DOI: 10.1111/ddi.12698

BIODIVERSITY RESEARCH

WILEY Diversity and Distributions

Correcting for bias in distribution modelling for rare species using citizen science data

Orin J. Robinson | Viviana Ruiz-Gutierrez | Daniel Fink

Cornell Laboratory of Ornithology, Ithaca, NY, USA

Correspondence
Orin J. Robinson, Cornell Laboratory of Ornithology, Ithaca, NY, USA.
Email: ojr7@cornell.edu

Funding information
Leon Levy Foundation; Wolf Creek Foundation; National Science Foundation, Grant/Award Number: DBI-1356308, CNS-1059284 and CCF-1522054

Editor: Risto Heikkinen

Abstract

Aim: To improve the accuracy of inferences on habitat associations and distribution patterns of rare species by combining machine-learning, spatial filtering and resampling to address class imbalance and spatial bias of large volumes of citizen science data.

Innovation: Modelling rare species' distributions is a pressing challenge for conservation and applied research. Often, a large number of surveys are required before enough detections occur to model distributions of rare species accurately, resulting in a data set with a high proportion of non-detections (i.e. class imbalance). Citizen science data can provide a cost-effective source of surveys but likely suffer from class imbalance. Citizen science data also suffer from spatial bias, likely from preferential sampling. To correct for class imbalance and spatial bias, we used spatial filtering to under-sample the majority class (non-detection) while maintaining all of the limited information from the minority class (detection). We investigated the use of spatial under-sampling with randomForest models and compared it to common approaches used for imbalanced data, the synthetic minority oversampling technique (SMOTE), weighted random forest and balanced random forest models. Model accuracy was assessed using kappa, Brier score and AUC. We demonstrate the method by evaluating habitat associations and seasonal distribution patterns using citizen science data for a rare species, the tricoloured blackbird (*Agelaius tricolor*).

Main Conclusions: Spatial under-sampling increased the accuracy of each model and outperformed the approach typically used to direct under-sampling in the SMOTE algorithm. Our approach is the first to characterize winter distribution and movement of tricoloured blackbirds. Our results show that tricoloured blackbirds are positively associated with grassland, pasture and wetland habitats, and negatively associated with high elevations or evergreen forests during both winter and breeding seasons. The seasonal differences in distribution indicate that individuals move to the coast during the winter, as suggested by historical accounts.

KEYWORDS
citizen science, class imbalance, random forest, spatial bias, species distribution model, tricoloured blackbird

1 | INTRODUCTION

Understanding the factors driving the geographical distribution of organisms is one of the fundamental motivators of ecological research. Species distribution models (SDM) are commonly used to describe and predict species distributions by relating a suite of environmental variables to geographical locations where searches have been conducted (Guisan & Zimmerman, 2000). Effectively modelling a species' distribution often requires a high-quality data set with enough presence–absence information at sites throughout the range of the species (Brotons, Thuiller, Arau′ jo, & Hirzel, 2004).

When the objective is to make inferences on rare, range-restricted or hard to detect species, the data requirements of distribution models can be especially challenging. Rare species often do not occupy all suitable habitat in a region, may be patchily distributed (e.g. locally abundant but regionally rare) and can be difficult to detect (McCune, 2016; Pacifici, Reich, Dorazio, & Conroy, 2015). This results in small numbers of positive detections, making it difficult to accurately predict distributions and making it easier to overfit them when used in statistical models, even with few number of explanatory variables (Vaughn & Ormerod, 2005). These challenges are exacerbated when the objective is to make inferences on the factors driving the distribution patterns of a species with habitat associations that differ throughout the annual life cycle of the species (e.g. migratory birds; Moore, 2000). While there have been advances in rare species sampling techniques (e.g. Conroy, Runge, Barker, Schofield, & Fonnesbeck, 2008; Guisan et al. 2006; Pacifici, Dorazio, & Conroy, 2012), there have been few developments focusing on how to utilize rare species data collected by large-scale citizen science projects. These projects have the potential to collect large numbers of positive detections; however, they also tend to collect even larger numbers of non-detections resulting in highly imbalanced presence–absence data sets.

Class imbalance occurs where the sample from one class (e.g. absences) is much larger than the sample from the other class (e.g. presences). Most often for rare event data, inferences are dependent on the information obtained from the minority class, such as the presence of a rare species, an incidence of an emergent disease or fraud detection. However, the volume of information for majority class, such as the number of healthy individuals, may overwhelm the model (Longadge, Dongre, & Malik, 2013). For example, if a rare disease is prevalent in only 0.5% of patients sampled, a model could simply choose the default strategy of always guessing that a patient is healthy (e.g. the negative class) and be correct 99.5% of the time. The degree of accuracy of the model is high, but the model will fail to make accurate predictions and will have little ability to identify the factors that drive the likelihood of contracting the disease or the class of interest. Further, Fithian and Hastie (2014) showed that logistic regression (often used in SDM) becomes less accurate as the classes move away from balance.

One approach to address bias related to class imbalance is to base inferences only on positive detections of individuals, otherwise known as presence-only models in the SDM literature. These models have become increasingly common (Hirzel, Hausser, Chessel, & Perrin, 2002; Phillips, Anderson, & Schapire, 2006), and many studies (36% in a review by Yackulic et al. 2013) have resorted to discarding absence data and using a presence-only method. However, throwing out absence data removes information useful for modelling, often increasing the risk of sampling bias (Fithian, Elith, Hastie, & Keith, 2015), and has been shown to be detrimental to the accuracy of the SDM, particularly when the majority of the data *is* absence data (Brotons et al., 2004).

Class imbalance issues are not unique to SDMs (He & Garcia 2009) and have been identified and long studied in fields as varied as oil spill detection from satellite images (Kubat, Holte, & Matwin, 1998), text classification (Lewis & Ringuette, 1994) and rare disease diagnosis (Woods et al., 1993). There are effective methods for handling class-imbalanced data. Kubat and Matwin (1997) proposed resampling observations from the minority class with replacement (oversampling), or randomly removing observations from the majority class (under-sampling or one-sided sampling), until the classes are balanced. Ling and Li (1998) and Japkowicz (2000) showed that under-sampling generally performed better than oversampling. Case–control sampling draws samples uniformly from each class and adjusts the sample to correct for imbalance. This method has been further developed to include local case–control sampling, which preferentially samples observations that are conditionally rare (Fithian & Hastie, 2014).

Chawla, Bowyer, Hall, and Kegelmayer (2002) proposed a method that creates synthetic examples from the minority class that are not exact copies, as in traditional oversampling, but that occupy the parameter space between a randomly drawn observation and a nearest neighbour, called the synthetic minority oversampling technique (SMOTE). The application of SMOTE has been shown to create data sets that produced more accurate classification models, when used with machine-learning classification methods, relative to traditional oversampling or under-sampling alone (Chawla et al., 2002). In addition, they found that SMOTE was most effective when used in conjunction with under-sampling.

Machine-learning methods, such as random forest (RF), are increasingly being applied to address ecological classification problems (Cutler, Edwards, Beard, Cutler, & Hess, 2007; Mi, Huttmann, Guo, Han, & Wen, 2017). RF is an ensemble method where a large number of individual decision tree models are induced by taking bootstrap samples of the data. Chen, Liaw, and Breiman (2004) proposed two simple modifications to RF that make it suitable to make inferences using highly imbalanced data sets One of the proposed modifications is a balanced RF (BAL), where the model first draws bootstrap samples from the minority class and then draws an equal number of samples from the majority class, thereby balancing the classes. Another modification is a weighted RF (WRF), and here, the model assigns a harsher penalty to the misclassification of the minority class than the majority class (often proportional to the class prevalences). Chen et al. (2004) were able to show that BAL and WRF were superior to one-sided sampling and SMOTE when applied to multiple data sets, but neither approach (BAL and WRF) outperformed the other.

The development of citizen science monitoring programs to survey species has proven to be a powerful tool towards addressing

data limitations of distribution models (Fink et al., 2014; Theobald et al., 2015). Citizen science monitoring programs are often able to accommodate or provide cost-effective surveys across numerous sites and times of the year (Tulloch et al. 2013). Although some citizen science monitoring programs have highly structured protocols (e.g. Weir & Mossman, 2005), most projects have very flexible and open levels of participation and are not limited to a specific sampling method or season. When guidelines for data collection are kept minimal, citizen science programs can engage large numbers of participants and collect large volumes of data that are useful for characterizing species distributions from the local to the continental scale (e.g. Fink et al., 2010). Although these characteristics of citizen science monitoring programs often generate relatively large numbers of species detections, class imbalance and other biases associated with citizen science data remain as challenges when modelling distributions for rare species with presence–absence data (Sullivan et al., 2014).

A major obstacle with citizen science programs is the irregular and often sparse spatial pattern of observations. Citizen scientists who participate in surveys are more likely to sample close to home (Luck, Ricketts, Daily, & Imhoff, 2004), convenient locations (e.g. roadsides; Kadmon, Farber, & Danin, 2004), or in accessible areas where biodiversity is known to be high (Prendergast, Wood, Lawton, & Eversham, 1993). This preferential sampling can translate into bias in spatial and environmental predictors associated with occurrence data (Geldmann et al., 2016). Citizen science data may also suffer from temporal bias, providing more information from specific times of the year when observers are more active, when certain species are present or when they are available to conduct surveys (e.g. weekend bias; Courter, Johnson, Stuyck, Lang, & Kaiser, 2012). This temporal bias may also add to spatial bias if observers are visiting specific locations for certain species during a given period of the annual cycle, such as repeatedly going to known breeding sites during the breeding season.

When not properly accounted for, preferential sampling may also lead to an increase in spatial autocorrelation, subsequent overfitting of the model and inaccurate model evaluation statistics (Boria, Olson, Goodman, & Anderson, 2014; Hijmans, 2012). Due to the lack of survey design in most citizen science programs, model-based approaches are the most effective method to account for bias inherent to these data, and modelling the sampling process itself can greatly improve accuracy of SDMs (e.g. Conn, Thorson, & Johnson, 2016). To date, most SDM approaches have been limited in their ability to improve our current capacity to accurately make inferences on factors driving the distribution of rare species. For example, Boria et al. (2014) proposed a spatial filtering technique to improve the accuracy of SDMs by only including species detections that were at least a given distance apart when using a presence-background method. Given that most rare species have few detections and that some may have clumped distributions where many detections are likely to occur close together in space, this spatial filtering technique may create a situation where too much information is lost when presence data are removed. Even though citizen science data also tend to increase information, the number of detections for a rare species will likely remain low even as the number of surveys increases. Therefore, there is a clear need to address both spatial bias and class imbalance in SDMs in order to effectively use citizen science survey data to model rare events.

In this study, we propose an approach to improve the accuracy of inferences on the habitat associations and distribution patterns of rare, range-restricted or hard to detect species by combining machine-learning, spatial filtering and resampling to address class imbalance and spatial bias of large volumes of citizen science data. We accomplish this using a RF approach that allows spatial filtering to direct the under-sampling of the majority class and compare our approach with other current methods used to deal with spatial bias and class imbalance. We illustrate the utility of our approach by modelling the winter and spring distributions of the tricoloured blackbird (*Agelaius tricolor*) using citizen science data from eBird (www.ebird.org; Sullivan et al., 2014). Tricoloured blackbirds are rare within their range but may be locally abundant at many sites. Their occupied region, rarity and patchy distribution make inferences on this species prone to spatial bias and class imbalance for presence–absence data, and thus, an ideal case study for our method.

2 | MATERIALS AND METHODS

2.1 | Study species

Tricoloured blackbirds are almost entirely restricted to California, with a small number (~1% of the total population) breeding in Oregon, Washington and Nevada in the United States, and Baja California in Mexico (Meese, Beedy, & Hamilton, 2014). Historically, tricoloured blackbirds used freshwater wetlands as breeding sites; however, more than 90% of the suitable wetlands were lost from 1780 to 1980 (Dahl, 1990). This has caused tricoloured blackbirds to seek alternative breeding sites, and many colonies are now found breeding in agricultural fields and invasive plants such as Himalayan blackberry (*Rubus armeniacus*) and milk thistle (*Silybum marianum*; Meese, 2009). While the reproductive success of colonies that breed in alternative habitats have been shown to be similar to, or greater than, those in the traditional wetland habitats, the population continues to decline (Holyoak, Meese, & Graves, 2014; Weintraub, George, & Dinsmore, 2016). Breeding tricoloured blackbirds may have multiple breeding attempts in a single season, potentially at a different site within the same breeding season, and are also not likely to use the same breeding sites each year but seem to prefer the same habitat types (DeHaven, Crase, & Woronecki, 1975).

Recent work has produced considerable knowledge of the distribution of tricoloured blackbirds during the breeding season, but the habitat requirements and distribution during the non-breeding period (winter) are poorly understood (Meese et al., 2014). There is speculation that tricoloured blackbirds withdraw from the northern parts of their range and move towards coastal California and the San Joaquin and Sacramento River deltas during winter (DeHaven et al., 1975). However, observational data for this portion of their annual cycle are

sparse, and their movements throughout the year are described as nomadic.

2.2 | Data

We collected data from eBird on checklists submitted in November-January and March-July of 2008–2014. Data were restricted to these time periods to capture the habitat associations of winter feeding and roosting patterns, and the time when the species is expected to be nesting and foraging to feed young (Meese, personal communication). Checklists were restricted to California as almost all of the known observations of tricoloured blackbirds are located within that state. More details about the filtering process for eBird data can be found in Appendix S1. We linked each checklist location with remotely sensed, spatial covariates from the Cropland Data Layer (CDL; Boryan, Yang, Mueller, & Craig, 2011; Han, Yang, Di, & Mueller, 2012). The National Agricultural Statistics Service of the US Department of Agriculture has produced the CDL each year since 1997, although for some states, the data begin in later years. It contains crop and land cover specific information that is georeferenced, verified via ground truth, at 30 or 56 m resolution and generally 85% to 95% accurate for most crop classes (Boryan et al., 2011; Han et al., 2012). We created 1.5 by 1.5 km pixels centred on each checklist location and used the CDL for the year in which the checklist was submitted to determine the per cent land cover of 105 crop or land cover classes and the elevation therein. We also included the effort variables of time, distance travelled, area searched and number of observers in each model to account for variation in detectability of tricoloured blackbirds. After constraining the data, we had 411,535 total checklists, 108,880 for the winter and 302,655 for the breeding season.

2.3 | Sampling methods and distribution models

Using the R package randomForest (Liaw & Wiener, 2002), we used random forest (RF) to separately model the breeding and winter occurrence of tricoloured blackbird's relationship to land cover and effort variables. As only ~1.5% of our eBird checklists recorded a "presence" for the species, we were concerned that class imbalance would potentially bias our results. To combat the class imbalance, we also tested BAL, WRF and SMOTE on our spring checklist data. Because our data contained such a low number of presence observations, we allowed the bootstrap sample drawn by BAL to equal the number of presence observations in the data set. For example, if the number of presence observations in the training data set for a BAL model was 1000, we allowed all 1000 detection observations to be used and drew 1000 random observations from those checklists that contained non-detection observations to run the model. For WRF, we weighted correctly classifying detection versus correctly classifying a non-detection as 100:1 (roughly the reciprocal of the class ratio). When using the SMOTE algorithm to resample the data, we doubled the number of presence observations by adding synthetic examples to the observed ones and randomly under-sampled the absences so that the classes were equally balanced (Chawla et al., 2002).

In addition to class imbalance, there was a noticeable spatial bias in our checklist locations that could potentially cause inaccuracies in our model results (Figure 1a). While the SMOTE algorithm thinned the data considerably while increasing the number of detections, it did not remedy the issue of spatial bias (Figure 1b). To improve the spatial balance of our data set, we created a bounding box (approximately 1053 km by 914.5 km) where each corner was one of the four combinations of maximum and minimum latitude and longitude of all

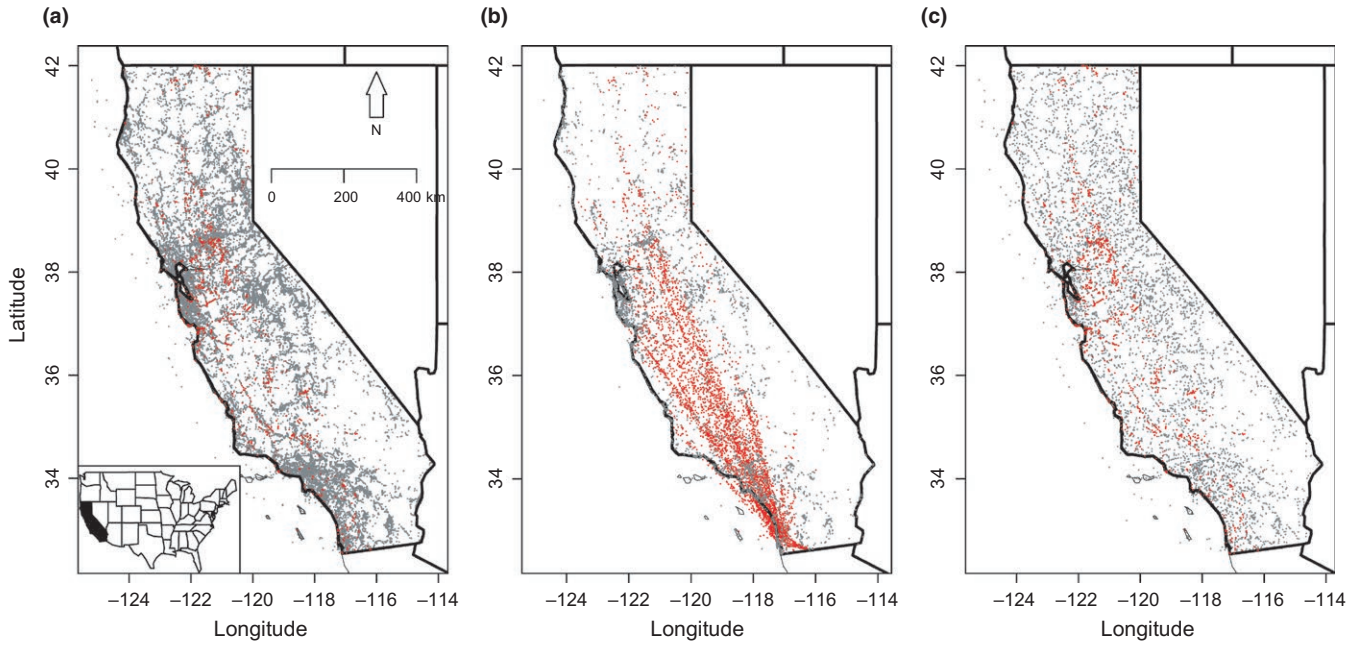

FIGURE 1 Breeding season checklists from eBird to be used in data analysis after no sampling method (a), SMOTE (b) and spatial undersampling (c). Grey points represent checklists that did not have a tricoloured blackbird detection, and red points represent those that did have a tricoloured blackbird detection

checklists included in our data set. We then created a 100 by 100 cell grid within the bounding box. To keep all of the information on presence observations, we split the data into two data sets: those checklists with a detection recorded for tricoloured blackbirds and those without a detection. We sampled one non-detection checklist at random from all the non-detection checklists within a given cell. This resulted in a more spatially balanced set of non-detection checklists. If a grid square had no non-detection checklists, no checklist was sampled from that cell. We then recombined the detection data with the grid-sampled non-detection data to create a spatially under-sampled detection/non-detection data set for analysis (Figure 2).

We examined accuracy measures for distribution models fit using RF 1) without a sampling technique applied to the data, 2) with the proposed techniques to deal with spatial bias and class imbalance (BAL, WRF and RF with SMOTE combined with random under-sampling; Chawla et al., 2002). In addition, we compared all four approaches described above under a scenario where the data were sampled via our spatial under-sampling method. For the analysis where we combined the SMOTE algorithm with our proposed technique, we used our spatial under-sampling method instead of random under-sampling. We split the data from eBird (for the first four methods) and the spatially under-sampled data (for the last four methods) into 50 random training and testing sets where half of the data was training and half testing. We ran each of the eight models on each training set and evaluated the accuracy of the models on each testing set. We used a spatially under-sampled testing set as the inferential target is detections across the state of California, where each location is equally important. Therefore, a spatially balanced testing set of locations is required. For each RF analysis (including all variations used), the number of classification trees in the ensemble was set to 1000 and the number of variables from which each model could select at each split for each tree was 11 ($\approx \sqrt{114}$), following the recommendation of using the square root of the number of variables included in the model (James, Witten, Hastie, & Tibshirani, 2013).

2.4 | Accuracy measures

We collected multiple model evaluation metrics at each run to characterize and compare the performance of the sampling methods. As RF is a Bootstrap AGgregation (BAGging) ensemble technique, it uses 60–70% of the data for each classification tree in the ensemble. For each model in the ensemble, the 30–40% of the data left out of the analysis, or out of bag, and used a test set to validate that model. The final out of bag estimates used for validation are averaged across the ensemble (James et al., 2013). The resulting metric is referred to as the out of bag error (OOB). This metric is highly sensitive to class

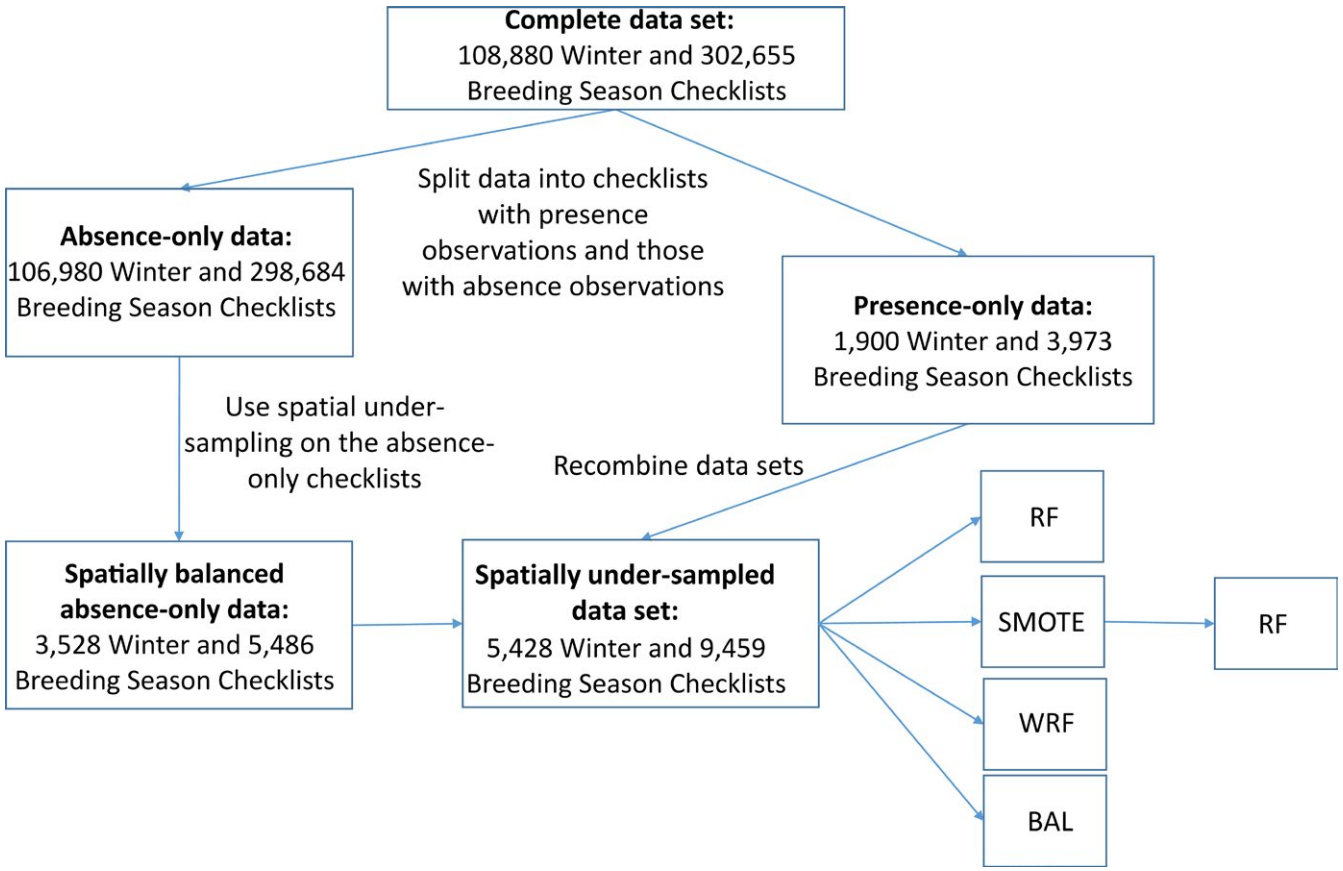

FIGURE 2 Schematic diagram of how we filtered the data for each season via spatial under-sampling before using it with one of the following methods: random forest (RF), synthetic minority oversampling (SMOTE) with RF, weighted random forest (WRF) and balanced random forest (BAL)

imbalance as it favours total accuracy over correctly predicting minority classes.

For this reason, we evaluated predictive performance on the independent test set. We used three statistics to characterize specific aspects of predictive performance. We used the area under the curve (AUC) to measure the model's ability to discriminate between positive and negative observations (Fielding & Bell 1997). The AUC is equal to the probability that the model will rank a randomly chosen positive observation higher than a randomly chosen negative one. Thus, AUC depends only on the ranking of the predictions.

To measure the models ability to accurately predict the binary detection/non-detection responses, we computed sensitivity (true positive rate; TPR), specificity (true negative rate; TNR) and Cohen's Kappa statistic (Cohen, 1960). Kappa (k) was designed to measure predictive performance taking into account the background rate of occurrence as

$$k = \frac{p_o - p_e}{1 - p_e}$$

where p_o is the observed agreement between the predictions made by the model and the test data

$$p_o = \text{TPR} + \text{TNR}$$

and p_e is a measure of how often the model and test data would agree by chance

$$p_e = m_1 \times d_1 + m_0 \times d_0$$

where m is the proportion of times the model predicted a presence (1) or absence (0), and d is the proportion of presences or absences in the test data set. Calculating kappa, TPR and TNR, required a threshold to convert the probabilities predicted by RF into the binary detection/non-detection responses. The threshold value chosen for each of the 50 random data sets was the value that maximized kappa for the data set used in an individual run.

Finally, to measure the quality for the predicted probabilities, we computed the Brier score (BS; Brier, 1950) as the mean squared error between probabilistic predictions from our model and the binary presence/absence data in the test set. Formally,

$$\text{BS} = \frac{1}{|M|} \sum_{i=1}^{|M|} \left(p_i - o_i\right)^2$$

where M is a set of i validation pairs $\left\{\left(p_1, o_1\right), \ldots, \left(p_i, o_i\right)\right\}$ and p_i is the probabilistic prediction made by the model for the i-th observation with observed presence/absence o_i. Brier scores are affected by discrepancies between predicted probabilities and empirical probabilities based on the observed data without requiring the choice of a threshold.

We then evaluated the models based on the predictive performance metrics above and chose the best method from which to create SDMs for tricoloured blackbirds in the winter and spring. Each SDM recorded the predictor importance statistics from each model for the time period over which they were evaluated. We evaluated the marginal effect of each variable on the probability of occurrence to determine the habitat associations for tricoloured blackbirds in each season. This approach has been shown to accurately estimate complex data generating processes with simulated and ecological data (Sethi, Dalton, & Hilborn, 2012). We then used the output of each SDM to create winter and breeding season distribution maps for tricoloured blackbirds in California.

3 | RESULTS

Tricoloured blackbirds were present in only 1.7% (1,900 detections on 108,880 checklists) of the winter and 1.3% (3,973 detections on 302,655 checklists) of the breeding season checklists. Our spatial under-sampling method greatly reduced the class imbalance and gave better spatial balance and representation to the data than no sampling technique or SMOTE (Figure 1). After running the models on the 50 randomly drawn data sets, the models using data that had been spatially under-sampled proved to be more accurate when compared to their non-spatially under-sampled counterpart with the exception of the OOB measure. All models had similar sensitivity and specificity. No one model was measurably better than another when using the same data sampling method (Figures 3 and 4). RF had the lowest OOB (mean = 0.001, *SD* = 0.0002), and WRF was the second lowest OOB (0.013, 0.004). These results are not surprising; RF and WRF used all of the checklists without any sampling technique applied. This kept the class imbalance intact in these data sets and OOB is highly influenced by the class imbalance.

We chose the RF model with spatial under-sampling for our SDMs for tricoloured blackbirds. Once spatially under-sampled and split in half to create training and validation data, our breeding season training data that went into the model contained 2005 detections on 4780 (42%) checklists and the winter data had 958 detections on 2760 checklists (35%). For making inferences on the occurrence patterns of tricoloured blackbird during the breeding season, the variables grass/pasture and evergreen forest had the highest variable importance score (VI; Appendix S2). The most important and positively associated variables were grass/pasture (#1 highest VI; Figure 4), developed/open (#4 VI) and herbaceous wetland (#5 VI). The most important and negatively associated variables were evergreen forest (#2 VI), shrubland (#3 VI) and elevation (#6 VI). The results are somewhat similar for winter occurrence; two variables that were clearly the most important were grass/pasture and herbaceous wetland (Appendix S2). The most important and positively associated variables were grass/pasture (highest overall ranked VI; Figure 5), herbaceous wetland (#2 VI), open water (#5 VI) and woody wetland (#7 VI). The most important and negatively associated variables were evergreen forest (#3 VI), shrubland (#4 VI) and elevation (#6 VI).

The distribution maps (Figure 6) predicted a higher probability of occurrence in the Central Valley and more "hotspots" overall, particularly in the interior of California for the breeding season. For the winter

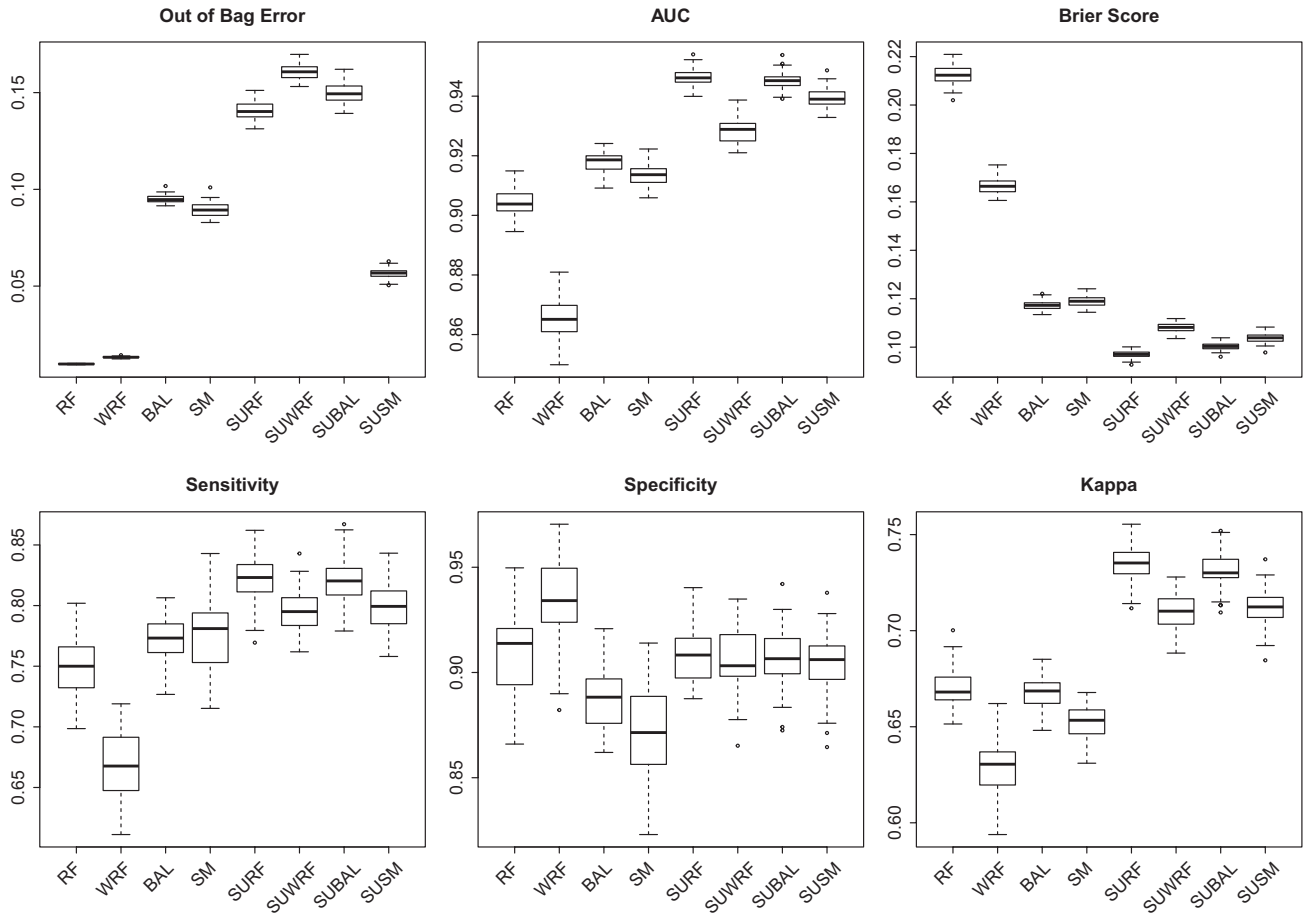

FIGURE 3 Accuracy metrics (OOB, AUC, Brier Score, Sensitivity, Specificity, and Kappa) for the different SDM models run over 50 randomly drawn tricoloured blackbird data sets. The models used were random forest (RF), weighted random forest (WRF), balanced random forest (BAL) and RF with synthetic minority oversampling technique (SM). Models with the prefix "SU" used our spatial under-sampling technique

distribution, the locations of highest probability of occurrence were along the coast, in the San Joaquin and Sacramento River deltas, and central Merced County; however, the overall probability of occurrence was lower in winter, even in the areas predicted to have the highest probability of occurrence during that season.

4 | DISCUSSION

Species distribution models play an important role in both ecological research, and more applied fields that inform conservation planning. However, data for rare or hard to sample species are costly to collect and are often lacking. Citizen science monitoring programs can be used to augment data sets for data-poor species, but it is not often used for reasons concerning biases in the collected data (Theobald et al., 2015). To address these challenges, we combined random forest with the data sampling techniques of under-sampling and spatial filtering. Our results show that spatial under-sampling can improve the accuracy measures of each of the SDM approaches that we tested, even when using a spatially biased and class-imbalanced citizen science data set. Model accuracy as measured by OOB was lower for the two models (RF and WRF) that made use of the entire imbalanced data set relative to the spatially under-sampled RF and WRF. This is attributable to the fact that the two latter models could have predicted an "absence" for each observation in the test data set for each run and been correct for more than 98% of the test observations. Thus, OOB is a poor measure of accuracy when using class-imbalanced data. While there was a small gain in sensitivity with spatial under-sampling, there was no loss of specificity. This is also reflected in our measure of kappa, as both specificity and sensitivity are used in its calculation. The loss of specificity was a concern as any under-sampling technique removes information used by the model, in our case, absences. True absence data in SDMs lead to higher accuracy in general and may be vital for ecologically meaningful evaluations of SDMs (Brotons et al., 2004; Václavík & Meentemeyer, 2009). Therefore, it was important not to lose specificity for the objective of gaining sensitivity.

The AUC for models using spatial under-sampling was higher than their non-spatially under-sampled counterparts, but the AUC for all models was relatively high (>0.85). We also computed Brier score to evaluate the accuracy of each model. Brier score allows the probabilistic predictions to be directly compared to the binary presence/absence data in the test set. The Brier score for each of the spatially under-sampled models consistently showed an improvement over the models not using that method.

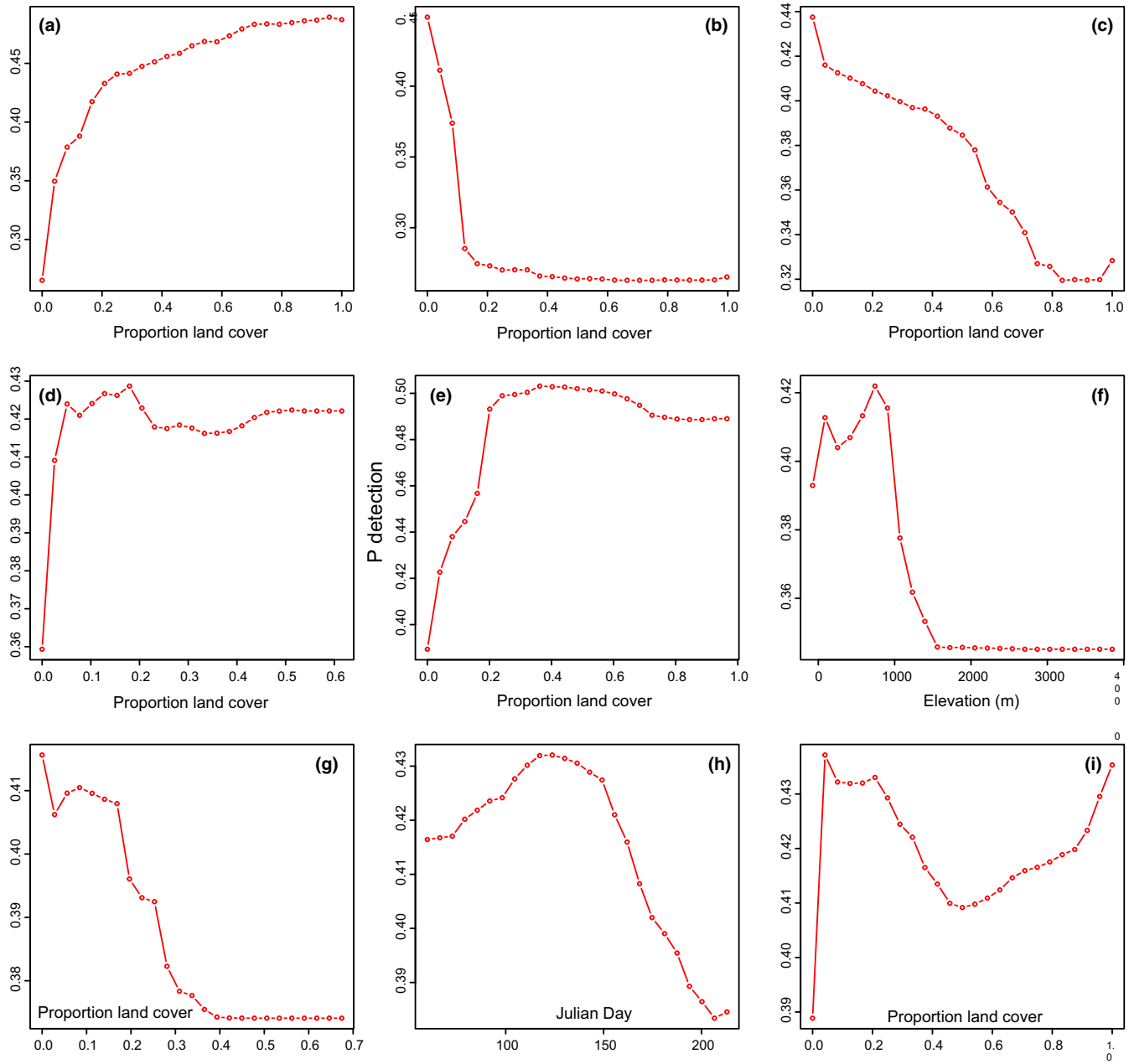

FIGURE 4 The partial dependence of detection probability (P detection) on the nine most important variables chosen in our breeding season habitat random forest analysis. (a) Grass/Pasture, (b) Evergreen Forest, (c) Shrubland, (d) Developed/Open Space, (e) Herbaceous Wetlands, (f) Elevation, (g) Developed/ Low Intensity, (h) Day, (i) Open Water

We used RF in concert with the spatial under-sampling technique to build winter and breeding season distribution maps for tricoloured blackbirds and to evaluate their habitat associations in each season. We chose this method because there was no discernable difference among the different RF models using spatial under-sampling. BAL would have further and randomly under-sampled the minority class, removing absence data that had been spatially stratified. The WRF model required considerably more computing time than the other models for no obvious gain, so we decided against its use. In addition, the SMOTE approach, combined with the spatial under-sampling, did not prove more accurate than the other models. Oversampling techniques may lead to the overfitting of a model (Weiss, 2004). SMOTE combats overfitting by creating synthetic examples rather than exact copies of the minority class, thereby creating more generality and reducing the chance of overfitting. However, as it did not improve the accuracy of our models, it was not necessary for our data.

Our results indicate that tricoloured blackbirds are positively associated with grassland, pasture and wetland habitats, and negatively associated with high elevations or evergreen forests during both the winter and breeding seasons. Our results also indicate a positive association with coastal areas and fewer definitive hotspots of

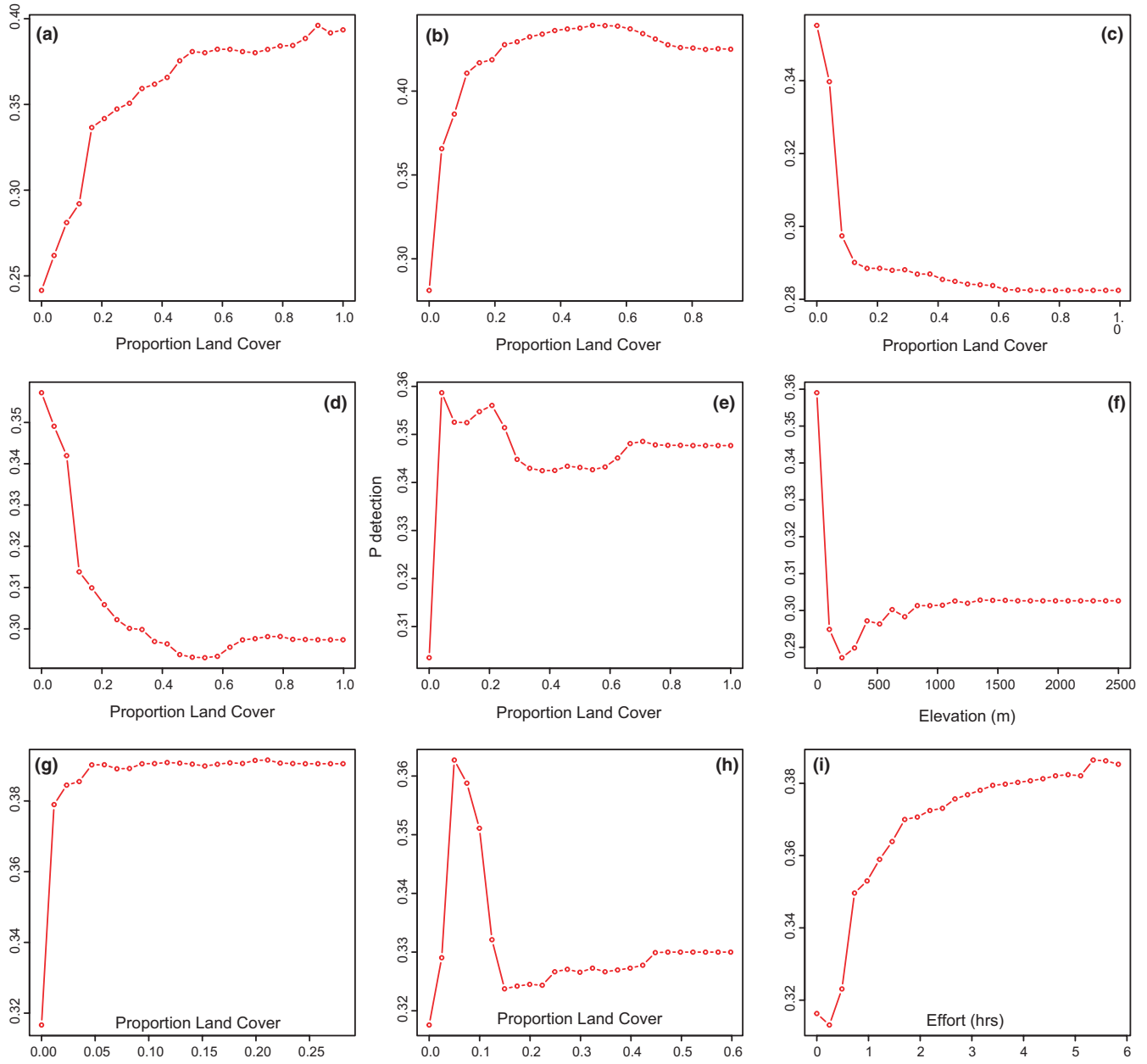

FIGURE 5 The partial dependence of detection probability (P detection) on the nine most important variables chosen in our Fall habitat random forest analysis. (a) Grass/Pasture, (b) Herbaceous Wetlands, (c) Evergreen Forest (d) Shrubland, (e) Open Water, (f) Elevation, (g) Woody Wetlands, (h) Developed/ Low Intensity Day, (i) Effort (hrs)

concentration overall, which agrees with anecdotal evidence and historical accounts (DeHaven et al., 1975; Meese et al., 2014).

We found that the non-crops were most highly associated with TRBL because of the extent and resolution of the analysis. Evergreen forest and high elevation essentially define the entire eastern boundary of the range. The grassland/pasture and herbaceous wetland habitats are also very large in comparison with the relatively small triticale and alfalfa patches as a fraction of the study extent. Thus, associations with these cover classes are generally stronger than those for the crop classes, as they effectively define the range boundaries and general distribution for TRBL. The crop classes act to modify the distribution within the range, and because of relatively small footprint of the crop classes associated with TRBL, these effects are relatively small. The highest ranked single crop is alfalfa (Figure S2.1), the fifteenth ranked predictor in the breeding season. Tricoloured blackbirds are known to breed in silage fields and forage for insects in alfalfa during the breeding season (Weintraub et al., 2016). We believe that with finer resolution eBird data, we may have been able to see stronger relative effects of these individual crops, reflecting the finer scale distribution of these crops in the landscape. However, using too fine of a resolution could lead to location error in eBird checklists.

Our predictions for the distribution during the breeding season suggest that tricoloured blackbirds move away from the coasts

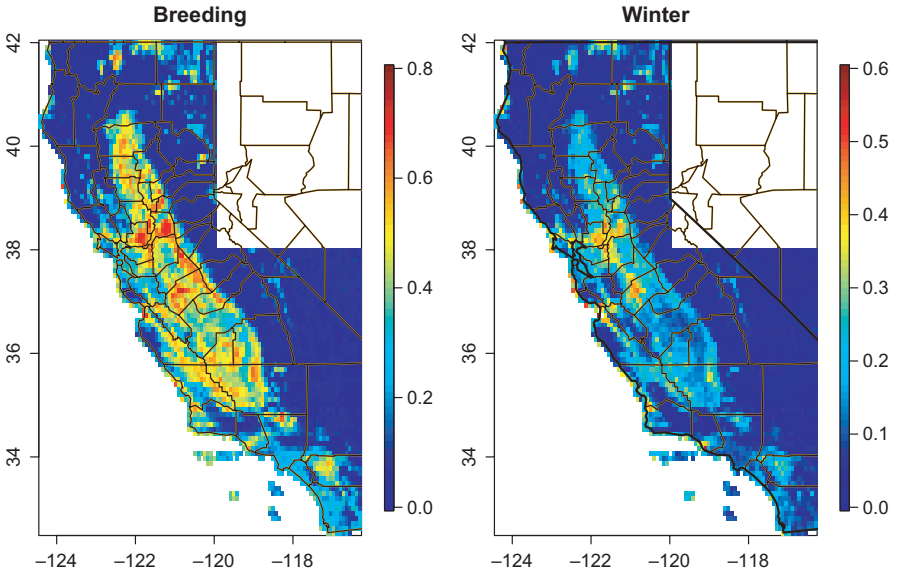

FIGURE 6 Predicted distribution maps for tricoloured blackbirds in the breeding season (left panel) and winter (right panel). The colours represent the probability of detection for a tricoloured blackbird at a given pixel. In pixels with warmer colours, there is a higher probability of detecting a tricoloured blackbird. Note the different scale for each panel

towards the interior of California, particularly the Central Valley, during this time of year. Our model identified central Merced County and the San Joaquin and Sacramento River deltas as important wintering areas (Figure 6). These areas are dense with wetland habitats, and a large portion of the land in central Merced County is managed to preserve grassland and wetland habitats. Much of the portion of Merced County that our SDM predicted as an area of high probability of occurrence is part of the Grassland Ecological Preserve and identified by Audubon as an Important Bird Area (http://www.audubon.org/important-bird-areas). This region is managed through a cooperation among private landowners, and state and federal agencies; continued efforts to manage these grassland and wetland habitats are likely to be important for the long-term persistence of tricoloured blackbird populations.

The general spatial extent of our predicted distributions shows a high degree of overlap between the breeding and non-breeding seasons. However, we did not find areas where occurrence is predicted to be relatively high during the winter. This supports current anecdotal information as historical accounts suggest that they form much smaller groups and flock with other blackbird species in the winter (Meese et al., 2014). This behaviour is also likely to make individuals more difficult to detect during this time of the year versus the breeding season, where they form large conspecific colonies (Meese et al., 2014). This is reflected in our variable importance measures as the effort variables representing the duration and distance travelled during a checklist were among the top 12 predictors (#9 and #12, respectively) for winter occurrence, while they were slightly less important predictors for the breeding season abundance (#12 and #13). What this suggests is that eBird users were required to search longer and travel greater distances to detect a tricoloured blackbird during the winter.

We also had fewer checklists for the winter than the breeding season. As such, there was more information available for training the spring distribution model, even after applying spatial under-sampling. This temporal bias in the citizen science data could have potentially led to a decrease in accuracy for the predicted winter distribution relative to the breeding season. One way to decrease this temporal bias in the data would be to incentivize eBird users to collect more checklists during the winter (e.g. Avicaching; Xue, Davies, Fink, Wood, & Gomes, 2016). Xue et al. (2016) showed that participants in citizen science could be influenced to collect data in under-sampled areas and that the data collected produced more accurate distribution maps than those based on data that were not part of the incentivized study. It may be possible to incentivize observers to sample during under-sampled periods of the year. Our results for the winter season may provide guidance on where to incentivize sampling during this time in the annual cycle of tricoloured blackbirds.

Estimating distributions for rare species often requires numerous surveys that often result in few detections, creating a challenge for the application of data-hungry SDM models to make inferences on habitat associations and distribution patterns of species. Detections of rare species are often limited, and many surveys rely on volunteers that cannot implement emerging techniques such as direct sampling (e.g. Conroy et al., 2008; Guisan et al. 2006; Pacifici et al., 2012). Citizen science is an underused means by which to add samples to a study or monitoring effort at low cost, and participation in such programs is rapidly increasing (Theobald et al., 2015). Species such as tricoloured blackbirds that have low site fidelity also present a challenge in SDMs because it is difficult to know where to sample for them in a given year. Citizen science data can improve sampling effort for these types of species because the spatial coverage of citizen science observations may be far greater than structured surveys with fixed spatial locations used in many studies. Citizen science data for any taxonomical group are subject to many of the same issues affecting all scientific research (Bird et al., 2014) and perhaps even more prone to them (Geldmann et al., 2016). Here, we show how citizen science data can be used to model the distribution of a rare species by combatting the issues of spatial bias and class imbalance by spatially under-sampling the data on which the model is trained.

ACKNOWLEDGEMENTS

We thank Robert Meese, Marcel Holyoak, Emilie Graves, Samantha Arthur, Russ Faucett, and Chad Wilsey for their helpful feedback, knowledge and support for this work. We also thank Tom Auer for his help with compiling the data used in the analyses, the eBird participants for their contributions, and both reviewers for their constructive suggestions. This work was funded by The Leon Levy Foundation, The Wolf Creek Foundation, and the National Science Foundation (ABI sustaining: DBI-1356308; computing support from CNS-1059284 and CCF-1522054).

DATA ACCESSIBILITY

All data used for these models may be requested from eBird (http://ebird.org/ebird/data/download).

ORCID

Orin J. Robinson http://orcid.org/0000-0001-8935-1242

REFERENCES

Bird, T. J., Bates, A. E., Lefcheck, J. S., Hill, N. A., Thomson, R. J., Edgar, G. J., … Frusher, S. (2014). Statistical solutions for error and bias in global citizen science datasets. *Biological Conservation*, *173*, 144–154. https://doi.org/10.1016/j.biocon.2013.07.037

Boria, R. A., Olson, L. E., Goodman, S. M., & Anderson, R. P. (2014). Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. *Ecological Modelling*, *275*, 73–77. https://doi.org/10.1016/j.ecolmodel.2013.12.012

Boryan, C., Yang, Z., Mueller, R., & Craig, M. (2011). Monitoring US agriculture: The US Department of Agriculture, National Agricultural Statistics Service. Cropland Data Layer Program. *Geocarto International*, *26*, 341–358. https://doi.org/10.1080/10106049.2011.562309

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

Brotons, L., Thuiller, W., Arau′ jo, M. B., & Hirzel, A. H. (2004). Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. *Ecography*, *27*, 437–448. https://doi.org/10.1111/j.0906-7590.2004.03764.x

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmayer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Reseaerch*, *16*, 321–357.

Chen, C., Liaw, A., & Breiman, L. (2004). *Using random forest to learn imbalanced data* (110 pp). Berkeley, CA: University of California.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46. https://doi.org/10.1177/001316446002000104

Conn, P. B., Thorson, J. T., & Johnson, D. S. (2016). Confronting preferential sampling in wildlife surveys: Diagnosis and model-based triage. *bioArxive*, https://doi.org/10.1101/080879

Conroy, M. J., Runge, J. P., Barker, R. J., Schofield, M. R., & Fonnesbeck, C. J. (2008). Efficient estimation of abundance for patchily distributed populations via two-phase, adaptive sampling. *Ecology*, *89*, 3362–3370. https://doi.org/10.1890/07-2145.1

Courter, J. R., Johnson, R. J., Stuyck, C. M., Lang, B. A., & Kaiser, E. W. (2012). Weekend bias in citizen science data reporting: Implications for phenology studies. *International Journal of Biometeorology*, *57*, 715–720.

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., & Hess, K. T. (2007). Random forests for classification in ecology. *Ecology*, *88*, 2783–2792. https://doi.org/10.1890/07-0539.1

Dahl, T. E. (1990). *Wetlands losses in the United States 1780's to 1980's. U.S. Department of Interior and U.S. Fish and Wildlife Service*. Retrieved from https://www.fws.gov/wetlands/Documents/Wetlands-Losses-in-the-United-States-1780s-to-1980s.pdf

DeHaven, R. W., Crase, F. T., & Woronecki, P. P. (1975). Movements of tricolored blackbirds banded in the Central Valley of California, 1965-1972. *Bird-Banding*, *46*, 220–229. https://doi.org/10.2307/4512139

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*, 38–49. https://doi.org/10.1017/S0376892997000088

Fink, D., Damoulas, T., Bruns, N. E., La Sorte, F. A., Hochachka, W. M., Gomes, C. P., & Kelling, S. (2014). Crowdsourcing meets ecology: Hemisphere wide spatiotemporal species distribution models. *AI Magazine*, *35*, 19–30. https://doi.org/10.1609/aimag.v35i2.2533

Fink, D., Hochachka, W. M., Zuckerberg, B., Winkler, D. W., Shaby, B., Munson, M. A., … 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., Elith, J., Hastie, T., & Keith, D. A. (2015). Bias correction in species distribution models: Pooling survey and collection data for multiple species. *Methods in Ecology and Evolution*, *6*, 424–438. https://doi.org/10.1111/2041-210X.12242

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

Geldmann, J., Heilmann-Clausen, J., Holm, T. E., Levinsky, I., Markussen, B., Olsen, K., … Tøttrup, A. P. (2016). What determines spatial bias in citizen science? Exploring four recording schemes with different proficiency requirements. *Diversity and Distributions*, *22*, 1139–1149. https://doi.org/10.1111/ddi.12477

Guisan, A., Broennimann, O., Engler, R., Vust, M., Yoccoz, N. G., Lehmann, A., & Zimmermann, N. E. (2006). Using niche-based models to improve the sampling of rare species. *Conservation Biology*, *20*, 501–511.

Guisan, A., & Zimmerman, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling*, *135*, 147–186. https://doi.org/10.1016/S0304-3800(00)00354-9

Han, W., Yang, Z., Di, L., & Mueller, R. (2012). CropScape: A web service based application for exploring and disseminating US conterminous geospatial cropland data products for decision support. *Computers and Electronics in Agriculture*, *84*, 111–123. https://doi.org/10.1016/j.compag.2012.03.005

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*, 1263–1284.

Hijmans, R. J. (2012). Cross-validation of species distribution models: Removing spatial sorting bias and calibration with a null model. *Ecology*, *93*, 679–688. https://doi.org/10.1890/11-0826.1

Hirzel, A. H., Hausser, J., Chessel, D., & Perrin, N. (2002). Ecological-niche factor analysis: How to compute habitat-suitability maps without absence data. *Ecology*, *83*, 2027–2036. https://doi.org/10.1890/0012-9658(2002)083[2027:ENFAHT]2.0.CO;2

Holyoak, M., Meese, R. J., & Graves, E. E. (2014). Combining site occupancy, breeding population sizes and reproductive success to calculate time-averaged reproductive output of different habitat types: An application to tricolored blackbirds. *PLoS ONE*, *9*, e96980. https://doi.org/10.1371/journal.pone.0096980

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*, Vol. *6*. New York, NY: Springer. https://doi.org/10.1007/978-1-4614-7138-7

Japkowicz, N. (2000). The Class Imbalance Problem: Significance and Strategies. In Proceedings of the 2000 International Conference on

Artificial Intelligence (IC-AI'2000): Special Track on Inductive Learning. Las Vegas, Nevada.

Kadmon, R., Farber, O., & Danin, A. (2004). Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. *Ecological Applications*, *14*, 401–413. https://doi.org/10.1890/02-5364

Kubat, M., Holte, R., & Matwin, S. (1998). Machine learning for the detection of oil spills in satellite radar images. *Machine Learning*, *30*, 195–215. https://doi.org/10.1023/A:1007452223027

Kubat, M., & Matwin, S. (1997). Addressing the Curse of Imbalanced Training Sets: One Sided Selection. In Proceedings of the Fourteenth International Conference on Machine Learning, pp. 179–186 Nashville, Tennesse. Morgan Kaufmann.

Lewis, D., & Ringuette, M. (1994). A Comparison of Two Learning Algorithms for Text Categorization. In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, pp. 81–93.

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

Ling, C., & Li, C. (1998). Data Mining for Direct Marketing Problems and Solutions. In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98). New York, NY: AAAI Press.

Longadge, R., Dongre, S. S., & Malik, L. (2013). Class imbalance problem in data mining: Review. *International Journal of Computer Science and Network*, *2*, 83–87.

Luck, G. W., Ricketts, T. H., Daily, G. C., & Imhoff, M. (2004). Alleviating spatial conflict between people and biodiversity. *Proceedings of the National Academy of Sciences USA*, *101*, 182–186. https://doi.org/10.1073/pnas.2237148100

McCune, J. L. (2016). Species distribution models predict rare species occurrences despite significant effects of landscape context. *Journal of Applied Ecology*, *53*, 1871–1879. https://doi.org/10.1111/1365-2664.12702

Meese, R. J. (2009). *Detection, monitoring, and fates of Tricolored Blackbird colonies in 2009 in the Central Valley of California*. Report submitted to California Department of Fish and Game and U.S. Fish and Wildlife Service, Sacramento, CA, USA. Retrieved from http://tricolor.ice.ucdavis.edu/reports?quicktabs_1=1

Meese, R. J., Beedy, E. C., & Hamilton III, W. J. (2014). Tricolored Blackbird (Agelaius tricolor), The Birds of North America. In P. G. Rodewald (Ed.). Ithaca, NY: Cornell Lab of Ornithology. Retrieved from the Birds of North America: https://birdsna.org/SpeciesAccount/bna/species/tribla https://doi.org/10.2173/bna.423

Mi, C., Huttmann, F., Guo, Y., Han, X., & Wen, L. (2017). Why choose random forest to predict rare species distribution with few samples in large undersampled areas? Three Asian crane species models provide supporting evidence. *PeerJ*, *5*, e2849. https://doi.org/10.7717/peerj.2849

Moore, F. R. (2000). Stopover ecology of nearctic–neotropical landbird migrants: Habitat relations and conservation implications. *Studies in Avian Biology*, *20*, pp. 133.

Pacifici, K., Dorazio, R. M., & Conroy, M. J. (2012). A two-phase sampling design for increasing detections of rare species in occupancy surveys. *Methods in Ecology and Evolution*, *3*, 721–730. https://doi.org/10.1111/j.2041-210X.2012.00201.x

Pacifici, K., Reich, B. J., Dorazio, R. M., & Conroy, M. J. (2015). Occupancy estimation for rare species using a spatially-adaptive sampling design. *Methods in Ecology and Evolution*, *7*, 285–293.

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, *190*, 231–259. https://doi.org/10.1016/j.ecolmodel.2005.03.026

Prendergast, J. R., Wood, S. N., Lawton, J. H., & Eversham, B. C. (1993). Correcting for variation in recording effort in analyses of diversity hotspots. *Biodiversity Letters*, *2*, 39–53. https://doi.org/10.2307/2999649

Sethi, S. A., Dalton, M., & Hilborn, R. (2012). Quantitative risk measures applied to Alaskan commercial fisheries. *Canadian Journal of Fisheries and Aquatic Sciences*, *69*, 487–498. https://doi.org/10.1139/f2011-170

Sullivan, B.L., Aycrigg, J.L., Barry, J.H., Bonney, R.E., Bruns, N., Cooper, C.B., Damoulas, T., Dhondt, A.A., Dietterich, T., Farnsworth, A., & Fink, D. (2014). The eBird enterprise: An integrated approach to development and application of citizen science. *Biological Conservation*, *169*, 31–40. https://doi.org/10.1016/j.biocon.2013.11.003

Theobald, E. J., Ettinger, A. K., Burgess, H. K., DeBey, L. B., Schmidt, N. R., Froehlich, H. E., … Prrish, J. K. (2015). Global change and local solutions: Tapping the unrealized potential of citizen science for biodiversity research. *Biological Conservation*, *181*, 236–244. https://doi.org/10.1016/j.biocon.2014.10.021

Tulloch, A. I. T., Possingham, H. P., Joseph, L. N., Szabo, J., & Martin, T. (2013). Realising the full potential of citizen science monitoring programs. *Biological Conservation*, *165*, 128–138.

Václavík, T., & Meentemeyer, R. K. (2009). Invasive species distribution modeling (iSDM): Are absence data and dispersal constraints needed to predict actual distributions? *Ecological Modelling*, *220*, 3248–3258. https://doi.org/10.1016/j.ecolmodel.2009.08.013

Vaughn, I. P., & Ormerod, S. J. (2005). The continuing challenges of testing species distribution Models. *Journal of Applied Ecology*, *42*, 720–730. https://doi.org/10.1111/j.1365-2664.2005.01052.x

Weintraub, K., George, T. L., & Dinsmore, S. J. (2016). Nest survival of tricolored blackbirds in California's Central Valley. *The Condor: Ornithological Applications*, *118*, 850–861. https://doi.org/10.1650/CONDOR-16-56.1

Weir, L. A., & Mossman, M. J. (2005). North American Amphibian Monitoring Program (NAAMP). In M. J. Lannoo (Ed.), *Amphibian declines: Conservation status of United States species* (pp. 307–313). Berkeley, CA: University of California Press. https://doi.org/10.1525/california/9780520235922.001.0001

Weiss, G. M. (2004). Mining with rarity: A unifying framework. *Sigkdd Explorations*, *6*, 7–19. https://doi.org/10.1145/1007730.1007734

Woods, K., Doss, C., Bowyer, K., Solka, J., Priebe, C., & Kegelmeyer, P. (1993). Comparative evaluation of pattern recognition techniques for detection of microcalcifications in mammography. *International Journal of Pattern Recognition and Artificial Intelligence*, *7*, 1417–1436. https://doi.org/10.1142/S0218001493000698

Xue, Y., Davies, I., Fink, D., Wood, C., & Gomes, C. P. (2016). Behavior identification in two-stage games for incentivizing citizen science exploration. Proceedings of the 22nd International Principles and Practice of Constraint Programming. 707-719.

Yackulic, C. B., Chandler, R., Zipkin, E. F., Royle, J. A., Nichols, J. D., Grant, E. H. C., & Veran, S. (2013). Presence-only modelling using MAXENT: When can we trust the inferences? *Methods in Ecology and Evolution*, *4*, 236–243.

BIOSKETCHES

Orin Robinson is a postdoctoral researcher at the Cornell Lab of Ornithology interested in using and developing quantitative tools to learn about vertebrate population and community ecology, and using lessons learned to inform conservation.

Viviana Ruiz-Gutierrez is a quantitative ecologist at the Cornell Lab of Ornithology in Conservation Science and Bird Population Studies. Viviana's research is focused on applying novel statistical models to field observations to ask questions about the dynamics

of bird populations. Viviana is most interested in how anthropogenic drivers (e.g. habitat fragmentation, land use change) affect large- and local-scale patterns of habitat use and distribution.

Daniel Fink is a senior research associate at the Cornell Lab of Ornithology interested in using data collected by citizen science projects to understand ecological phenomena across scales. He is also developing data-mining and semiparametric data analysis techniques for use in analysis of large-scale ecological data.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Robinson OJ, Ruiz-Gutierrez V, Fink D. Correcting for bias in distribution modelling for rare species using citizen science data. *Divers Distrib*. 2017;00:1–13. https://doi.org/10.1111/ddi.12698