Ecological Modelling 422 (2020) 108927

Contents lists available at ScienceDirect

Ecological Modelling

journal homepage: www.elsevier.com/locate/ecolmodel

Estimating species distributions from spatially biased citizen science data

Alison Johnston[a,*], Nick Moran[a], Andy Musgrove[a], Daniel Fink[b], Stephen R. Baillie[a]

[a] British Trust for Ornithology, The Nunnery, Thetford, IP24 2PU, UK
[b] Cornell Lab of Ornithology, 159 Sapsucker Woods Road, Cornell University, Ithaca, NY, 14850, USA

ARTICLE INFO

Keywords:
BirdTrack
Citizen science
Occupancy models
Preferential sampling
Spatial bias
Species distribution models

ABSTRACT

Ecological citizen science data are rapidly growing in availability and use in ecology and conservation. Many citizen science projects have the flexibility for participants to select where they survey, resulting in more participants, but also spatially biased data. It is important to assess the extent to which these spatially biased data can provide reliable estimates of species distributions. Here we quantify the extent of site selection bias in a citizen science project and the implications of this spatial bias in species distribution models. Using data from the BirdTrack citizen science project in Great Britain from 2007 to 2011, we modelled the spatial bias of data submissions. We next produced species occupancy models for 138 bird species, and assessed the impact of accounting for spatial bias. We compared the distributions to those produced using unbiased data from an Atlas survey from the same region and time period. Averaging across 138 species, models with spatially biased data produced accurate and precise estimates of species occupancy for most locations in Great Britain. However, these distributions were both less accurate and less precise in the Scottish Highlands, showing on average a positive bias. Accounting for the spatially biased sampling with weights led to on average greater accuracy in the Scottish Highlands, but did not increase precision. This region is both distinct in environmental characteristics and has a low density of observations, making it difficult to characterise environmental relationships with species occupancy. Accounting for the spatially biased sampling did not affect average accuracy or precision throughout most of the country. Spatially biased citizen science data can be used to estimate species occupancy in regions with stationary environmental relationships and good sampling across environmental space. The reliability of estimated species distributions from spatially biased data should be further validated and tested under a range of different scenarios.

1. Introduction

Datasets collected by citizen scientists are increasingly being used to answer a wide range of ecological questions, partly due to their cost effectiveness relative to professional surveys (Powney and Isaac, 2015; Silvertown, 2009). Ecological citizen science programmes range widely in geographic scope from a single island (e.g. White et al., 2015) to global (e.g. Newson et al., 2016; Sullivan et al., 2009), in species scope from single species (e.g. Howard and Davis, 2009) to all species (e.g. Karns et al., 2006) and from highly structured methods (e.g. Harris et al., 2016; Newson et al., 2015) to completely unstructured data collection (e.g. Pocock et al., 2015). Unstructured surveys that allow participants substantial flexibility in location and style of survey will often attract a large number of participants with a wide range of skill and expertise (Kelling et al., 2015; Pocock et al., 2017). However, this flexibility results in participants choosing sites they enjoy visiting and this introduces spatial bias into the resulting data.

In these flexible and unstructured surveys, citizen scientists select sites that they visit based on one or several criteria. Observers may record from particular sites because they are easily accessible; either near to their home (Dennis and Thomas, 2000; Luck, 2007) or close to roads and paths (Botts et al., 2011; Hijmans et al., 2000; Kadmon et al., 2004; Keller and Scallan, 1999; Mair and Reute, 2016; Reddy and Dávalos, 2003; Tiago et al., 2017). Alternatively they may select sites for ecological reasons; selecting protected areas (Boakes et al., 2010; Botts et al., 2011; Freitag et al., 1998; Reddy and Dávalos, 2003; Tulloch et al., 2013), sites with a high species diversity (Dennis and Thomas, 2000; Hijmans et al., 2000; Prendergast et al., 1993; Tulloch et al., 2013), or sites with expected presence of a particularly interesting species (Boakes et al., 2010; Booth et al., 2011; Greenwood, 2007; Tulloch et al., 2013; Tulloch and Szabo, 2012). In reality, each citizen scientist will likely consider several of these factors and the

* Corresponding author. Present address: Cornell Lab of Ornithology, 159 Sapsucker Woods Road, Cornell University, Ithaca, NY, 14850, USA
E-mail addresses: aj327@cornell.edu (A. Johnston), nick.moran@bto.org (N. Moran), andy.musgrove@bto.org (A. Musgrove), daniel.fink@cornell.edu (D. Fink), stephen.baillie@bto.org (S.R. Baillie).

https://doi.org/10.1016/j.ecolmodel.2019.108927
Received 11 September 2019; Received in revised form 24 December 2019; Accepted 24 December 2019
Available online 04 March 2020

importance of each factor will vary among participants (Booth et al., 2011; Tulloch and Szabo, 2012). However, we expect there will be common patterns among individuals in factors determining site selection. Aggregating across all participants, these site selection preferences can lead to strong spatial biases in recording locations (Bird et al., 2014; Dickinson et al., 2010; Mair and Reute, 2016). Furthermore, such spatial bias is likely to be particularly strong when survey guidelines are flexible and therefore the participants comprise a large number of people with a broad range of skills and motivations (Geldmann et al., 2016).

In analyses of citizen science data it is important to consider the impact of spatial bias, which can lead to biased estimates of species distributions (Boakes et al., 2010; Conn et al., 2013; Yang et al., 2013). Site selection bias can be put into two broad categories: 1) biases that are independent of the response variable and 2) biases that are correlated with the species response (Diggle et al., 2010). Biases in category 1 will lead to inferred environmental relationships between species and habitat that are dominated by regions with more surveys. However, these category 1 biases can often be separated from the response, given the independence between the response and the bias. Biases in category 2 can cause problems because the inferred environmental relationships are confounded with the spatial bias in site selection. For example, many observers may choose to survey wetland habitat that is highly correlated with the presence of wetland species. This spatial bias is more challenging to deal with, because it is harder to separate the bias and the response. In citizen science surveys with a larger scope than a single species, observers are likely to target a variety of sites with high species richness (Dennis and Thomas, 2000; Prendergast et al., 1993), which may be correlated with, but not directly related to the occurrence of any individual species.

When estimating species distributions from presence-only data, it is important that the spatial bias of presence-only data match the spatial bias among the selected pseudo-absence data. For this reason, models for presence-only data have a longer history of addressing spatial bias, because it is critical to account for the difference between species absence and lack of sampling effort (Beck et al., 2014; Fithian et al., 2015; Mair and Reute, 2016; Phillips et al., 2009). A common method to account for spatial bias is to generate pseudo-absences with the same spatial bias as the presence data (e.g. Higa et al., 2014; Phillips and Dudík, 2008). Where it is possible to statistically characterise the mechanisms of spatial bias, this process will generate a dataset of presences and pseudo-absences that are all from the same spatial process.

When dealing with detection/non-detection data (or presence/absence data), the non-detections are already generated with the same process of spatial bias as the detections. Therefore, the dataset is already at the point that many correction methods target for presence-only data. However, even with this situation there is potential for further impacts of spatial bias. For example, relationships between species occurrence and environment can vary across space (Zuckerberg et al., 2016). When the detection/non-detection data are spatially biased, then the inferred relationships will be dominated by the locations or habitats that contribute the most data. In this paper we examine the effects of such spatial bias, where the non-detections are already generated with the same process as the detections.

Two main analytical approaches have been used to account for spatial bias in detection/non-detection data. First, spatial filtering can reduce the spatial bias by selecting a subset of data that has a more even spatial distribution (Araújo and Guisan, 2006). Spatial filtering improves the spatial evenness of the data, but at the cost of reducing the sample size. Spatial filtering is effective at reducing a predominance of data from certain regions by reducing the larger scale spatial bias. However, it will usually not reduce smaller-scale bias caused by people selecting certain habitats. Spatial filtering often has only a small effect on ecological conclusions (Beck et al., 2014; Geldmann et al., 2016; Kadmon et al., 2004). A second option is to estimate the probabilities of site selection based on environmental covariates and then use these probabilities to adjust or balance the empirical distribution of the covariates. In practice, these probabilities are either incorporated as model weights (Rosenbaum and Rubin, 1983) or used to produce a *post-hoc* stratification to correct the analysis (Van Turnhout et al., 2008). When the site selection process can be assumed to be independent of the species response (category 2 biases) theoretical results show that using weights can eliminate bias (Heckmann, 1979). The use of weighting methods to address this problem has been studied in a number of disciplines, including in statistics where the weights are known as propensity scores (Guo and Fraser, 2014) and in machine learning where the problem is known as covariate shift (Sugiyama and Kawanabe, 2012). A third emerging option for spatially biased data is to jointly model the site selection process and the ecological response, which is an extension of the second option (Conn et al., 2016; Diggle et al., 2010; Pati et al., 2011).

Here we characterise spatial bias in site selection in BirdTrack, a flexible citizen science scheme, and quantify the impact on species distributions of accounting for this bias. We model the process of site selection bias and produce a statistical description of the spatial bias. Preliminary analyses revealed there were many covariates associated with the spatial bias and that the relationships could not be adequately described by a simple model. This precluded option three; the joint distribution and observation model. Therefore we modelled spatial bias using a flexible machine learning approach and used the predictions to weight subsequent species distribution models (option two above). We quantified the differences in predictive performance of estimated species distributions when we accounted for the spatial bias in site selection. We validated the models against high quality data with minimal spatial bias and identified situations where inference with spatially biased data is less reliable.

2. Methods

Our methods comprise four analytical stages: In stage 1, we describe the BirdTrack data used for the analysis. In stage 2, we classify locations (1 km squares) into two categories based on how popular they are to visit. In stage 3, we model the probability of a BirdTrack list in each location, using a variety of environmental variables. In stage 4, we use occupancy models to quantify the distributions of 138 species. We run one standard model and one model using the results from stage 3 as weights. Further details of the variables and models are provided in Appendix A. We compare the results of the occupancy models to models produced using a standardised and unbiased dataset from the same time period.

2.1. Stage 1: data selection

The main dataset for the analysis was BirdTrack, a citizen science project in which participants collect records of species they detected at given locations. The data collection protocol is relaxed and many participants contribute to this project, representing a wide range of motivations, skill, knowledge, and behaviour. In this study we used a subset of the full dataset, which included records of bird species recorded in Great Britain between 2007–2011. We used only 'complete' lists, in which observers reported all the species they were able to detect and identify, and we converted counts of birds to a detection/non-detection response variable. We also applied a number of other data filters in order to produce a more standardised and consistent dataset. See Appendix A for details.

2.2. Stage 2: site classification

We wanted to understand the environmental factors that determine the types of locations from which BirdTrack users submit lists. However, we expected that there would be considerable heterogeneity in the reasons that BirdTrack users decide to visit certain locations. For

example, one location may be near to their home, whilst another might have a good variety of habitats, and another may often host rare birds. These different types of sites are likely to be associated with different environmental features. In this stage we wanted to identify different types of sites, using a clustering algorithm. The data showed strong evidence of clustering (Hopkin's statistic = 0.07) (Hopkins, 1954) and the silhouette method suggested there were two clusters (Kaufman and Rousseeuw, 1990). We therefore used a k-means clustering algorithm (Haritgan and Wong, 1979) to split the locations into two clusters based on seven covariates describing the popularity of a site. The covariates for each site (1 km square) used in the clustering algorithm were: number of lists, number of observers, average number of lists per observer, median distance from the observer's home, median number of species per list, and two rarity scores which describe a) the average rarity of species listed at those locations and b) the propensity for the locations to host particularly rare species. All analyses were conducted in R (R Core Team, 2016) using package 'factoextra' (Kassambara and Mundt, 2016).

2.3. Stage 3: Models of site selection

We aimed to predict the locations at which BirdTrack users create lists. We achieved this by characterising the environments that are typical for sites with lists. The multinomial response variable was whether a site was: cluster 1, cluster 2, or had no lists. We associated this response with a number of environmental variables. These variables were mostly summarised within 3 km × 3 km squares and were standardised before modelling. The covariates described characteristics of landcover (25 variables), geography (easting, northing, elevation), housing density (postcode density at five spatial scales), roads (total road length, distance to nearest road) and nature reserves (combinations of bird reserves, nature reserves, and designated sites).

Preliminary analyses revealed that parametric models did not sufficiently characterise the relationship between the environment and the response. Therefore we used a Generalized Boosted Regression Tree (GBRT) for this model, because it enables a wide variety of non-parametric relationships, a large number of interactions between variables, and can accommodate a large number of potentially correlated variables. The regression tree identifies non-parametric relationships between the predictors and the probability of the categorical responses. These relationships can be more difficult to interpret than standard parametric models, but here we were more concerned with good descriptions of the patterns than an understanding of the environmental relationships.

We wanted to validate the models with independent data that were not included in the model training. We therefore removed approximately 25 % of the locations in the UK. To ensure a more rigorous test of the environmental relationships (Fourcade et al., 2018), we divided Great Britain into 100 km × 100 km squares and systematically selected one in every 4 squares (Fig. S1). Data within these 25 % of squares were removed prior to modelling and were used to validate the resulting model. By removing large squares we ensure that validation cannot be driven by small-scale spatial autocorrelation (Hochachka et al., 2010). We used the model built using 75 % of the squares to predict the locations of lists in the 25 % of squares. We used Area Under the Curve (AUC) and the True Skill Statistic (TSS) to compare these predictions to the observed locations of lists in the 25 % validation set.

The multinomial GBRT models were fitted in the R package 'gbm' (Ridgeway, 2013). We used the GBRT model to estimate the probability of at least one BirdTrack list from each 1 km square in Great Britain. See Appendix A for further details of the model covariates and parameters.

2.4. Stage 4. Species distribution models

2.4.1. BirdTrack species distributions

To estimate the occurrence of each species within a 1 km square, we used occupancy models (MacKenzie et al., 2002). We wanted to compare the models using BirdTrack data to models using data from the Bird Atlas 2007–11 (Balmer et al., 2013) (hereafter, the 'Atlas'), so we selected BirdTrack observations that fell within the Atlas breeding season monitoring period (April–July, 2008–2011). Occupancy modelling relies on multiple visits to the same location. As repeat visits, we used visits by the same observer to the same location (1 km square). We selected any observer-location combination that had at least one visit in the breeding season. Sites with only one visit do not contribute to the estimate of detectability, but they do contribute to the estimate of occupancy. The resultant BirdTrack dataset (April–July, 2008–2011), had 31,911 lists from 13,279 unique location-observer combinations across 10,023 distinct locations.

Each BirdTrack list was located within a 1 km × 1 km square, and as covariates of occupancy we summarised information from the 3 km × 3 km surrounding square. The rationale for this was firstly that observations could be from outside the boundaries of the 1 km square, and secondly that birds move around the landscape into nearby environments. Using environmental data from the 3 km square will therefore increase the accuracy of the environmental information (relative to the bird observations), but reduces the precision. The environmental covariates used were: geographic location (easting and northing), island group, mean elevation (Jarvis et al., 2008), the area of 27 different landcover types (Morton et al., 2011), total seasonal rainfall (Met Office, 2016), and seasonal mean temperature (Met Office, 2016). The equation defining the occupancy (ψ) was:

$$logit(\hat{\psi}_k) = \alpha_0 + \sum_{g=1}^{4} \alpha_g island_{gk} + \alpha_5 elevation_k + \alpha_6 elevation_k^2 + \alpha_7 easting_k + \alpha_8 northing_k + \sum_{h=1}^{27} \beta_h area_{hk} + \sum_{s=1}^{4} \gamma_s rainfall_{sk} + \sum_{s=1}^{4} \delta_s temp_{sk}$$

Where ψ_k is the occupancy of location k, α β γ δ are estimated coefficients, $island_{gk}$ is a binary variable with 1 if location k is within island group g (Scilly, Orkney, Shetland, Outer Hebrides) and 0 otherwise, $elevation_k$ is the mean elevation within location k, $easting_k$ and $northing_k$ are the central geographic coordinates of location k on the UK OS grid system, $area_{hk}$ is the area of landcover h in location k, $rainfall_{sk}$ is the total rainfall in season s at location k, and $temp_{sk}$ is the mean temperature in season s at location k.

As covariates of detectability we used: duration of the observation period for the list, a quadratic function of day and a quadratic function of time. The equation defining the detectability (p) was:

$$logit(\hat{p}_i) = \lambda_0 + \lambda_1 duration_i + \lambda_2 day_i + \lambda_3 day_i^2 + \lambda_4 time_i + \lambda_5 time_i^2$$

Where p_i is the detectability of list i, λ are estimated coefficients, $duration_i$ is the duration of the list (in hours), day_i is the day of the year that the list was conducted, and $time_i$ is the start time of the list (in decimal hours). To conform with standard analyses of unstructured citizen science data, we did not account for variation in participant expertise in the detectability model.

An assumption of occupancy modelling is that the occurrence of a species (presence/absence at a site) does not change during the period of sampling. However, during April–July we expect that the presence of many species will change, due to mortality, productivity, and migratory movements. Although this violates the assumption of closure of a standard occupancy model, this assumption can be relaxed as long as the estimated parameters are interpreted correctly (MacKenzie and Royle, 2005) and where appropriate covariates that describe the patterns of variation are added to the model (Kéry et al., 2010). We allow for temporal changes in occupancy by the quadratic relationship with *day* in the detectability part of the model. Therefore, any seasonal changes in occupancy will be assigned to seasonal variation in

detectability. We therefore interpret the estimated occupancy as the occupancy from the part of the breeding season with highest occupancy. Combining data across four years also involves the assumption that occurrence did not change among years. We therefore interpret estimated occupancy as the species occupancy in the year with highest occupancy and this estimate will be robust if sampling sites were selected randomly in each year. If a species had a rapid change in occupancy *and* sampling was temporally and spatially non-random (for example all northern sites in year 1), then combining across years could produce a biased estimate of the distribution. Therefore, an assumption of our models is that no species were affected by rapid change in distribution *and* strongly biased spatio-temporal sampling.

We used the information on probability of sampling to account for the spatially biased data within the occupancy model. For the standard models, we used a maximum likelihood occupancy model from R package 'unmarked' (Fiske and Chandler, 2011). For each species we ran a second occupancy model where we included a likelihood weighting based on the probability of sampling estimated at any location that was calculated as described in stage 3 of this paper. We adapted the code in package 'unmarked' to include a weighting that multiplied each data point (visit) by a given weight in the likelihood (based on sampling probability). Weightings were the reciprocals of the estimated probabilities of locations having a list, in order to account for differences in sampling intensities in different environments. Each weighting was therefore specific to a given spatial location k and did not vary with visit. Using this procedure, each list from a sparsely-surveyed environment contributes more to the likelihood than a list from densely-surveyed environments. Two models were produced for each species – an unweighted model and a weighted model. We expect the unweighted model to be biased towards the environments with more BirdTrack lists.

Predictions of occupancy were made for each 1 km square in Great Britain with covariate data. These estimated occupancy rates ($\widehat{b}_{l,s}$) for 1 km × 1 km square l and species s were accumulated across each 10 km × 10 km square. For each 10 km × 10 km square L and species s we calculated the average estimated occupancy ($\widehat{B}_{L,s}$), i.e. the mean of the estimated occupancy of species s in each constituent 1 km square:

$$\widehat{B}_{L,s} = \frac{1}{100}\sum_{1}^{100} \widehat{b}_{l \in L,s}$$

2.4.2. Atlas species distribution models

To validate the distribution models from BirdTrack data, we constructed similar distribution models from Atlas data. The Atlas used experienced citizen scientists, standardised survey methods, and crucially had even coverage across the whole country, reducing the impact of spatial bias and site selection. We used data for the breeding season and across Great Britain, which included 58,285 1 -h visits to 18,736 2 km × 2 km squares (tetrads). This Atlas dataset contained data from considerably more locations than the BirdTrack dataset and had even coverage across 10 km × 10 km squares (Fig. 2).

We followed a very similar approach for modelling the Atlas data to that taken with the BirdTrack data, to ensure comparable methodology. As above the model results were aggregated to the metric of average occupancy within each 10 km × 10 km square L for each species s ($A_{L,s}$). See Appendix A for further model details.

2.4.3. Comparing distributions from BirdTrack and Atlas data

In order to assess the quality of the distributions produced from BirdTrack, we made the assumption that the occupancy models produced from Atlas data are the closest estimate we can achieve to true occupancy distributions using this analytical approach and with these species. The standardised and spatially intensive sampling for the Atlas represents one of the highest quality ecological datasets available across a region. There were some species for which Atlas occupancy models did not produce reliable results and we removed these from the suite of species for comparison (keeping those with AUC > 0.80; which was 138 of the 159 species modelled). For each of these 138 species we calculated the difference in mean estimated occupancy in each 10 km × 10 km square from the two BirdTrack occupancy models (unweighted and weighted) and the Atlas occupancy model.

$$D_{L,s} = \widehat{B}_{L,s} - \widehat{A}_{L,s}$$

Where $D_{L,S}$ is the difference between the estimated occupancy from the BirdTrack and Atlas distribution models for location L (10 km × 10 km square) and species s. We calculated the mean and standard deviation of the differences between BirdTrack and Atlas distributions across all species, for each 10 km × 10 km square (D_L). These results will tell us on average across all species, whether BirdTrack data lead to biased estimates in particular locations.

For each species we also calculated the average deviation between BirdTrack and Atlas distributions across all 10 km squares (D_s). To assess how absolute sample size was a factor in the estimated Atlas and BirdTrack distributions, we compared the BirdTrack sample size for each species with the average deviation for each species, $mean(D_s)$. We also compared the BirdTrack sample size for each 10 km square with the average deviation for each 10 km square averaged across species, $mean(d_L)$. These results will tell us on average across all locations, whether volume of BirdTrack data affect the degree of bias in estimates for particular species.

3. Results

3.1. Stage 1: data selection

After filtering, the year-round BirdTrack dataset was comprised of 170,723 lists and over 3.7 million observations of bird species in Great Britain, during 2007–2011 (Table S1).

3.2. Stage 2: site categorisation

We categorised sites with lists into two clusters. Cluster 1 sites comprised 19 % of all the sites with lists and they were characterised by higher site rarity scores and higher median distance from home postcodes of observers (Figs. S5–S7). This suggests that cluster 1 sites have a more interesting bird assemblage and that observers will travel further to visit these locations. There were many cluster 1 sites on the coast and in less populated areas (Fig. S9). Conversely, this suggests that cluster 2 sites are visited for convenience of travel and generally have a less interesting bird assemblage.

3.3. Stage 3: Models of site selection

We associated environmental covariates with the probability that locations were in cluster 1, cluster 2, and or had no lists. There were seven variables that had high relative influence in determining whether a site was cluster 1, cluster 2 or had no lists: three landcover variables (suburban land, improved grassland, and landcover diversity), two house density variables (5 km and 25 km radii), density of roads, and altitude. All of the reserve variables and most of the landcover variables had a low relative influence on probability of a BirdTrack list. The model performed moderately well against the validation data from the removed 100 km × 100 km squares. With the validation data AUC = 0.78 and TSS = 0.39. With the modelled data AUC = 0.82 and TSS = 0.48. The similarity between model performance in the modelled data and validation data suggests that spatial autocorrelation did not have a large effect on the model.

Using the model from stage 3, we predicted the probability that each location (1 km squre) would be either a cluster 1 or a cluster 2 site. This is the equivalent to calculating the probability that the site has a

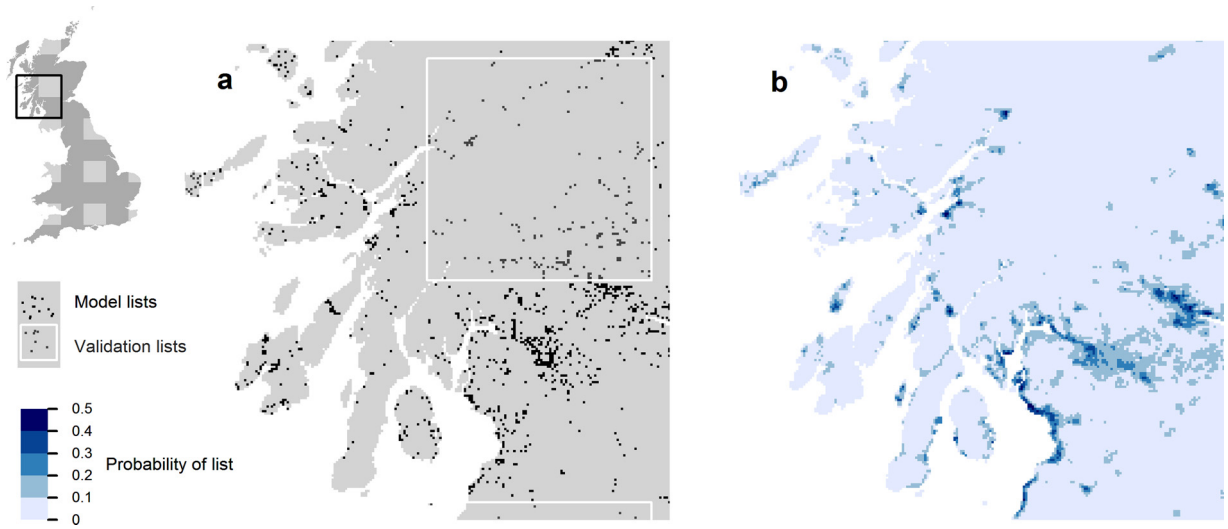

Fig. 1. Observed and modelled estimates of spatial bias in BirdTrack lists.
a) Locations with BirdTrack lists and b) modelled probability of a list submission from a location. The small GB map indicates the zoom selection with a black-bordered square. It also shows in paler grey the 100 km × 100 km squares from which data were withheld from model fitting for validation. The validation data are also evident on map a) within a white-bordered square. Note that only one 1 km × 1 km square within each 3 km × 3 km square was used for model fitting, so approximately one-ninth of the list locations were used for model fitting. The estimated probability of a list is the combined probability of location having a cluster 1 type *or* a cluster 2 type list.

list. All 1 km squares had a probability of a list between 0 and 0.5, with most sites having a low probability (Fig. 1).

3.4. Stage 4. Species distribution models

We present an example of one species distribution, marsh harrier *Circus aeruginosus* (Fig. 2). This species has a relatively restricted breeding distribution within Great Britain. The unweighted model shows a large predicted population in the Scottish Highlands, where there are few BirdTrack lists and the modelled environmental relationships have extrapolated to a population in the Highlands. The distribution from the weighted model does not have a predicted population in the Highlands and closely matches the distribution from Atlas data. This positive effect of weighting by the probability of lists is evident in a number of other species; however, there were species for which the weighting had no effect or a negative effect, producing a predicted distribution that was a worse match to the distribution from Atlas data. In Appendix B we present eight species that demonstrate a range of responses to the weighting and a range of agreement between the distributions estimated from BirdTrack and Atlas data.

Averaging across all 138 species, BirdTrack distributions had positively biased occupancy estimates in the Scottish Highlands (Fig. 3). The BirdTrack distributions also had negatively biased occupancy estimates in coastal locations. The models that were weighted by BirdTrack sampling probability had less positive bias in the Scottish Highlands, but slightly greater negative bias around the coastal squares (Fig. 3). The standard deviation of differences between species was highest in the Scottish Highlands and this remained the case when using the weighted BirdTrack models (Fig. 3). Overall, weighting by the probability of sampling increased the accuracy of the estimated occupancies across species (3b is less pink than 3a), but did not improve precision (3d has similar dark grey area to 3c).

The average difference between estimated occupancy from BirdTrack and Atlas distributions was higher in squares with few BirdTrack observations, and was reduced in the weighted models (which was evident in the Scottish Highlands) (Fig. 4). At sites with a very high number of BirdTrack observations there was a propensity for the BirdTrack occupancy to be negatively biased (which was evident in coastal sites), although the median difference was still zero (Fig. 4). Across species, the average difference at a given location was relatively small. However, within species average differences were higher, suggesting some species had positive biased occupancy estimates in many locations, whilst other species had negatively biased occupancy estimates in many locations. From the total of 138 species, 87 had an average difference less than 0.05 and 124 species less than 0.10. These numbers rose only slightly to 91 and 126 with the weighted models. For 65 species the weighted models were closer to the Atlas distributions and for 73 species the unweighted models were closer to the Atlas distributions. The median difference between the closeness of weighted and unweighted models to the Atlas distributions was 0.00, demonstrating roughly equal number of species had improvements and deteriorations by adding the model weights. The median absolute difference was 0.01, demonstrating that changes were in general very small with the addition of the weights. There was no clear pattern in species positive observations in relation to bias of estimates (Fig. 4), therefore at this scale and with these data and species, the accuracy of BirdTrack models was not directly related to number of positive observations.

4. Discussion

We have demonstrated an analytical approach to model spatial bias in citizen science data and to account for this spatial bias in species distribution models. On average, weighting by sampling density improved the accuracy, but not the precision of estimated species occurrence. However, there were species-specific differences; for some species the weighting improved the estimated occupancy and for others it made the estimates worse. The variation across species was not closely predicted by the species rarity in surveys.

Given the strong spatial biases evident in BirdTrack site selection, it is notable that there is not a larger effect on estimated species distributions. We suggest three reasons for this. Firstly spatial bias would mainly lead to problems if there was non-stationarity in environmental relationships. We expect that on the scale of Great Britain, there is consistency in environmental relationships. Secondly, although there is strong spatial bias, the BirdTrack dataset is relatively large and even in poorly surveyed areas there is a reasonable sample size. Thirdly, there is relatively good coverage of environmental space, which reduces the impact of spatial bias (Higa et al., 2014; Kadmon et al., 2004). The Scottish Highlands are an exception to both the second and third reasons here – they have low sampling intensity and there are several

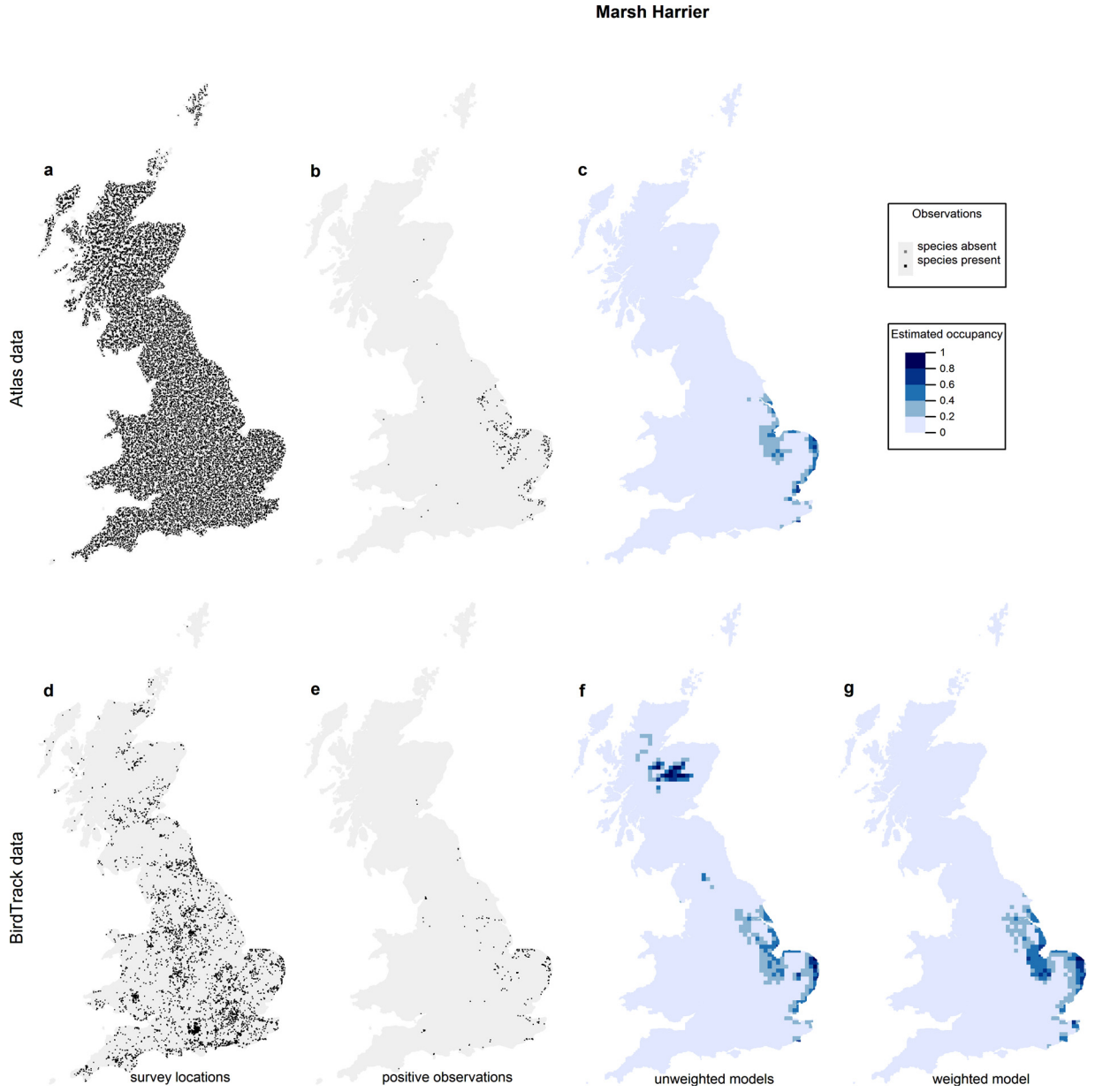

Fig. 2. Observed and modelled occupancy of marsh harrier Circus aeruginosus from Atlas and BirdTrack data.
Data and estimated distributions of marsh harrier from Atlas data (top row) and BirdTrack data (bottom row). Columns show sampling locations (a,d), locations of positive marsh harrier observations (b,e), and estimated distributions (c,f,g). Map (c) is the estimated distribution from Atlas data. Map (f) is the estimated distribution from BirdTrack data and map (g) is the estimated distribution from BirdTrack data, weighting by the probability of sampling to account for the spatial bias in surveyed sites.

environmental variables that have extreme values in this region, leading to relatively novel environmental space with respect to the rest of Great Britain (Fig. S11). We expect that the poor coverage of these novel environmental spaces has reduced the accuracy and precision of estimated species distributions in this area. Locally weighted regression or spatially localised models will naturally adapt to large-scale bias and non-stationarity (Fink et al., 2010; Fotheringham et al., 2003). However, these modelling structures will also tend towards poor estimation when situations two and three occur (small sample size and poor coverage of environmental space) and they will still be susceptible to small-scale spatial bias in site-selection.

Several previous studies also found that accounting for spatial bias did not have a large effect on estimated distributions (Beck et al., 2014; Higa et al., 2014; Kadmon et al., 2004). Although there was not a large effect of spatial bias in this study, we argue that spatial bias is an important consideration in many analyses of citizen science data, particularly those at large spatial scales and where there are substantial differences in sampling intensity between habitats. For example, it may be more critical when overall sampling density is lower (El-Gabbas and Dormann, 2017). We expect that the effect of spatial bias on estimated trends could be substantial (Isaac et al., 2014; Kamp et al., 2016; Zbinden et al., 2014), due to strong spatial patterns in population trends (Massimino et al., 2015). Spatial differences in population trend are likely to be much larger than spatial differences in environmental relationships and are likely to vary within smaller geographic scales.

The greatest benefit of weighting data by sampling density appeared in predictions for regions with the lowest densities of data. For 90 % of species, the distributions from BirdTrack data were close to the Atlas distributions, with an average difference in estimated occupancy of less than 0.10. However, although there was agreement on this broad scale

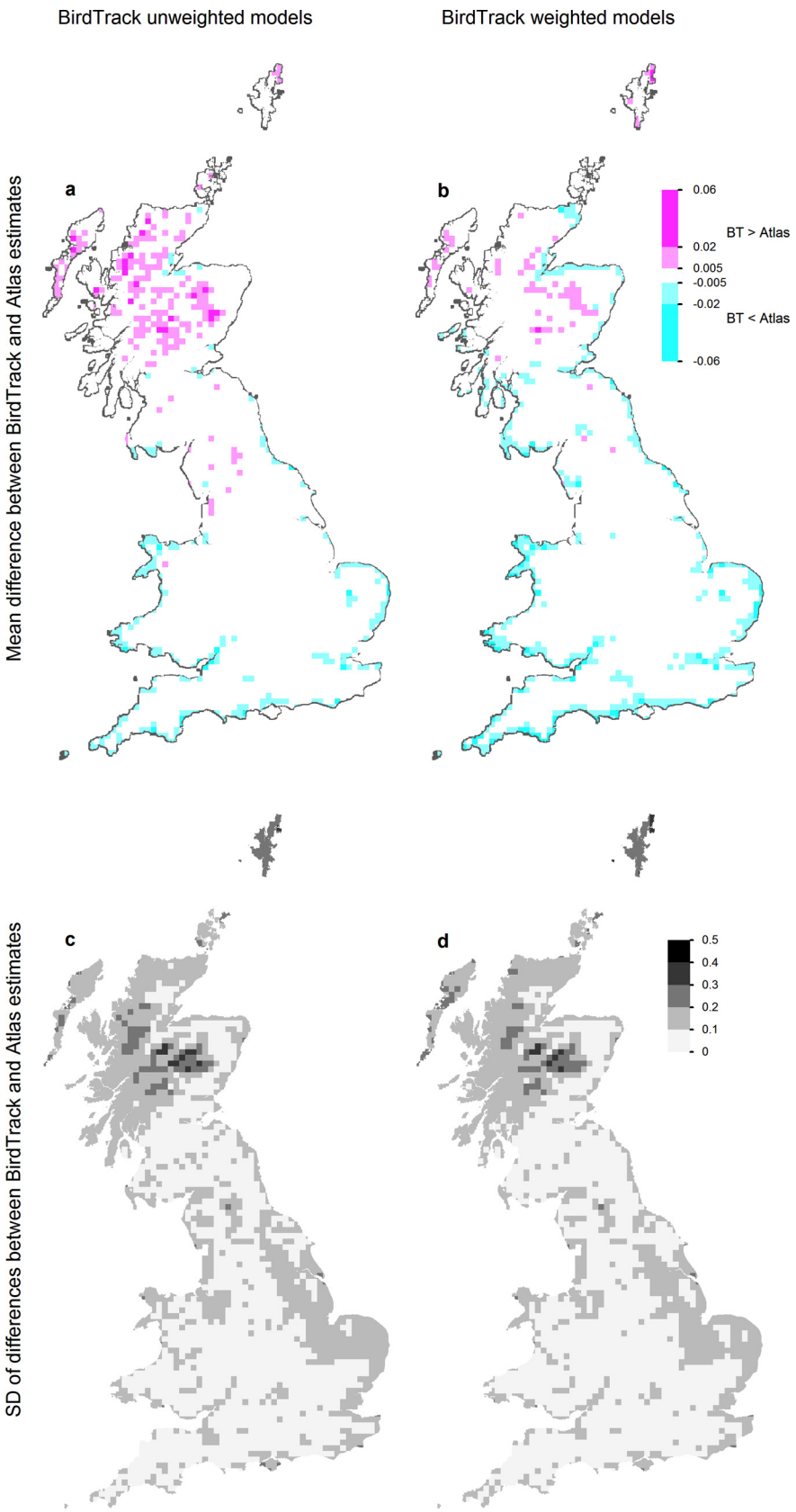

Fig. 3. Average differences between estimated occupancy with Atlas and BirdTrack data.
Differences between estimated distributions from Atlas and BirdTrack data, averaged across all 138 species. The top row indicates mean differences across all species in estimated occupancy (a,b) and the bottom row standard deviation across all species, of differences in estimated occupancy (c,d). The left column compares Atlas models to unweighted BirdTrack models (a,c) and the right column compares Atlas models to weighted BirdTrack models (b,d).

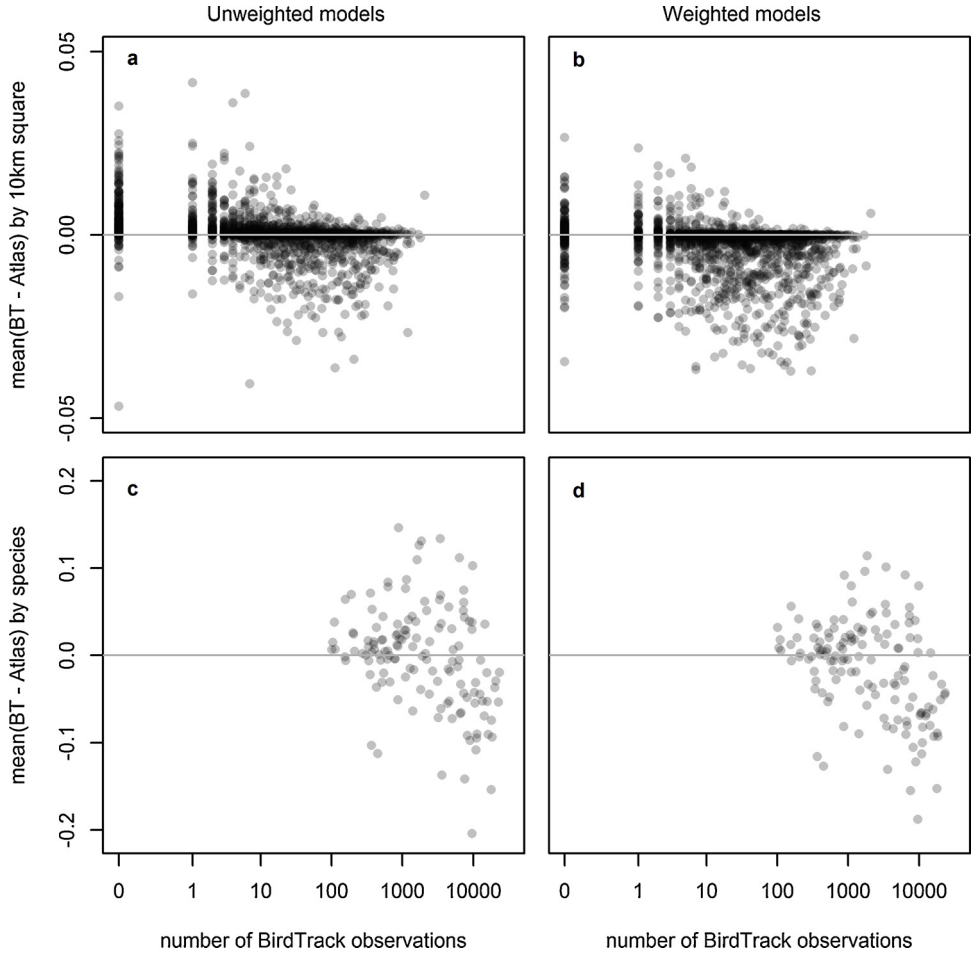

Fig. 4. Sample size and difference between estimated occupancy with Atlas and BirdTrack data.
Number of BirdTrack sampling locations in relation to differences between estimated distributions from Atlas and BirdTrack data, averaged across 10 km × 10 km squares and across all 138 species. The top row shows the average differences within each 10 km square, averaged across species. The bottom row shows the average differences within each species, averaged across locations. The minimum sample size for species inclusion was 100 positive observations.

(averaging across the country for each species), there were notable spatial patterns in the biases (averaging across species for different areas of the country). The Scottish Highlands had higher bias and lower precision than the rest of the country. The weightings reduced the bias (Pati et al., 2011), but did not affect the precision. This suggests that estimates of occupancy from the Scottish Highlands are not accurate within species, which is also evident from some individual species maps (Appendix B).

Estimates of species distributions from BirdTrack data had a negative bias around many coastal sites. This may be because different observers visit different locations. Birdwatchers vary in their expertise (Kelling et al., 2015) and those with different expertise may visit different types of sites, which can lead to bias in estimated species distributions (Johnston et al., 2018). Coastal sites often have high species diversity and attract rare species, which makes them attractive to birdwatchers (Boakes et al., 2010; Booth et al., 2011; Tulloch et al., 2013). These sites would be expected to attract less experienced birdwatchers at a higher rate than other sites, giving the participant pool at these sites a lower average experience. If less experienced birdwatchers are not as skilled at detecting and identifying birds, the lower average experience would give these sites a negative estimated occupancy using BirdTrack data. Overall, greater attention should be given to discriminating the quality of species distributions between regions.

Sites visited by citizen scientists could be categorised into two groups; one cluster of sites had more unusual bird species reported and observers travelled further to visit these sites. This suggests that observers will incur a greater cost (i.e. travel time and costs) to visit a site with a more interesting bird assemblage (Kolstoe and Cameron, 2017). These results concur with previous studies that demonstrate observers show a bias towards rare, threatened or attractive species (Boakes et al., 2010; Dennis et al., 2006; Greenwood, 2007; Tulloch and Szabo, 2012) and that birdwatchers are willing to travel further to see rare species (Booth et al., 2011). In addition, a number of studies have found that there are higher densities of observations in urban areas (Botts et al., 2011; Hijmans et al., 2000; Mair and Reute, 2016; Reddy and Dávalos, 2003), presumably due to proximity to the residences of the majority of observers (Dennis and Thomas, 2000). Many of these studies involve bird records and it is unclear how many of these biases might also apply to other taxa (Reddy and Dávalos, 2003).

Models of site selection associated environmental covariates with the presence of a submitted list and separate relationships were estimated for locations from the two clusters. Across both clusters, habitat and accessibility were important in determining the probability of a list. The habitat variables that had high relative influence for determining the probability of a list were suburban, improved grassland, and landcover diversity. Areas with high suburban landcover are often accessible to many observers, but contain more biodiversity than the truly urban areas. Areas with a high diversity of landcover will often provide a greater range of species in a small area, which would be attractive to many birdwatchers (Tulloch et al., 2013; Kolstoe and Cameron, 2017).

The accessibility variables with high relative influence were road density, altitude, and housing density within 25 km radius and 5 km

radius (although the 5 km radius may also describe aspects of habitat). A wide range of taxa have demonstrated higher rates of sampling near roads (Botts et al., 2011; Hijmans et al., 2000; Kadmon et al., 2004; Keller and Scallan, 1999; Mair and Reute, 2016; Reddy and Dávalos, 2003). In Great Britain, altitude is strongly linked to accessibility, as higher altitude locations are usually more difficult to access. Housing density likely describes the accessibility of a site to local observers. Altogether, the road density, altitude, and housing density all represent the accessibility of sites. The accessibility of sites is more strongly associated with the location of sites in cluster 2 which are nearer to observers' homes and have more common bird assemblages (Figs. S6, S7, S9).

We did not find that any of the reserve variables were important for determining the probability of a list. This result contrasts with many previous studies that have found higher density of sampling within protected areas (Boakes et al., 2010; Botts et al., 2011; Freitag et al., 1998; Reddy and Dávalos, 2003; Tulloch et al., 2013). In Great Britian there are a large number of bird reserves, nature reserves and designated sites. Even within a single category (e.g. bird reserves), there is a very wide range of site characteristics. Some bird reserves are small patches of woodland, whilst others have diverse habitat, a diverse community of birds, and facilities such as a shop and café. We suggest that this wide range of reserve attributes for each reserve category will reduce the effect of the reserve variables in the model. Additionally in Great Britain the difference between reserves and non-reserve areas might be less stark than in some other countries.

5. Conclusions

Overall we demonstrate that spatially biased citizen science data can be used to produce accurate estimates of species distributions, which show consistent bias across species only in locations with both low sampling density and unique environments. Accounting for the spatial bias with weights led to an overall reduction in bias in this region with poor estimation, but no increase in precision. These results demonstrate the utility of large citizen science datasets for estimating species distributions, despite the strong spatial bias in site selection. The approaches set out in this paper provide a methodology that can be applied to the validation and spatial modelling of other citizen science datasets. However, with other datasets, regions, species, or analytical approaches, the effects of spatial bias are likely to be different and we therefore urge caution in directly applying these results to other situations without further testing.

CRediT authorship contribution statement

Alison Johnston: Conceptualization, Formal analysis, Methodology, Project administration, Resources, Writing - original draft. **Nick Moran:** Data curation, Validation, Writing - review & editing. **Andy Musgrove:** Funding acquisition, Writing - review & editing. **Daniel Fink:** Methodology, Writing - review & editing. **Stephen R. Baillie:** Conceptualization, Funding acquisition, Supervision, Writing - review & editing.

Declaration of Competing Interest

The authors have no interests which might be perceived as posing a conflict or bias.

Acknowledgements

We thank the many thousands of citizen scientists who contributed bird records to BirdTrack or Bird Atlas 2007–11. We are grateful to supporters of British Trust for Ornithology (BTO)'s BirdTrack Research Appeal and a legacy from Diana Gay Carr for financial support. BirdTrack is operated by the BTO, and supported by the Royal Society for the Protection of Birds, BirdWatch Ireland, Scottish Ornithologists' Club, the Welsh Ornithological Society and BirdLife International. Bird Atlas 2007–11 was a joint project between BTO, BirdWatch Ireland and the Scottish Ornithologists' Club. We thank all of the organisers and staff of both BirdTrack and the Atlas, particularly Stephen McAvoy, Simon Gillings and Dawn Balmer. We thank Mark Eaton, Graeme Buchanan, Paul Donald, Steffen Oppel, Wesley Hochachka and, Philipp Boersch-Supan for input to the analysis and comments on the manuscript. We thank two anonymous reviewers for comments that improved the manuscript.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.ecolmodel.2019.108927.

References

Araújo, M.B., Guisan, A., 2006. Five (or so) challenges for species distribution modelling. J. Biogeogr. 33, 1677–1688.
Balmer, D.E., Gillings, S., Caffrey, B., Swann, B., Downie, I., Fuller, R.J., 2013. Bird Atlas 2007–2011: The Breeding and Wintering Birds of Britain and Ireland. BTO, Thetford, UK.
Beck, J., Böller, M., Erhardt, A., Schwanghart, W., 2014. Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. Ecol. Inform. 19, 10–15.
Bird, T.J., Bates, A.E., Lefcheck, J.S., Hill, N.A., Thomson, R.J., Edgar, G.J., Stuart-Smith, R.D., Wotherspoon, S., Krkosek, M., Stuart-Smith, J.F., Pecl, G.T., Barrett, N., Frusher, S., 2014. Statistical solutions for error and bias in global citizen science datasets. Biol. Conserv. 173, 144–154.
Boakes, E.H., McGowan, P.J.K., Fuller, R.A., Chang-qing, D., Clark, N.E., O'Connor, K., Mace, G.M., 2010. Distorted views of biodiversity: spatial and temporal bias in species occurrence data. PLoS One 8, e1000385.
Booth, J.E., Gaston, K.J., Evans, K.L., Armsworth, P.R., 2011. The value of species rarity in biodiversity recreation: a birdwatching example. Biol. Conserv. 144, 2728–2732.
Botts, E.A., Erasmus, B.F.N., Alexander, G.J., 2011. Geographic sampling bias in the South African Frog Atlas Project: implications for conservation planning. Biodivers. Conserv. 20, 119–139.
Conn, P.B., McClintock, B.T., Cameron, M.F., Johnson, D.S., Moreland, E.E., Boveng, P.L., 2013. Accommodating species identification errors in transect surveys. Ecology 94, 2607–2618.
Conn, P.B., Thorson, J.T., Johnson, D.S., 2016. Confronting preferential sampling in wildlife surveys: diagnosis and model-based triage. bioRxiv. https://doi.org/10.1101/080879.
Dennis, R.L.H., Shreeve, T.G., Isaac, N.J.B., Roy, D.B., Hardy, P.B., Fox, R., Asher, J., 2006. The effects of visual apparency on bias in butterfly recording and monitoring. Biol. Conserv. 128, 486–492.
Dennis, R.L.H., Thomas, C.D., 2000. Bias in butterfly distribution maps: the influence of hot spots and recorder's home range. J. Insect Conserv. 4, 73–77.
Dickinson, J.L., Zuckerberg, B., Bonter, D.N., 2010. Citizen science as an ecological research tool: challenges and benefits. Annu. Rev. Ecol. Evol. Syst. 41, 149–172.
Diggle, P.J., Menezes, R., Su, T., 2010. Geostatistical inference under preferential sampling. J. R. Stat. Soc. Ser. C-Appl. Stat. 59, 191–232.
El-Gabbas, A., Dormann, C.F., 2017. Improved species-occurrence predictions in data-poor regions: using large-scale data and bias correction with down-weighted Poisson regression and Maxent. Ecography. https://doi.org/10.1111/ecog.03149.
Fink, D., Hochachka, W.M., Zuckerberg, B., WInkler, D.W., Shaby, B., Munson, M.A., Hooker, G.J., Riedewald, M., Sheldon, D., Kelling, S., 2010. Spatiotemporal exploratory models for large-scale survey data. Ecol. Appl. 20, 2131–2147.
Fiske, I., Chandler, R., 2011. Unmarked: an R package for fitting hierarchical models of wildlife occurrence and abundance. J. Stat. Softw. 43, 1–23.
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 Ecol. Evol. 6, 424–438.
Fotheringham, A.S., Brunsdon, C., Charlton, M., 2003. Goegraphically Weighted Regression: The Analysis of Spatially Varying Relationships. John Wiley and Sons, Ltd, Chichester.
Fourcade, Y., Besnard, A.G., Secondi, J., 2018. Paintings predict the distribution of species, or the challenge of selecting environmental predictors and evaluation statistics. Glob. Ecol. Biogeogr. 27, 245–256.
Freitag, S., Hobson, C., Biggs, H.C., van Jaarsveld, A.S., 1998. Testing for potential survey bias: the effect of roads, urban areas and nature reserves on a southern African mammal data set. Anim. Conserv. 1, 119–127.
Geldmann, J., Heilmann-Clausen, J., Holm, T.E., Levinsky, I., Markussen, B., Olsen, K., Rahbek, C., Tøttrup, A.P., 2016. What determines spatial bias in citizen science? Exploring four recording schemes with different proficiency requirements. Divers. Distrib. 22, 1139–1149.
Greenwood, J.J.D., 2007. Citizens, science and bird conservation. J. Ornithol. 148, 77–124.
Guo, S., Fraser, M.W., 2014. Propensity Score Analysis. Sage.

Haritgan, J.A., Wong, M.A., 1979. Algorithm AS 136: a K-means clustering algorithm. J. R. Stat. Soc. Ser. C-Appl. Stat. 28, 100–108.
Harris, S.J., Massimino, D., Newson, S.E., Eaton, M.A., Marchant, J.H., Balmer, D.E., Noble, D.G., Gillings, S., Procter, D., Pearce-Higgins, J.W., 2016. The Breeding Bird Survey 2015 (BTO Reserach Report No. 687). British Trust for Ornithology, Thetford, UK.
Heckmann, J.J., 1979. Sample selection bias as a specification error. Econometrica 47, 153–161.
Higa, M., Yamaura, Y., Koizumi, I., Yabuhara, Y., Senzaki, M., Ono, S., 2014. Mapping large-scale bird distributions using occupancy models and citizen data with spatially biased sampling effort. Divers. Distrib. 21, 46–54.
Hijmans, R.J., Garrett, K.A., Huaman, Z., Zhang, D.P., Scheuder, M., Bonierbale, M., 2000. Assessing the geographic representativeness of gene bank collections: the case of Bolivian wild potatoes. Conserv. Biol. 14, 1755–1765.
Hochachka, W.M., Fink, D., Kelling, S., 2010. Checklist programs as a source of data for Bird Monitoring: designing analyses and model validations to account for unequal spatial and temporal sampling effort. In: Bird Numbers 2010: Monitoring, Indicators and Targets. Presented at the 18th Conference of the European Bird Census Council. Cáceres, Spain. pp. 9–20.
Hopkins, B., 1954. A new method of determining the type of distribution of plant individuals. Ann. Bot. 18, 213–226.
Howard, E., Davis, A.K., 2009. The fall migration flyways of monarch butterflies in eastern North America revealed by citizen scientists. J. Insect Conserv. 13, 279–286.
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 Ecol. Evol. 5, 1052–1060.
Jarvis, A., Reuter, H.I., Nelson, A., Guevara, E., 2008. Hole-filled seamless SRTM data v4. International Centre for Tropical Agriculture (CIAT) [WWW Document]. . URL http://srtm.csi.cgiar.org (Accessed 30 May 2012).
Johnston, A., Fink, D., Hochachka, W.M., Kelling, S., 2018. Estimates of observer expertise improve species distributions from citizen science data. Methods Ecol. Evol. 9, 88–97.
Kadmon, R., Farber, O., Danin, A., 2004. Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. Ecol. Appl. 14, 401–413.
Kamp, J., Oppel, S., Heldbjerg, H., Nyegaard, T., Donald, P.F., 2016. Unstructured citizen science data fail to detect long-term population declines of common birds in Denmark. Divers. Distrib. 22, 1024–1035.
Karns, D.R., Ruch, D.G., Brodman, R.D., Jackson, M.T., Rothrock, P.E., Scott, P.E., Simon, T.P., Whitaker Jr, J.O., 2006. Results of a short-term bioblitz of the aquatic and terrestrial habitats of Otter Creek, Vigo County, Indiana. Proc. Indiana Acad. Sci. 115, 82–88.
Kassambara, A., Mundt, F., 2016. Factoextra: Extract and Visualise the Results of Multivariate Data Analyses.
Kaufman, L., Rousseeuw, P.J., 1990. Finding Groups in Data: An Introduction to Cluster Analysis, Wiley Series in Probability and Statistics. John Wiley and Sons, Inc, Hoboken, New Jersey.
Keller, C.M.E., Scallan, J.T., 1999. Potential roadside biases due to habitat changes along breeding bird survey routes. Condor 101, 50–57.
Kelling, S., Johnston, A., Hochachka, W.M., Iliff, M., Fink, D., Gerbracht, J., Lagoze, C., La Sorte, F.A., Moore, T., Wiggins, A., Wong, W.-K., Wood, C., Yu, J., 2015. Can observation skills of citizen scientists be estimated with species accumulation curves? PLoS One 10, e0139600.
Kéry, M., Gardner, B., Monnerat, C., 2010. Predicting species distributions from checklist data using site-occupancy models. J. Biogeogr. 37, 1851–1862.
Kolstoe, Sonja, Cameron, Trudy Ann, 2017. The non-market value of birding sites and the marginal value of additional species: biodiversity in a random utility model of site choice by eBird members. Ecol. Econ. 137, 1–12. https://doi.org/10.1016/j.ecolecon.2017.02.013.
Luck, G.W., 2007. A review of the relationships between human population density and biodiversity. Biol. Rev. 82, 607–645.
MacKenzie, D.I., Nichols, J.D., Lachman, G.B., Droege, S., Royle, J.A., Langtimm, C.A., 2002. Estimating site occupancy rates when detection probabilities are less than one. Ecology 83, 2248–2255.
MacKenzie, D.I., Royle, J.A., 2005. Designing occupancy studies: general advice and allocating surey effort. J. Appl. Ecol. 42, 1105–1114.
Mair, L., Reute, A., 2016. Explaining spatial variation in the recording effort of citizen science data across multiple taxa. PLoS One 11, e0147796.
Massimino, D., Johnston, A., Noble, D.G., Pearce-Higgins, J.W., 2015. Multi-species spatially-explicit indicators reveal spatially strutured trends in bird communities. Ecol. Indic. 58, 277–285.
Met Office, 2016. UKCP09: Gridded Observation Data Sets [WWW Document]. URL http://www.metoffice.gov.uk/climatechange/science/monitoring/ukcp09/ (Accessed 5 January 2016). .
Morton, D., Rowland, C., Wood, C., Meek, L., Marston, C., Smith, G., Wadsworth, R., Simpson, I.C., 2011. Final Report for LCM2007 - The New UK Land Cover Map (CS Technical Report No. 11/07). Centre for Ecology & Hydrology.
Newson, S.E., Evans, H.E., Gillings, S., 2015. A novel citizen science approach for large-scale standardised monitoring of bat activity and distribution, evaluated in eastern England. Biol. Conserv. 191, 38–49.
Newson, S.E., Moran, N.J., Musgrove, A.J., Pearce-Higgins, J.W., Gillings, S., Atkinson, P.W., Miller, R., Grantham, M.J., Baillie, S.R., 2016. Long-term changes in the migration phenology and UK breeding birds detected by large-scale citizen science recording schemes. Ibis 158, 481–495.
Pati, D., Reich, B.J., Dunson, D.B., 2011. Bayesian geostatistical modelling with informative sampling locations. Biometrika 98, 35–48.
Phillips, S.J., Dudík, M., 2008. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography 31, 161–175.
Phillips, S.J., Dudík, 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. Ecol. Appl. 19, 181–197.
Pocock, M.J.O., Roy, H.E., Preston, C.D., Roy, D.B., 2015. The Biological Records Centre: a pioneer of citizen science. Biol. J. Linn. Soc. 115, 475–493.
Pocock, M.J.O., Tweddle, J.C., Savage, J., Robinson, L.D., Roy, H.E., 2017. The diversity and evolution of ecological and environmental citizen science. PLoS One 12, e0172579.
Powney, G.D., Isaac, N.J.B., 2015. Beyond maps: a review of the applications of biological records. Biol. J. Linn. Soc. 115, 532–542.
Prendergast, J.R., Wood, S.N., Lawton, J.H., Eversham, B.C., 1993. Correcting for variation in recording effort in analyses of diversity hotspots. Biodivers. Lett. 1, 39–53.
R Core Team, 2016. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
Reddy, S., Dávalos, L.M., 2003. Geographical sampling bias and its implications for conservation priorities in Africa. J. Biogeogr. 30, 1719–1727.
Ridgeway, G., 2013. gbm: Generalized Boosted Regression Models.
Rosenbaum, P.R., Rubin, D.B., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41–55.
Silvertown, J., 2009. A new dawn for citizen science. Trends Ecol. Evol. 24, 467–471.
Sugiyama, M., Kawanabe, M., 2012. Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation. MIT Press.
Sullivan, B.L., Wood, C.L., Iliff, M.J., Bonney, R.E., Fink, D., Kelling, S., 2009. eBird: a citizen-based bird observation network in the biological sciences. Biol. Conserv. 142, 2282–2292.
Tiago, P., Ceia-Hasse, A., Marques, T.A., Capinha, C., Pereira, H.M., 2017. Spatial distribution of citizen science casuistic observations for different taxonomic groups. Sci. Rep. 7, 12832. https://doi.org/10.1038/s41598-017-13130-8.
Tulloch, A.I.T., Mustin, K., Possingham, H.P., Szabo, J.K., Wilson, K.A., 2013. To boldly go where no volunteer has gone before: predicting volunteer activity to prioritize surveys at the landscape scale. Divers. Distrib. 19, 465–480.
Tulloch, A.I.T., Szabo, J.K., 2012. A behavioural ecology approach to understand volunteer surveying for citizen science datasets. Emu 112, 313–325.
van Turnhout, C.A.M., Willems, F., Plate, C., van Strien, A., Teunissen, W., van Dijk, R., Foppen, R., 2008. Monitoring common and scarce breeding birds in the Netherlands: applying a post-hoc stratification and weighting procedure to obtain less biased population trends. Rev. Catalana Ornitol. 24, 15–29.
White, E.R., Myers, M.C., Mills Flemming, J., Baum, J.K., 2015. Shifting elasmobranch community assemblage at Cocos island - an isolated marine protected area. Conserv. Biol. 29, 1186–1197.
Yang, H., Zhang, J., Roe, P., 2013. Reputation modelling in citizen science for environmental acoustic data analysis. Soc. Netw. Anal. Min. 3, 419–435.
Zbinden, N., Kéry, M., Häfliger, G., Schmid, H., Keller, V., 2014. A resampling-based method for effort correction in abudance trend analyses from opportunistic biological records. Bird Study 61, 506–517.
Zuckerberg, B., Fink, D., La Sorte, F.A., Hochachka, W.M., Kelling, S., 2016. Novel seasonal land cover associations for eastern North American forest birds identified through dynamic species distribution modelling. Divers. Distrib. 22, 717–730.