics, soil types and human disturbance (Lee, 2006). We identified (when possible) remotely sensed environmental variables that corresponded to the strongest habitat affinities reported for each focal species in Lee (2006). All six of our focal species were reported to respond strongly to urban vs. rural land use classifications in Lee (2006). Four species (*M. palicola*, *B. guttulatus*, *G. marginata*, *C. punctatus*) showed strong relationships with woodland land cover in Lee (2006). At least one species, *G. marginata*, may be unable to tolerate low temperatures (Lee, 2006), which might exclude it from higher elevations in Ireland.

The environmental variables we used included elevation, two climate variables and five land cover variables (Table 1). We calculated the value of each predictor variable in 1×1 km grid cells covering Ireland. We used the mean elevation of each grid square (calculated by interpolating using ordinary kriging) from the ETOPO1 Global Relief Model (Amante & Eakins, 2009). For the land cover variables, we calculated the proportion of each grid cell covered by "artificial surfaces," "forest and semi-natural areas," "wetlands," "pasture" and "arable land" classes from the CORINE Land Cover database (CORINE, 2012). We downloaded gridded climate variables from the E-OBS European Climate Assessment and Dataset EU project (Haylock et al., 2008; van den Besselaar et al., 2011), and, for each 1 km^2 grid cell, calculated the mean annual precipitation for the years 1995 to 2016 (excluding years 2010 through 2012 because of missing data), and the mean annual low temperature across

TABLE 1 Environmental predictor variables used to model millipede species in Ireland. For each millipede species, the number of positive detections with a spatial precision of 1 km or less is shown, along with the proportion of all checklists on which the species was detected. The environmental predictor variables used for each species are indicated with an "X".

Species	Number of detections	Proportion of checklists with a detection	Mean annual minimum temperature	Mean annual precipitation	Elevation	Artificial surfaces land cover	Arable land cover	Wetlands land cover	Forest and semi-natural areas land cover	Pasture land cover
Macrosternodesmus palicola	61	0.035		X		X	X			
Boreoiulus tenuis	73	0.042		X	X	X	X			
Ommatoiulus sabulosus	77	0.044		X	X	X		X		
Blaniulus guttulatus	143	0.081	X	X	X	X	X		X	
Glomeris marginata	325	0.185	X	X	X	X	X		X	
Cylindroiulus punctatus	615	0.350	X	X	X	X	X		X	X

years 1995 to 2016. Annual low temperature was taken to be the 2% quantile of daily temperatures from a year. The 2% quantile was used to prevent erroneous extreme data values from influencing the results. Low temperature was used rather than other temperature variables because Kime (1999) expected low temperature to be an important determinant of millipede distribution in northern Europe, while high summer temperature was expected to be an important determinant in southern Europe. Cold winters are believed to limit the distribution of *G. marginata* (Kime, 2004), though other species have behavioural responses (e.g. burrowing in soil or dead wood) that allow them to survive cold periods (Kime, 2004).

2.4 | Spatially stratified under-sampling and unstratified under-sampling

Our spatially stratified under-sampling process followed Robinson et al. (2018). We first split the millipede checklists into two groups: non-detection checklists, on which the focal species was not detected, and detection checklists, on which the focal species was detected. We then generated a randomly positioned 30 × 30 km grid over the study extent using the "blockCV" R package (Valavi et al., 2019). We under-sampled the non-detection checklists by randomly selecting one non-detection checklist from each 30 × 30 km grid cell (provided a grid cell contained at least one non-detection checklist). We kept all detection checklists. We then combined the spatially stratified under-sampled non-detection checklists with the detection checklists to create a spatially stratified under-sampled dataset. We used this spatially stratified under-sampled dataset as training data for SDMs. We evaluated spatial evenness in the raw and spatially stratified under-sampled data by measuring Simpson's evenness for the number of checklists in 30 × 30 km grid squares.

To assess the separate effects of adjusting the class balance and the effects of adjusting the spatial pattern of non-detection data, we created a dataset in which under-sampling was done in an unstratified, spatially naive way. Non-detection checklist were under-sampled randomly from the entire study extent until the class balance matched the class balance in spatially stratified under-sampled datasets.

All spatial data processing used R version 3.6 (R Core Team, 2020) and the packages "sf" (Pebesma, 2018), "fasterize" (Ross, 2018), "raster" (Hijmans, 2018), "rgdal" (Bivand et al., 2018), "gstat" (Gräler et al., 2016) and "tidyverse" (Wickham, 2017).

2.5 | Species distribution models

We trained random forest SDMs with the "randomForest" function in R (Liaw & Wiener, 2002) using threefold cross-validation (three folds were used because using five folds often resulted in folds with no or few detections, which made model testing difficult or impossible in those folds). The units of analysis were sampling event checklists (n = 1757), on which each species was either detected or not detected (van Strien et al., 2010). We modelled species detections as a function of predictor variables expected to influence species occupancy and/or detectability. We did not separately estimate occupancy and detectability, as is done in hierarchical occupancy/detection models (MacKenzie et al., 2002) because our data contained few repeat survey visits within a year, but we included predictor variables that we expected would primarily influence either occupancy or detectability. Occupancy covariates were geographic coordinates and environmental covariates. The month of each observation was included primarily as a detectability covariate because most millipede species in our study have seasonal changes in behaviour and/or seasonal life cycles that make them easier to detect and identify at some times of year. To allow for the periodic variation in detectability with month, we represented months as integers (1 through 12) and used cosine and sine transformations of month to create two separate transformed month variables that we provided to the random forest SDMs (James, 2011; see Appendix S1). Checklist length was used as a proxy for sampling effort (Isaac et al., 2014; Szabo et al., 2010) and was thus primarily a covariate for detectability, though checklist length likely also varied with environmental conditions and species richness in our study (see Section 4), and is therefore potentially related to occupancy as well as detectability. We expected the probability of detecting each focal species to increase with checklist length.

We used different environmental predictor variables for each species based on the environmental and habitat affinities reported in Lee (2006). The limited amount of occurrence data available for rare millipede species made over-fitting a concern. Each model included variables indicating the month in which the record was collected, and checklist length. For models including environmental covariates, we selected the number of predictor variables to use in each model based on the number of positive detections for each species, so that there were at least 10 detections of the focal species per predictor variable (Table 1).

We calculated prediction performance measures on spatially stratified under-sampled test datasets, because the goal of SDMs was to predict species occurrence at all locations in Ireland, where all locations are equally important (Robinson et al., 2018). Using spatially stratified under-sampled test data reduces the extent to which prediction performance measures are dominated by how models predict in the most heavily sampled areas (Fink et al., 2010; Robinson et al., 2018). Each checklist in the raw data was randomly assigned to one of three cross-validation folds. We used the same cross-validation folds to train and evaluate all types of models, so that all models were tested on identical test data. We used random forests for classification to predict the detection or non-detection of the focal species on each checklist. For each random forest model, we grew 1000 trees with a terminal node size of one, using the largest integer less than the square root of the number of predictor variables as the number of variables to consider for splitting at each split. For each model (each combination of species, data type and model type), we performed 33 modelling runs, each time fitting models with threefold cross-validation. This produced a total of 1188 fitted models for each species (3 CV folds × 33 modelling runs × 3 data types × 4 model types = 1188 models). We generated a new spatially stratified under-sampled dataset for each of the 33 model fitting iterations.

We assessed prediction performance of each fitted model by predicting to checklists in the cross-validation test fold. We measured the ability of models to accurately discriminate between detections and non-detections using the area under the receiver operating characteristic curve (AUC) (Fielding & Bell, 1997), Cohen's Kappa (Cohen, 1960), and sensitivity (true positive rate) and specificity (true negative rate) at the threshold that maximized Cohen's Kappa. Sensitivity measured the ability of models to correctly predict which checklists had detections, while specificity measured the ability of models to correctly predict which checklists had non-detections. We also measured the calibration of the predicted probabilities of models using the Brier score (Brier, 1950) following Robinson et al. (2018).

We averaged the variable importance (mean decrease in node impurity over all trees due to splitting on each variable, measured using the Gini index) and partial dependence measures produced by random forest models over all 99 iterations of each model (see Appendix S1). We made predicted distribution maps for each species using the most complex model (ENVIRONMENT + COORDINATES + SEASON + LIST LENGTH), fixing the checklist length at two (the median in the observed data) and generating predictions for every month. We then averaged the monthly predictions over the entire annual cycle from all 99 model iterations to get the average predicted probability of detecting the focal species on a checklist of length two for each grid cell. The variability in model predictions for each grid cell was visualized by mapping the standard error of the mean annual predictions from the 99 model iterations (Fink et al., 2014). The standard errors show the variability in the mean prediction due to differences in which data were included in the cross-validation training sets, but do not show variation in predictions due to changes over the annual cycle.

3 | RESULTS

3.1 | Spatially stratified under-sampling

Spatially stratified under-sampling improved both the class balance (Figure S1) and spatial evenness (Figure S2) of training data for all species. Using spatially stratified under-sampled training data generally improved overall discrimination performance (AUC) for rarer species (Figure 2), though there was minimal or no improvement for some models of the rarest species (Figure 2). For the two more common species (*G. marginata* and *C. punctatus*), spatially stratified under-sampling did not improve prediction performance as much as it did for rarer species, and notably reduced the overall performance of the simplest model (Figure 2).

Prediction performance of models trained with spatially stratified under-sampled data differed from performance of models trained with unstratified under-sampled data (Figures 3 and 4a–f).

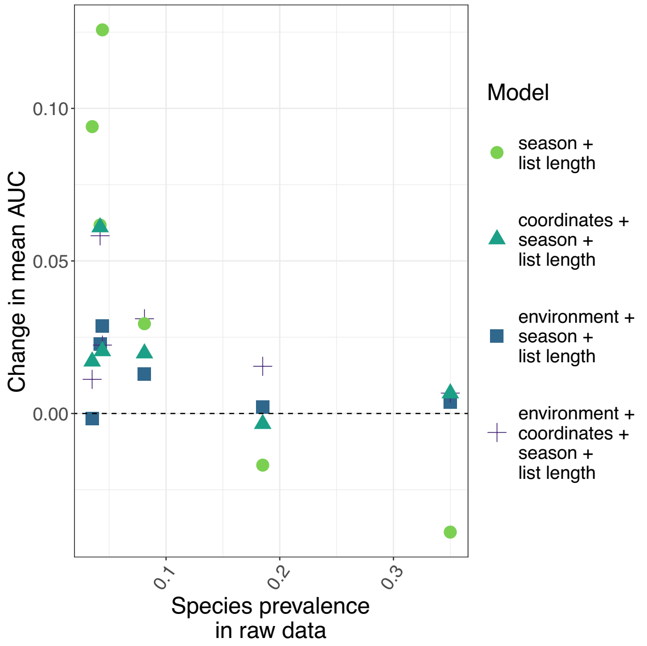

FIGURE 2 Change in mean prediction performance as a function of species prevalence in the original data, when species distribution models (SDMs) were trained using spatially stratified under-sampled rather than raw data. Results are shown for random forest SDMs for six millipede species in Ireland. Points above the horizontal dotted line indicate that spatially stratified under-sampling the training data improved model prediction performance. Prediction performance was measured using the area under the receiver operating characteristic curve (AUC). Four models with different sets of predictor variables were tested.

Unstratified under-sampling improved performance of models for rare species, but reduced performance of most models for more common species (Figure 3a). Spatially stratified under-sampling, in which the same class balance was achieved by removing non-detections in a spatially stratified way, provided additional improvements in prediction performance for most (but not all) models, beyond the improvements due to adjusting class balance (Figure 3b). For common species, spatially stratified under-sampling generally improved model performance relative to unstratified under-sampling. This offset the negative effects of adjusting class balance for common species, so that models trained with spatially stratified under-sampled data had prediction performance roughly similar to models trained with raw data for common species (Figure 2). For the rarest species, spatially stratified under-sampling reduced model performance for some species but improved model performance for other species, relative to models trained with unstratified under-sampled data (Figure 3b).

Prediction performance of the most complex model (ENVIRONMENT + COORDINATES + SEASON + LIST LENGTH) was generally improved by spatially stratified under-sampling training data according to most performance metrics, including threshold-dependent (Kappa, sensitivity) and -independent (AUC) discrimination metrics, and Brier score (Figure 5). Sensitivity was notably improved both by adjusting class imbalance and by improving the spatial evenness of the non-detection data (Figure 5). Both unstratified and spatially stratified under-sampling reduced model specificity. The decrease in specificity with under-sampled training data were greatest for the most common species and smallest for the rarer species (Figure 5). For rare species, unstratified under-sampling seemed to be an effective way of increasing model sensitivity without sacrificing too much specificity. Overall, the additional effects of performing the under-sampling in a spatially stratified way were not consistently positive for rare species (Figure 3).

Rankings of variable importance changed when models were trained with spatially stratified under-sampled rather than raw data (Figures S3 and S4).

3.2 | Environmental vs. spatial models

The spatial model (COORDINATES + SEASON + LIST LENGTH) was better than the environmental model (ENVIRONMENT + SEASON + LIST LENGTH) for *O. sabulosus* (Figure 4c).

The simplest model (SEASON + LIST LENGTH) generally performed worse than more complex models that included geographic coordinates, environmental variables or both (Figure 4, mean difference in AUC between the simplest and most complex models for each species when using spatially stratified under-sampled data = −0.09, range = −0.12 to −0.03). A notable exception was for the common species *C. punctatus*, for which the SEASON + LIST LENGTH model trained with raw data was among the best models (Figure 4f). This suggests that information about the location of a checklist was not important for predicting the probability of recording *C. punctatus*.

3.3 | Effects of covariates on species occurrence and detection

Partial dependence plots of the marginal effect of each variable from the most complex model showed plausible relationships. The probability of detecting the focal species on a checklist generally increased in a decelerating curve with checklist length, as expected (Figure 6a–h, Figures S5–S10). Checklist length was among the most important variables for all species except *O. sabulosus*, when assessing variable importance for the most complex model trained with spatially stratified under-sampled data (Figures S3 and S4).

Seasonal changes in the probability of detection were clearly visible in partial dependence plots for the month variable for the three rarest species, with increased probability of detection in winter for *M. palicola* (Figure S5) and *Boreoiulus tenuis* (Figure S6) and increased probability of detection in summer for *O. sabulosus* (Figure 6b, Figure S7). For *Blaniulus guttulatus*, seasonal patterns of detectability were less clear, but detectability appeared lowest in summer and highest in spring and fall (Figure S8). There were no clear seasonal patterns in detectability for *G. marginata* (Figure S9), or *C. punctatus* (Figure S10).

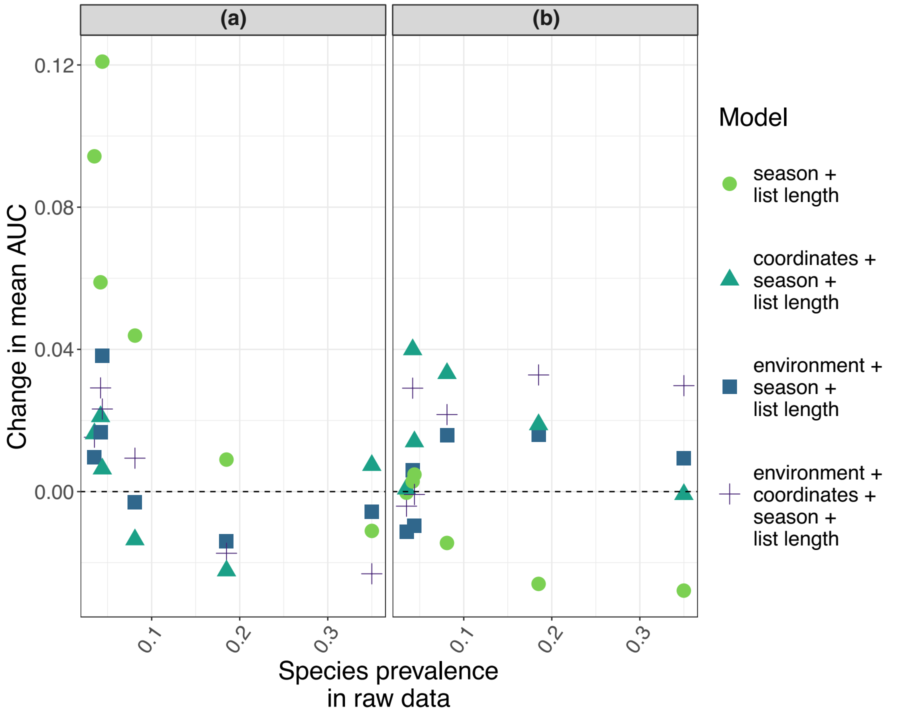

FIGURE 3 The separate effects of adjusting class balance (a) and spatial bias in the non-detection data (b) on prediction performance of species distribution models for Irish millipedes. Panel (a) shows the change in average AUC when models were trained with data that had been under-sampled in an unstratified way compared to models that had been trained with the raw data. Panel (b) shows the additional change in average AUC when models were trained with spatially stratified under-sampled data rather than unstratified under-sampled data. If the changes in performance shown in (a) are added to the changes shown in (b), the overall change in performance would be as shown in Figure 2.

4 | DISCUSSION

We set out to answer five questions: (1) Does spatially stratified under-sampling of training data improve the predictive performance of SDMs for rare millipedes? (2) Does spatially stratified under-sampling of training data improve predictive performance of SDMs more than unstratified under-sampling? (3) Does the effectiveness of spatially stratified under-sampling depend on the rarity of the species? (4) Do models using environmental predictors have better prediction performance than models using geographic coordinates as predictor variables? (5) Does spatially stratified under-sampling change the apparent relative importance of predictor variables compared to models trained with raw data? Briefly, the answers to these questions were: (1) yes; (2) usually, but not always; (3) yes; (4) usually, but not always and (5) usually, but not always.

4.1 | Adjusting spatial bias during under-sampling

Under-sampling data to address class imbalance has been explored in the machine learning literature, and is used in a wide variety of applied settings (reviewed in Haixiang et al., 2017). The innovation of Robinson et al. (2018) was to perform the under-sampling using a spatial grid to simultaneously improve the spatial evenness of data. We tested the usefulness of Robinson et al.'s (2018) method for modelling distributions of rare invertebrates using a small dataset. We also compared spatially stratified under-sampling to unstratified under-sampling in order to determine the effect of adjusting the spatial evenness of data, separate from the effect of adjusting the class balance. We found that adjusting class balance using unstratified under-sampling improved prediction performance for rare species, but the effects of performing under-sampling in a spatially stratified way were mixed. Manipulating the spatial pattern of the data during the spatially stratified under-sampling process sometimes erased the performance gains from improving class balance (Figure 3). Spatially stratified under-sampling improved most measures of prediction performance of our random forest SDMs, and was most effective for rarer species (Figure 2), but for some rare species models would have been even better if the under-sampling had been done in an unstratified way (Figure 3).

Adjusting the spatial pattern of the non-detection data during spatially stratified under-sampling caused changes in model performance that were additional to the changes caused by adjusting class balance (Figures 3–5), but those changes were not consistently helpful (Figure 3b). Robinson et al. (2018) and Robinson et al. (2020) used spatially stratified under-sampling, but did not compare it to unstratified under-sampling. Our results suggest that the effect of manipulating the spatial pattern of non-detection data to make it more spatially even is not always beneficial, and that unstratified under-sampling may sometimes be preferable to spatially stratified under-sampling when using spatially biased data. Phillips et al. (2009) described potential problems caused by training models with data in which spatial bias differs between detection points and "background" points. In Phillips et al. (2009), the difference in spatial bias between detection and background data is due to generating background points in a spatially even way, while using detection data that is spatially biased. During spatially stratified under-sampling, the process is different, but the end result is the same: the spatial biases in the data used to train models differ between the detection and

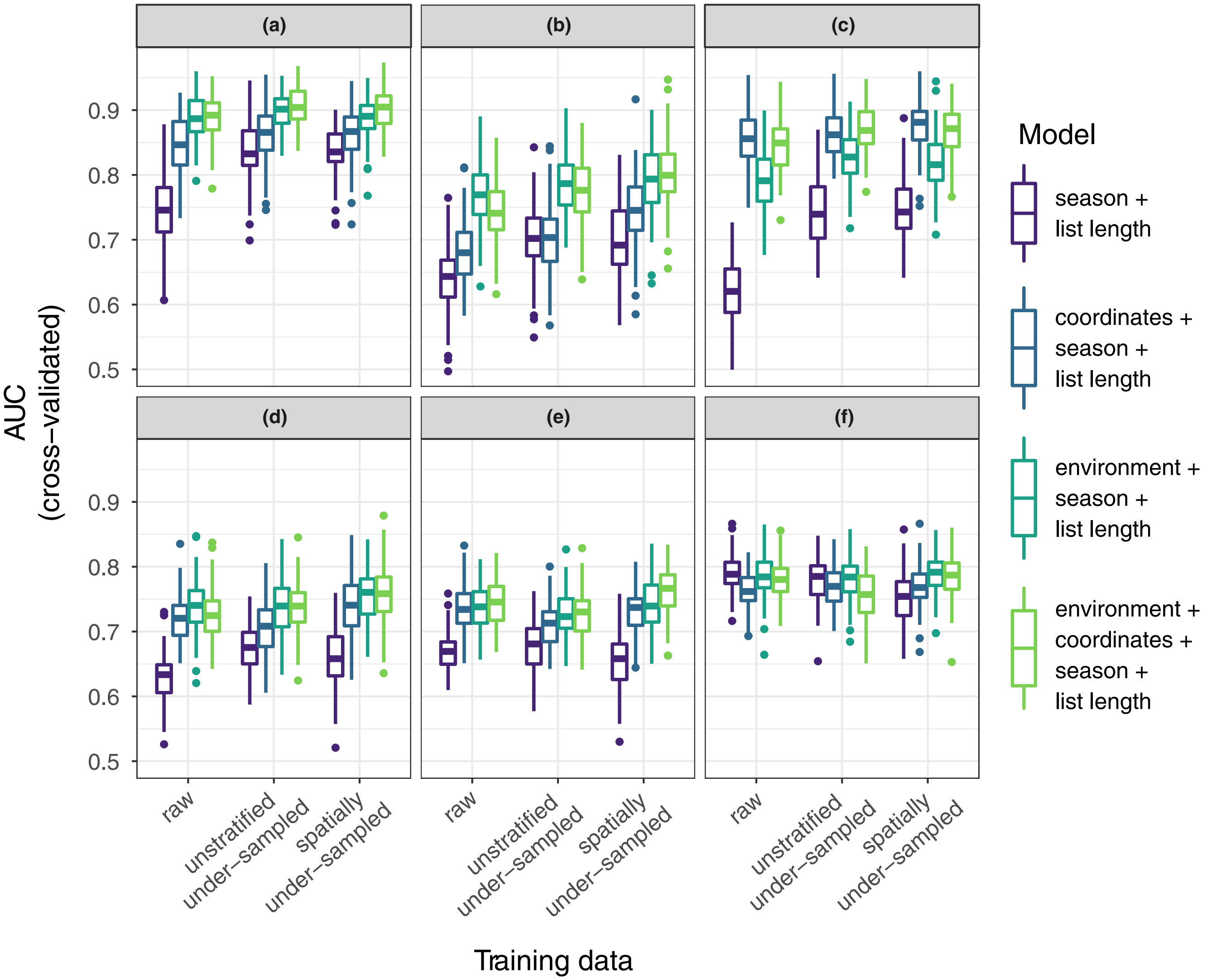

FIGURE 4 Prediction performance (AUC) of random forest species distribution models for six millipede species in Ireland. Results are shown for models trained with raw, unstratified under-sampled and spatially stratified under-sampled data (left, middle and right box plots, respectively, within each panel). The six modelled species, arranged from rarest to most common in our data, were *Macrosternodesmus palicola* (a), *Boreoiulus tenuis* (b), *Ommatoiulus sabulosus* (c), *Blaniulus guttulatus* (d), *Glomeris marginata* (e) and *Cylindroiulus punctatus* (f). Box plots show the distribution of AUC values for 99 replicates of each model; boxes contain the middle 50% of the data, the horizontal line within each box shows the median.

non-detection data. In our case study, we found inconsistent effects of manipulating the spatial bias in the non-detection data. In many cases, under-sampling the data in a spatially stratified way provided additional improvements in model performance (Figure 3b), but in a few cases the spatial stratification was harmful, and models were better when trained with unstratified under-sampled data. Like our study, Robinson et al. (2018) and Robinson et al. (2020) tested spatially stratified under-sampling using a case study, with real, opportunistically collected data. To the best of our knowledge, using spatially stratified under-sampled data to train SDMs has not been assessed using simulations or using systematically collected test data. We suggest that spatially stratified under-sampling be tested using data in which the truth is perfectly known (i.e. simulated data), or using systematically collected data, which would provide a better test of whether spatially stratified under-sampling improves the ability of SDMs to predict the true distribution of species.

4.2 | Sensitivity vs. specificity

Both spatially stratified and unstratified under-sampling increased sensitivity at the expense of specificity in our models. Under-sampling (spatially stratified or not) will therefore be most useful in applications where false positive predictions are not particularly problematic. High-sensitivity SDMs can guide targeted surveys for rare millipedes, which would help fill knowledge gaps and could facilitate production of a national Red List for Irish millipedes. High sensitivity SDMs could also be used to produce a “short list” of locations that are candidates for conservation or management of

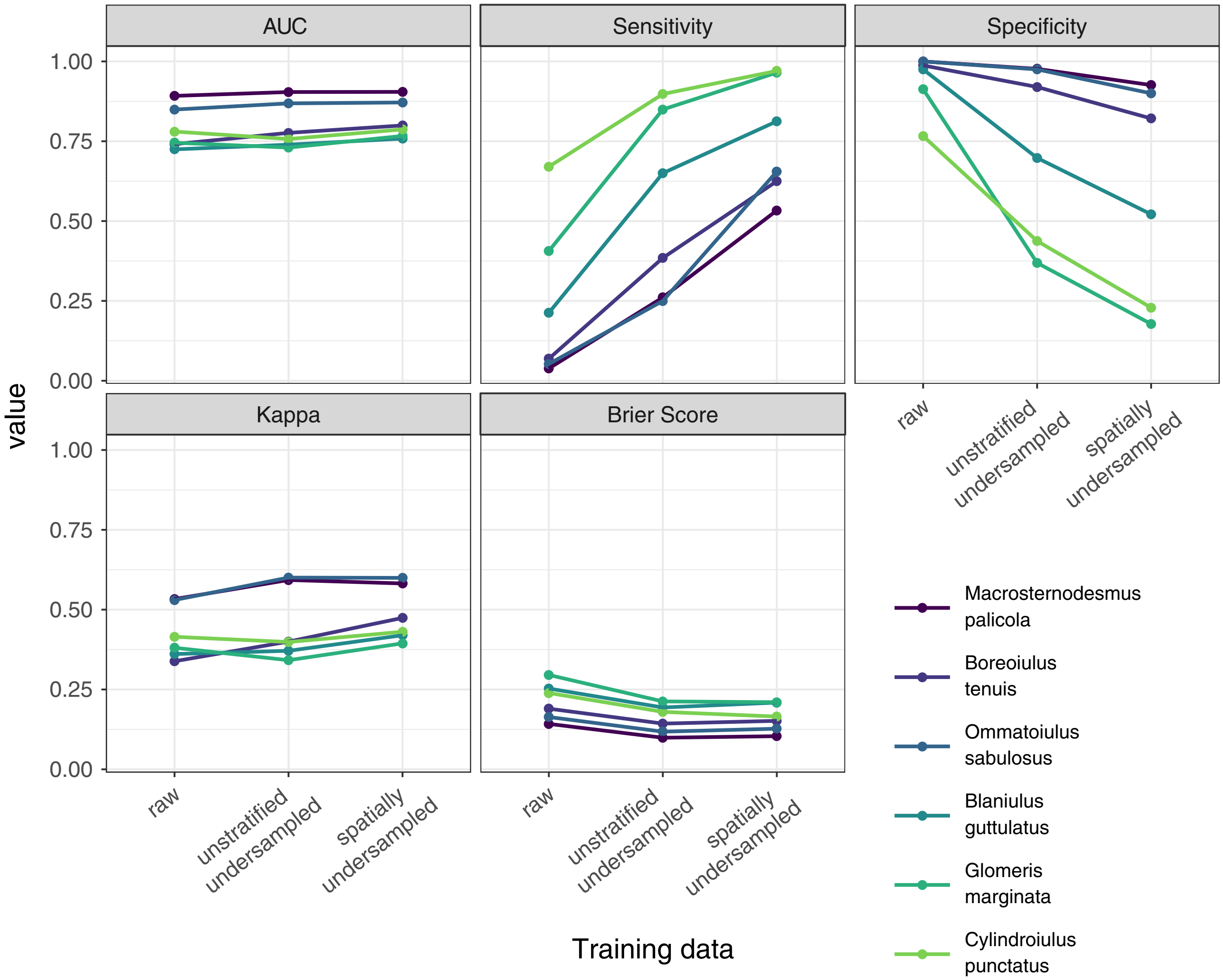

FIGURE 5 Prediction performance of species distribution models trained with raw (left), with unstratified under-sampled (middle) and with spatially stratified under-sampled (right) data for six species of millipede in Ireland. Points show the median value of each prediction performance measure from 99 replicate models (33 replicates of threefold cross-validation) fit to each species with each type of training data. Darker coloured points and lines indicate rarer species, and lighter colours indicate more common species. Sensitivity was substantially improved for all species when training data were under-sampled. Under-sampling caused an undesirable decrease in specificity for all species, but the decrease was smaller for rarer species and greater for more common species.

focal species. Locations on a "short list" could subsequently be surveyed more intensively to confirm focal species presence. Any false positive predictions (locations in which further surveying revealed that the species is absent despite the model's positive prediction), could then be excluded, resulting in a final list of locations for conservation action. False positive predictions from models trained with under-sampled data could be problematic if conservation or policy decisions ignore false positives. For example, deciding not to list a species as threatened because a model predicted a large distribution would be a bad decision if that model had high sensitivity and low specificity. The trade-off between sensitivity and specificity can be adjusted by changing the threshold value used to dichotomize predictions. The improvements we saw in AUC when training models with under-sampled data suggest that, in many cases, both spatially stratified and unstratified under-sampling resulted in models with better overall discrimination than did models trained with raw data. However, the improvements in AUC were small; in most practical applications the impact of improving model AUC by about 0.025 or less (as for most models in Figure 2) is minimal.

The millipede dataset we used had about 0.02 checklists per km^2—an order of magnitude less data than the dataset used by Robinson et al. (2018), which had about 0.26 to 0.71 checklists per km^2 (for their winter and summer models, respectively). Robinson et al. (2018) reported no loss of specificity when using spatially stratified under-sampled training data. In contrast, models for all of our species had some loss of specificity when using spatially stratified under-sampled training data. Perhaps the limited amount of information available in our smaller dataset meant that there was less "spare" non-detection data to be discarded before model specificity

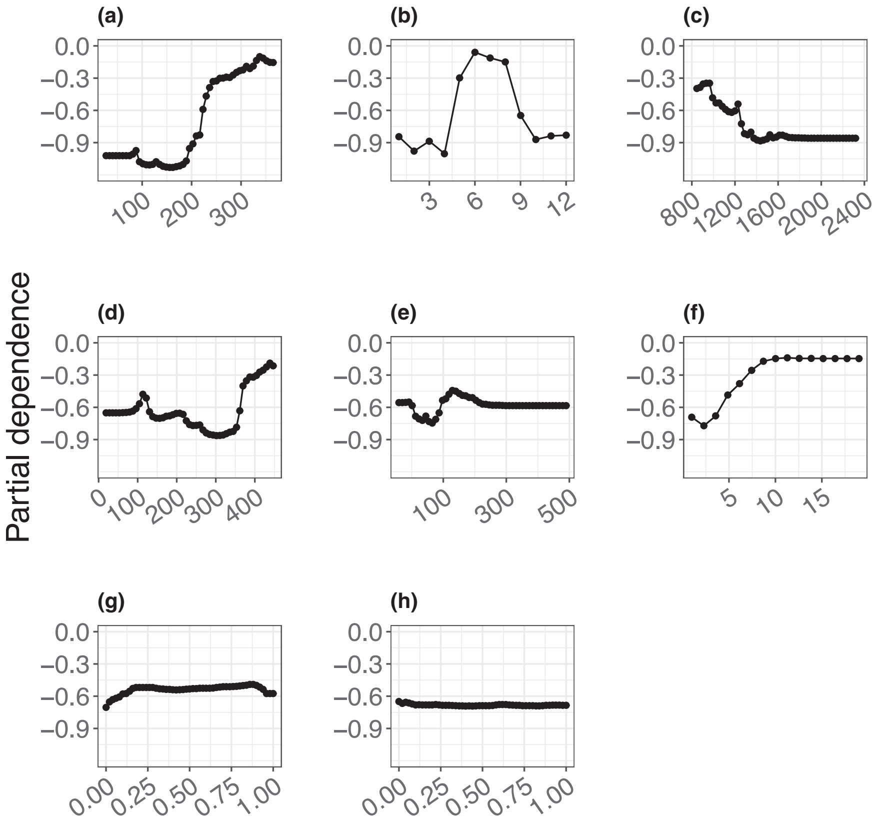

FIGURE 6 Partial dependence plots showing the effect of predictor variables on the probability of detecting the millipede *Ommatoiulus sabulosus* in Ireland. Results are shown for the *ENVIRONMENT + COORDINATES + SEASON + LIST LENGTH* model trained with spatially stratified under-sampled data. The vertical axes show the partial dependence measure (Appendix S1), with higher values indicating a higher probability of detecting the species. The horizontal axes show the value of the predictor variables: (a) kilometres east of the origin point of the TM75 Irish Grid Reference system; (b) month (one indicates January); (c) annual precipitation (mm); (d) kilometres north of the origin point of the TM75 Irish Grid Reference system; (e) elevation (m); (f) checklist length (number of records); (g) proportion of grid cell area covered by artificial surfaces; and (h) proportion of grid cell area covered by wetlands.

declined. Large reductions in specificity indicated that models trained with spatially stratified under-sampled data over-estimated distributions of the more common species in our study.

4.3 | Tuning the under-sampling procedure

Given the mixed performance of spatially stratified under-sampling in our study, it is worth exploring how to optimally tune the procedure, and when to avoid spatially stratified under-sampling. The size of the spatial grid used for stratified under-sampling, and the class balance of the under-sampled dataset could be systematically explored and tuned using cross-validation, as is done for other model parameters in machine learning settings (Hastie et al., 2009).

4.4 | Geographic coordinates as predictor variables

The spatial model (*COORDINATES + SEASON + LIST LENGTH*) clearly outperformed the environmental model (*ENVIRONMENT + SEASON + LIST LENGTH*) for *O. sabulosus*. For that species, the environmental covariates did not have any more predictive power than did information about spatial location. The distribution of *O. sabulosus* in Ireland might be determined primarily by non-environmental factors, such as dispersal or biotic interactions. Alternatively, it could be that *O. sabulosus* distribution is determined by environmental variables that were not in our model, but which were spatially auto-correlated so that geographic coordinates were effective proxies.

For all species, the model including both environmental variables and geographic coordinates as predictors was among the best models. Because our model evaluations used cross-validation, this is unlikely to be an artefact in which the most complex model appears best because it is over-fitted. We therefore suggest that a reasonable strategy for selecting variables to include in SDMs is to begin with geographic coordinates (and a sampling effort covariate such as checklist length, if appropriate), and then add environmental variables, limiting the number of environmental variables based on sample size to avoid overfitted models. Many studies have noted the benefits of explicitly including space in SDMs (Beale et al., 2010; Dormann, 2007; Lennon, 2000); we suggest taking advantage of the nearly universal presence of spatial autocorrelation in species distributions by including geographic coordinates as part of a "base model" to which environmental covariates can be added. Our goal in including geographic coordinates as predictor variables was to take advantage of spatial structure for prediction (Bahn & McGill, 2007), rather than to better estimate the effects of predictor variables or control for pseudo-replication or spatial structure in the errors (Beale et al., 2010). Using geographic coordinates as predictor variables has the advantage that decent predictive models can likely be constructed even for species for which the most relevant environmental drivers of distribution are not known.

The good performance of the spatial models is encouraging for predictive models, but discouraging for attempts to identify biologically meaningful environmental drivers of distributions. Geographic coordinates were regularly in the top half of variables ranked by importance for our SDMs (Figures S3 and S4), highlighting the fact that the usefulness of a variable for prediction does not provide insight about whether the variable is a causative determinant of distribution.

Dispersal is likely an important determinant of distributions for some millipede species. Millipedes may be dispersed long distances by humans in soil and plant material, but once established in new locations, local dispersal may be much slower (Baker, 1978), though some species can be mobile and disperse readily (David & Handa, 2010). Limited local dispersal could lead to fragmented distributions in which large parts of suitable environmental space are not occupied, obscuring any signal of environmental suitability from SDMs. Models using geographic coordinates as predictors may be better able to capture such fragmented distributions, and may predict well when interpolating between sampling locations, even though they cannot extrapolate to new geographic areas.

4.5 | Variable importance

The ranking of the predictor variables by relative importance in the random forest models changed when models were trained with spatially stratified under-sampled rather than raw data (Figures S3 and S4). Knowledge of the life history and ecology of millipedes in Ireland is patchy, with almost nothing known about some species (Lee, 2006). Variable importance measures (and partial dependence plots) from our SDMs can be used to suggest hypotheses about factors that influence the distribution and/or detectability of species (Kelling et al., 2009). However, because the variable importance rankings for most species changed when using spatially stratified under-sampled data, we have low confidence in interpreting variable importance rankings in terms of ecology.

4.6 | Checklist length as a proxy for sampling effort

We used checklist length as a proxy for sampling effort, but checklist length probably also varied with factors not related to sampling effort, including species richness (Warton et al., 2013). Checklist length can be used as a sampling effort covariate in the detection submodel within hierarchical occupancy-detection models (MacKenzie et al., 2002), or to account for detectability in non-hierarchical models (Isaac et al., 2014; Szabo et al., 2010). The intuition we followed is the same. To the extent that checklist length successfully captured variability in sampling effort, the predictions generated using a standardized checklist length (Figure 7a–d, Figures S11–S15) represent the relative probability of recording the focal species when sampling effort is constant in each grid cell. Checklist length was the most important variable in models for three out of our six focal species (Figures S3 and S4). Partial dependence plots showed that relationships between checklist length and relative probability of the focal species being recorded were generally positive, with

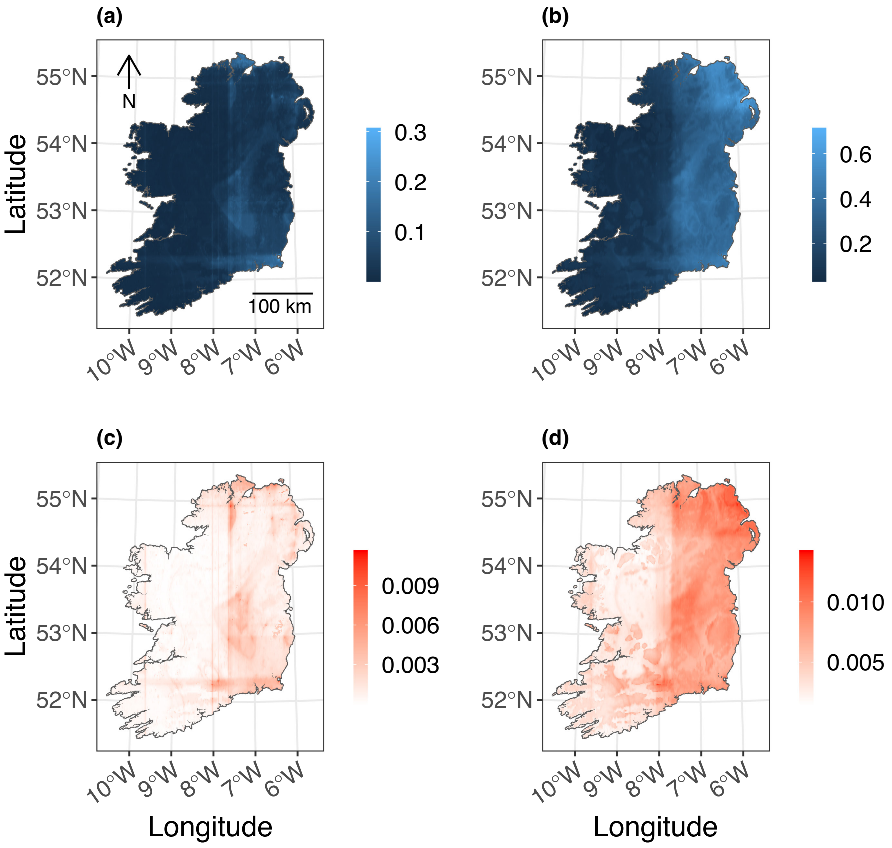

FIGURE 7 Predicted distribution of the millipede *Ommatoiulus sabulosus* in Ireland. The top maps show mean predicted relative probability of detecting *O. sabulosus* on a checklist of length two, from the *ENVIRONMENT + COORDINATES + SEASON + LIST LENGTH* model trained with (a) raw data and (b) spatially stratified under-sampled data. Bottom maps show: (c) standard error of the mean predictions from (a); and (d) standard error of the mean predictions from (b). The standard errors of the predictions in each grid cell (c) and (d) show how much model predictions varied based on which records were included or excluded from the cross-validation training dataset. Structure from the predictor variables is visible in model predictions and standard errors, for example as vertical and horizontal lines (artefacts of the geographic coordinate predictor variables) and contour lines (e.g. in the southeast quadrant of (a), an artefact of the precipitation variable).

probability of the species being recorded increasing more quickly when checklist length was small, as expected (Figure 6f, Figures S5–S10).

Checklist length is not a perfect proxy for sampling effort, because it confounds sampling effort and the number of species available to be recorded (Warton et al., 2013). The number of species available for detection is probably not constant through the year in many locations in Ireland because of seasonal patterns in detectability. Likewise, the total number of millipede species present in 1×1 km grid cells is almost certainly not constant across all grid squares in Ireland. Checklist length is therefore determined by sampling effort, species richness (which varies spatially) and detectability (which varies seasonally). In our data, checklist length was not correlated with any of the environmental predictor variables (absolute value of Spearman's correlation coefficient was always <0.1 for correlations between checklist length and other predictor variables), and there were no obvious geographic patterns in checklist length (Figure S16). Most locations in Ireland probably have multiple millipede species available for detection at most times of year (i.e. adults of multiple species present), but, as is common in biological records datasets, many of the checklists in our dataset had checklist lengths of one (37% of checklists) or two (23% of checklists). We suspect that checklists of only one or two species resulted from limited sampling effort rather than intensive surveys in locations with only one or two species. Our use of checklist length provided no information about cases in which surveys were conducted but no species were found. There may be times of the year and/or 1 km^2 grid squares in which there truly are no millipedes to be found (e.g. in grid squares dominated by bog), but our training data do not contain information about those areas. Because of this, we expect our models to over-estimate the probability of millipedes occurring in those areas.

5 | CONCLUSION

Species distribution models for rare millipede species had better prediction performance according to most metrics when the training data were under-sampled in a spatially stratified way by discarding non-detection data to improve class balance and spatial bias. In some cases, training data that were under-sampled in an unstratified way produced better predictive models than did spatially stratified under-sampled data. Models trained with spatially stratified under-sampled data had worse specificity (true negative rate) than models trained with raw data, but the decrease in specificity was small for the rarest species and was accompanied by large improvements in sensitivity (true positive rate).

A comparison of models using geographic coordinates as predictor variables and models using environmental predictor variables showed that neither set of variables always outperformed the other. Notably, the distribution of *O. sabulosus* was better predicted using geographic coordinates rather than environmental variables. Models combining both geographic coordinates and environmental variables as predictors were consistently among the best performing models.

We tested spatially stratified under-sampling using a smaller, sparser dataset than was used in previous tests (Robinson et al., 2018, 2020). Modelling distributions using sparse datasets is beneficial when it is difficult or expensive to collect additional observational data about species occurrence. Traditional survey methods (including citizen science) will likely never produce large occurrence datasets for taxa that are small, difficult to identify and/or non-charismatic (though other approaches including eDNA may be able to provide large amounts of data about such taxa). Our results suggested that under-sampling can improve SDMs of rare, non-charismatic, poorly sampled taxa, including invertebrates, for which biological recording effort is limited. Manipulating the spatial pattern of non-detection data during spatially stratified under-sampling sometimes improved, but sometimes reduced, model performance. Under-sampling is worth considering when modelling distributions of rare, poorly sampled taxa, but more guidance is needed about the spatial pattern of under-sampling. The effects of manipulating the spatial pattern of data during under-sampling needs careful testing using simulations and/or systematically collected test data in order to provide guidance.

ACKNOWLEDGEMENTS

The authors thank Tomás Murray and the Irish National Biodiversity Data Centre (NBDC) for advice about the biological records data, and Declan Doogue for information about millipedes in Ireland. The authors thank the British Myriapod and Isopod Group and the many citizen scientists who collected and contributed data. Open access funding provided by IReL.

FUNDING INFORMATION

This publication emanated from research supported in part by Science Foundation Ireland (grant number 15/IA/2881). The opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Science Foundation Ireland. This research used CORINE data made available with funding by the European Union.

CONFLICT OF INTEREST

All authors declare that they have no conflicts of interest.

DATA AVAILABILITY STATEMENT

All code used to perform analyses and produce plots is available at https://doi.org/10.5281/zenodo.6872414. Millipede occurrence data is available to download from GBIF at https://doi.org/10.15468/dl.833k97. CORINE land cover data is available from https://www.eea.europa.eu/ds_resolveuid/ecb838dabf4849838ba5f3dc81ca6b0e [8 Aug 2016]. E-OBS climate data is available from http://www.ecad.eu/download/ensembles/downloadchunks.php. ETOPO1 elevation data is available from https://www.ngdc.noaa.gov/mgg/global/relief/ETOPO1/data/ice_surface/grid_registered/netcdf/ [accessed 8 May 2019].

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/ddi.13619.

ORCID

Willson Gaul https://orcid.org/0000-0002-0821-9256

REFERENCES

Amano, T., & Sutherland, W. J. (2013). Four barriers to the global understanding of biodiversity conservation: Wealth, language, geographical location and security. *Proceedings of the Royal Society B: Biological Sciences*, *280*, 20122649. https://doi.org/10.1098/rspb.2012.2649

Amante, C., & Eakins, B. W. (2009). *ETOPO1 1 arc-minute global relief model: Procedures, data sources and analysis*. NOAA Technical Memorandum NESDIS NGDC-24. National Geophysical Data Center, NOAA. https://doi.org/10.7289/V5C8276M

Austin, M. (2007). Species distribution models and ecological theory: A critical assessment and some possible new approaches. *Ecological Modelling*, *200*, 1–19. https://doi.org/10.1016/j.ecolmodel.2006.07.005

Bahn, V., & McGill, B. J. (2007). Can niche-based distribution models outperform spatial interpolation? *Global Ecology and Biogeography*, *16*, 733–742. https://doi.org/10.1111/j.1466-8238.2007.00331.x

Baker, G. H. (1978). The distribution and dispersal of the introduced millipede, *Ommatoiulus moreletii* (Diplopoda: Iulidae), in Australia. *Journal of Zoology*, *185*, 1–11. https://doi.org/10.1111/j.1469-7998.1978.tb03309.x

Barbet-Massin, M., Jiguet, F., Albert, C. H., & Thuiller, W. (2012). Selecting pseudo-absences for species distribution models: How, where and how many? *Methods in Ecology and Evolution*, *3*, 327–338. https://doi.org/10.1111/j.2041-210X.2011.00172.x

Bardgett, R. D. (2005). *The biology of soil: A community and ecosystem approach*. Oxford University Press.

Bardgett, R. D., & Wardle, D. A. (2010). *Aboveground-belowground linkages: Biotic interactions, ecosystem processes, and global change*. Oxford University Press.

Beale, C. M., Lennon, J. J., & Gimona, A. (2008). Opening the climate envelope reveals no macroscale associations with climate in European birds. *Proceedings of the National Academy of Sciences of the United States of America*, *105*, 14908–14912. https://doi.org/10.1073/pnas.0803506105

Beale, C. M., Lennon, J. J., Yearsley, J. M., Brewer, M. J., & Elston, D. A. (2010). Regression analysis of spatial data. *Ecology Letters*, *13*, 246–264. https://doi.org/10.1111/j.1461-0248.2009.01422.x

Biological Records Centre, UK (2017). *Millipedes of Ireland* [dataset]. Waterford, Ireland: National Biodiversity Data Centre. https://maps.biodiversityireland.ie/Dataset/52

Bivand, R., Keitt, T., & Rowlingson, B. (2018). *Rgdal: Bindings for the 'geospatial' data abstraction library*. R package versions 1.3-9 and 1.4-4.

Boakes, E., Gliozzo, G., Seymour, V., Harvey, M., Smith, C., Roy, D. B., & Haklay, M. (2016). Patterns of contribution to citizen science biodiversity projects increase understanding of volunteers' recording behaviour. *Scientific Reports*, *6*, 33051. https://doi.org/10.1038/srep33051

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

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1–3. https://doi.org/10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2

Cardoso, P., Borges, P. A. V., Triantis, K. A., Ferrández, M. A., & Martín, J. L. (2011). Adapting the IUCN red list criteria for invertebrates. *Biological Conservation*, *144*(10), 2432–2440. https://doi.org/10.1016/j.biocon.2011.06.020

Cardoso, P., Borges, P. A. V., Triantis, K. A., Ferrández, M. A., & Martín, J. L. (2012). The underrepresentation and misrepresentation of invertebrates in the IUCN red list. *Biological Conservation*, *149*(1), 147–148. https://doi.org/10.1016/j.biocon.2012.02.011

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46.

CORINE (2012). *CORINE Land Cover database*. Version 18. © European Union, Copernicus Land Monitoring Service 2016, European Environment Agency (EEA). https://www.eea.europa.eu/ds_resolveuid/ecb838dabf4849838ba5f3dc81ca6b0e

Currie, D. J., Pétrin, C., & Boucher-Lalonde, V. (2019). How perilous are broad-scale correlations with environmental variables? *Frontiers of Biogeography*, *12*(2), e44842. https://doi.org/10.21425/f5fbg44842

David, J. F., & Handa, I. T. (2010). The ecology of saprophagous macroarthropods (millipedes, woodlice) in the context of global change. *Biological Reviews*, *85*(4), 881–895. https://doi.org/10.1111/j.1469-185X.2010.00138.x

Donaldson, M. R., Burnett, N. J., Braun, D. C., Suski, C. D., Hinch, S. G., Cooke, S. J., & Kerr, J. T. (2016). Taxonomic bias and international biodiversity conservation research. *FACETS*, *1*(1), 105–113. https://doi.org/10.1139/facets-2016-0011

Dormann, C. F. (2007). Effects of incorporating spatial autocorrelation into the analysis of species distribution data. *Global Ecology and Biogeography*, *16*(2), 129–138. https://doi.org/10.1111/j.1466-8238.2006.00279.x

El-Gabbas, A., & Dormann, C. F. (2018). Improved species-occurrence predictions in data-poor regions: Using large-scale data and bias correction with down-weighted Poisson regression and Maxent. *Ecography*, *41*, 1161–1172. https://doi.org/10.1111/ecog.03149

Fielding, A. H., & Bell, J. F. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, *24*(1), 38–49.

Fink, D., Damoulas, T., Bruns, N. E., La Sorte, F. A., Hochachka, W. M., Gomes, C. P., & Kelling, S. (2014). Crowdsourcing meets ecology: Hemispherewide spatiotemporal species distribution models. *AI Magazine*, *35*(2), 19–30.

Fink, D., Hochachka, W. M., Zuckerberg, B., Winkler, D. W., Shaby, B., Munson, M. A., Hooker, G., Riedewald, M., Sheldon, D., & Kelling, S. (2010). Spatiotemporal exploratory models for broad-scale survey data. *Ecological Applications*, *20*, 2131–2147. https://doi.org/10.1890/09-1340.1

Fithian, W., & Hastie, T. (2014). Local case-control sampling: Efficient subsampling in imbalanced data sets. *Annals of Statistics*, *42*(5), 1693–1724. https://doi.org/10.1214/14-AOS1220

Fourcade, Y., Besnard, A. G., & Secondi, J. (2018). Paintings predict the distribution of species, or the challenge of selecting environmental predictors and evaluation statistics. *Global Ecology and Biogeography*, *27*, 245–256. https://doi.org/10.1111/geb.12684

Fourcade, Y., Engler, J. O., Rödder, D., & Secondi, J. (2014). Mapping species distributions with MAXENT using a geographically biased sample of presence data: A performance assessment of methods for correcting sampling bias. *PLoS One*, *9*, e97122. https://doi.org/10.1371/journal.pone.0097122

Gaul, W., Sadykova, D., White, H. J., Leon-Sanchez, L., Caplat, P., Emmerson, M. C., & Yearsley, J. M. (2020). Data quantity is more important than its spatial bias for predictive species distribution modelling. *PeerJ*, *8*, e10411. https://doi.org/10.7717/peerj.10411

GBIF.org (2021). *GBIF occurrence download*. https://doi.org/10.15468/dl.833k97

Gräler, B., Pebesma, E., & Heuvelink, G. (2016). Spatio-temporal interpolation using gstat. *The R Journal*, *8*, 204–218. https://doi.org/10.32614/RJ-2016-014

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, *73*, 220–239. https://doi.org/10.1016/j.eswa.2016.12.035

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference and prediction* (2nd ed.). Springer.

Haylock, M. R., Hofstra, N., Klein Tank, A. M. G., Klok, E. J., Jones, P. D., & New, M. (2008). A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006. *Journal of Geophysical Research Atmospheres*, *113*, D20119. https://doi.org/10.1029/2008JD010201

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. https://doi.org/10.1007/978-3-030-04663-7_4

Heberling, M. J., Miller, J. T., Noesgaard, D., Weingart, S. B., & Schigel, D. (2021). Data integration enables global biodiversity synthesis. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(6), e2018093118. https://doi.org/10.1073/pnas.2018093118

Hijmans, R. J. (2018). *Raster: Geographic data analysis and modelling*. R package versions 2.8-4 and 2.9-23.

Isaac, N. J. B., van Strien, A. J., August, T. A., de Zeeuw, M. P., & Roy, D. B. (2014). Statistics for citizen science: Extracting signals of change from noisy ecological data. *Methods in Ecology and Evolution*, *5*, 1052–1060. https://doi.org/10.1111/2041-210X.12254

IUCN (2012). *IUCN red list categories and criteria*. Version 3.1 (2nd ed.). Gland, Switzerland and Cambridge, UK: IUCN. iv + 32pp.

James, J. F. (2011). *A student's guide to fourier transforms: With applications in physics and engineering*. Cambridge University Press.

Johnston, A., Moran, N., Musgrove, A., Fink, D., & Baillie, S. R. (2020). Estimating species distributions from spatially biased citizen science data. *Ecological Modelling*, *422*, 108927. https://doi.org/10.1016/j.ecolmodel.2019.108927

Kelling, S., Hochachka, W. M., Fink, D., Riedewald, M., Caruana, R., Ballard, G., & Hooker, G. (2009). Data-intensive science: A New paradigm for biodiversity studies. *Bioscience*, *59*(7), 613–620. https://doi.org/10.1525/bio.2009.59.7.12

Kime, R. D. (1999). The continental distribution of British and Irish millipedes. *Bulletin of the British Myriapod Group*, *15*, 33–76.

Kime, R. D. (2001). The continental distribution of British and Irish millipedes, part 2. *Bulletin of the British Myriapod and Isopod Group*, *17*, 7–42.

Kime, R. D. (2004). The Belgian millipede fauna (Diplopoda). *Bulletin de l'Institution Royal Des Sciences Naturelles de Belgique Entomologie*, *74*, 35–68.

Lee, P. (2006). *Atlas of the millipedes (Diplopoda) of Britain and Ireland*. Pensoft Publishers.

Lennon, J. J. (2000). Red-shifts and red herrings in geographical ecology. *Ecography*, *23*, 101–113. https://doi.org/10.1111/j.1600-0587.2000.tb00265.x

Liaw, A., & Wiener, M. (2002). Classification and regression by random-Forest. *R News*, *2*, 18–22.

MacKenzie, D. I., Nichols, J. D., Lachman, G. B., Droege, S., Royle, A. A., & Langtimm, C. A. (2002). Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, *83*, 2248–2255. https://doi.org/10.1890/0012-9658(2002)083[2248:ESORWD]2.0.CO;2

Maes, D., Isaac, N. J. B., Harrower, C. A., Collen, B., van Strien, A. J., & Roy, D. B. (2015). The use of opportunistic data for IUCN red list assessments. *Biological Journal of the Linnean Society*, *115*(3), 690–706. https://doi.org/10.1111/bij.12530

Mammola, S., Riccardi, N., Prié, V., Correia, R., Cardoso, P., Lopes-Lima, M., & Sousa, R. (2020). Towards a taxonomically unbiased European Union biodiversity strategy for 2030. *Proceedings of the Royal Society B: Biological Sciences*, *287*(1940), 20202166. https://doi.org/10.1098/rspb.2020.2166

Oliveira, U., Paglia, A. P., Brescovit, A. D., de Carvalho, C. J. B., Silva, D. P., Rezende, D. T., Leite, F. S. F., Batista, J. A. N., Barbosa, J. P. P. P., Stehmann, J. R., Ascher, J. S., Vasconcelos, M. F., De Marco, P., Löwenberg-Neto, P., Dias, P. G., Ferro, V. G., & Santos, A. J. (2016). The strong influence of collection bias on biodiversity knowledge shortfalls of Brazilian terrestrial biodiversity. *Diversity and Distributions*, *22*, 1232–1244. https://doi.org/10.1111/ddi.12489

Pebesma, E. (2018). Simple features for R: Standardized support for spatial vector data. *The R Journal*, *10*, 439–446.

Phillips, S. J., Dudik, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., & Ferrier, S. (2009). Sample selection bias and presence-only distribution models: Implications for background and pseudo-absence data. *Ecological Applications*, *19*, 181–197. https://doi.org/10.1890/07-2153.1

Potts, S. G., Imperatriz-Fonseca, V., Ngo, H. T., Aizen, M. A., Biesmeijer, J. C., Breeze, T. D., Dicks, L. V., Garibaldi, L. A., Hill, R., Settele, J., & Vanbergen, A. J. (2016). Safeguarding pollinators and their values to human well-being. *Nature*, *540*(7632), 220–229. https://doi.org/10.1038/nature20588

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.r-project.org/

Risch, A. C., Ochoa-Hueso, R., van der Putten, W. H., Bump, J. K., Busse, M. D., Frey, B., Gwiazdowicz, D. J., Page-Dumroese, D. S., Vandegehuchte, M. L., Zimmermann, S., & Schütz, M. (2018). Size-dependent loss of aboveground animals differentially affects grassland ecosystem coupling and functions. *Nature Communications*, *9*(1), 3684. https://doi.org/10.1038/s41467-018-06105-4

Robinson, O. J., Ruiz-Gutierrez, V., & Fink, D. (2018). Correcting for bias in distribution modelling for rare species using citizen science data. *Diversity and Distributions*, *24*, 460–472. https://doi.org/10.1111/ddi.12698

Robinson, O. J., Ruiz-Gutierrez, V., Reynolds, M. D., Golet, G. H., Strimas-Mackey, M., & Fink, D. (2020). Integrating citizen science data with expert surveys increases accuracy and spatial extent of species distribution models. *Diversity and Distributions*, *26*, 1–11. https://doi.org/10.1111/ddi.13068

Ross, N. (2018). *fasterize: Fast polygon to raster conversion*. R package version 1.0.0.

Szabo, J. K., Vesk, P. A., Baxter, P. W. J., & Possingham, H. P. (2010). Regional avian species declines estimated from volunteer-collected long-term data using list length analysis. *Ecological Applications*, *20*, 2157–2169. https://doi.org/10.1890/09-0877.1

Thibaud, E., Petitpierre, B., Broennimann, O., Davison, A. C., & Guisan, A. (2014). Measuring the relative effect of factors affecting species distribution model predictions. *Methods in Ecology and Evolution*, *5*, 947–955. https://doi.org/10.1111/2041-210x.12203

Valavi, R., Elith, J., Lahoz-Monfort, J. J., & Guillera-Arroita, G. (2019). Block CV: An r package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. *Methods in Ecology and Evolution*, *10*, 225–232. https://doi.org/10.1111/2041-210X.13107

van den Besselaar, E. J. M., Haylock, M. R., van der Schrier, G., & Klein Tank, A. M. G. (2011). A European daily high-resolution observational gridded data set of sea level pressure. *Journal of Geophysical Research*, *116*, D11110. https://doi.org/10.1029/2010JD015468

van Strien, A. J., Termaat, T., Groenendijk, D., Mensing, V., & Kéry, M. (2010). Site-occupancy models may offer new opportunities for dragonfly monitoring based on daily species lists. *Basic and Applied Ecology*, *11*, 495–503. https://doi.org/10.1016/j.baae.2010.05.003

Warton, D. I., Renner, I. W., & Ramp, D. (2013). Model-based control of observer bias for the analysis of presence-only data in ecology. *PLoS One*, *8*, e79168. https://doi.org/10.1371/journal.pone.0079168

Wickham, H. (2017). *tidyverse: Easily install and load the 'tidyverse'*. R package version 1.2.1.

BIOSKETCH

The coauthor team is a group of ecologists and biologists with diverse interests in biogeography, macroecology and species interactions. Willson Gaul is currently an ecologist and teacher on Saipan in the Northern Mariana Islands.

Author contributions: W.G. conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft. D.S. contributed code, advised on statistical methods, authored or reviewed drafts of the paper, and approved the final draft. H.J.W. contributed code, advised on statistical methods, authored or reviewed drafts of the paper, and approved the final draft. L.L.S., P.C. and M.C.E. authored or reviewed drafts of the paper, and approved the final draft. J.M.Y. conceived and designed the experiments, contributed code, advised on statistical methods, authored or reviewed drafts of the paper, and approved the final draft.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Gaul, W., Sadykova, D., White, H. J., León-Sánchez, L., Caplat, P., Emmerson, M. C., & Yearsley, J. M. (2022). Modelling the distribution of rare invertebrates by correcting class imbalance and spatial bias. *Diversity and Distributions*, *28*, 2171–2186. https://doi.org/10.1111/ddi.13619