Best practices for making reliable inferences from citizen science data: case study using eBird to estimate species distributions

A Johnston*[1,2], WM Hochachka[1], ME Strimas-Mackey[1], V Ruiz Gutierrez[1], OJ Robinson[1], ET Miller[1], T Auer[1], ST Kelling[1], D Fink[1]

* Corresponding author: aj327@cornell.edu

[1] Cornell Lab of Ornithology, Cornell University, 159 Sapsucker Woods Road, Ithaca, NY14850, USA

[2] Conservation Science Group, Dept of Zoology, University of Cambridge, The David Attenborough Building, Pembroke Street, Cambridge, CB2 3QZ, UK

Abstract

Citizen science data are valuable for addressing a wide range of ecological research questions, and there has been a rapid increase in the scope and volume of data available. However, data from large-scale citizen science projects typically present a number of challenges that can inhibit robust ecological inferences. These challenges include: species bias, spatial bias, variation in effort, and variation in observer skill.

To demonstrate key challenges in analysing citizen science data, we use the example of estimating species distributions with data from eBird, a large semi-structured citizen science project. We estimate three widely applied metrics for describing species distributions: encounter rate, occupancy probability, and relative abundance. For each method, we outline approaches for data processing and modelling that are suitable for using citizen science data for estimating species distributions.

Model performance improved when data processing and analytical methods addressed the challenges arising from citizen science data. The largest gains in model performance were achieved with two key processes 1) the use of complete checklists rather than presence-only data, and 2) the use of covariates describing variation in effort and detectability for each checklist. Including these covariates accounted for heterogeneity in detectability and reporting, and resulted in substantial differences in predicted distributions. The data processing and analytical steps we outlined led to improved model performance across a range of sample sizes.

When using citizen science data it is imperative to carefully consider the appropriate data processing and analytical procedures required to address the bias and variation. Here, we describe the consequences and utility of applying our suggested approach to semi-structured citizen science data to estimate species distributions. The methods we have outlined are also likely to improve other forms of inference and will enable researchers to conduct robust analyses and harness the vast ecological knowledge that exists within citizen science data.

Key words

Abundance, citizen science, detectability, eBird, occupancy model, species distribution model

Introduction

Citizen science data are increasingly making important contributions to basic and applied ecological research. One of the most common forms of citizen science data come from members of the public recording species observations. These observations are being collected for a diverse array of taxa, including butterflies (Howard, Aschen, & Davis, 2010), sharks (Vianna, Meekan, Bornovski, & Meeuwig, 2014), lichen (Casanovas, Lynch, & Fagan, 2014), bats (Newson, Evans, & Gillings, 2015), and birds (Sauer et al., 2017). The number of these citizen science projects has been growing exponentially, but they vary widely in complexity, flexibility, and participation (Wiggins & Crowston, 2011; Pocock, Tweddle, Savage, Robinson, & Roy, 2017). Projects occur on a spectrum from those with a predefined sampling structure that resemble more traditional survey designs, to those that are unstructured and collect observations 'opportunistically'. Projects with study designs and defined protocols generally produce data that are more informative for a particular objective, but are often limited to a specific time frame and region and have fewer participants. This can lead to a trade-off between the quality and quantity of data supported by citizen science projects (Bird et al., 2014; Pacifici et al., 2017). Semi-structured citizen science projects have unstructured data collected, but critically also collect data on the observation process, which can be used to address many of the issues arising from citizen science data (Kelling et al., 2018; Altwegg & Nichols, 2019). With the increasing popularity in the use and application of citizen-science data, we describe and evaluate best practices for data analysis, that maximise the value of semi-structured citizen science data (Sullivan et al., 2014).

Data consisting of species observations from citizen scientists present a number of challenges that are not as prevalent in conventional scientific data. Firstly, participants often have preferences for certain species, which may lead to preferential recording of some species over others (Tulloch & Szabo, 2012; Troudet, Grandcolas, Blin, Vignes-Lebbe, & Legendre, 2017). Secondly, the observation process is heterogeneous, as there is large variation in effort, time of day, observers, and weather, all of which can affect the detectability of species (Ellis & Taylor, 2018; Oliveira, Olmos, dos Santos-Filho, & Bernardo, 2018). Thirdly, the locations selected by participants to collect data usually contain strong spatial bias. For example, participants may preferentially visit locations that are close to where they live (Dennis & Thomas, 2000; Mair & Ruete, 2016), more accessible (Kadmon, Farber, & Danin, 2004; Botts, Erasmus, & Alexander, 2011), contain high species diversity (Hijmans et al., 2000; Tulloch, Possingham, Joseph, Szabo, & Martin, 2013), or are within protected areas (Tulloch et al., 2013). Fourthly, data are collected from participants with a wide variety of behaviour, experience, and skill in detecting and identifying species correctly (Cohn, 2008; Bird et al., 2014). However, citizen science data also contain a wealth of ecological knowledge and they are often the only source of biological information for many biodiverse

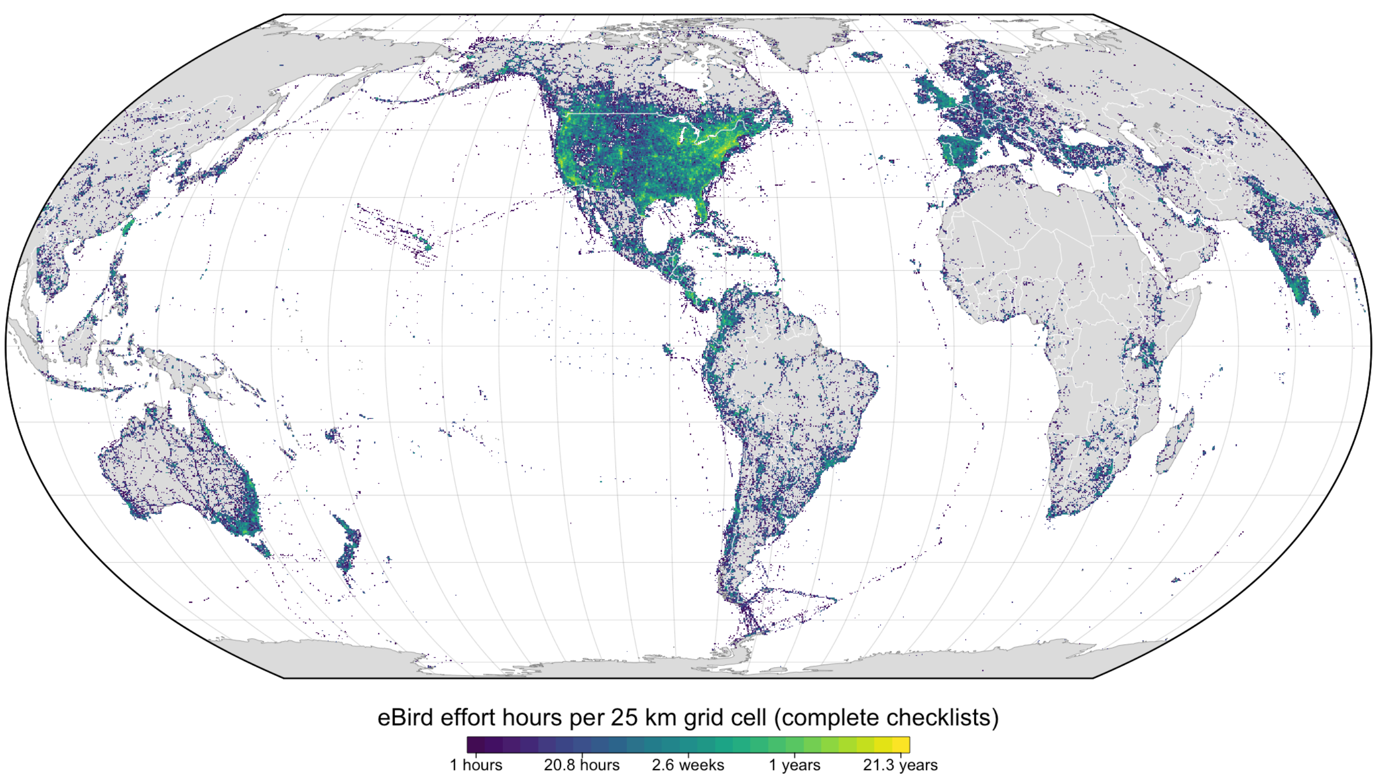

Figure 1. Cumulative duration of eBird complete checklists aggregated within 25 x 25 km grid cells. These aggregations are from eBird data submitted up to the end of Dec 2018.

regions. Therefore it is imperative to define approaches that can maximise the value of increasing volumes of citizen science observations.

There are two main approaches for addressing known challenges related to citizen-science data: 1) imposing a more structured protocol onto the dataset after collection via data filtering (Kamp, Oppel, Heldbjerg, Nyegaard, & Donald, 2016); 2) including covariates in a model to account for the variation (Miller, Pacifici, Sanderlin, & Reich, 2019). In this paper we advocate combining both of these approaches to increase the reliability of inferences made using citizen science observations. To do this, we describe best practices for using semi-structured citizen science data, using the example of estimating species distributions. We assess the impact on estimated distributions when these practices are not followed. Our recommendations focus on the use of eBird data, although they also apply to similar citizen science datasets.

eBird: an example of a semi-structured citizen-science program

Data in eBird

We use the example of the semi-structured citizen science programme eBird (Sullivan et al., 2014), which was originally created as a comprehensive tool and database for collecting high quality bird observations. eBird provides high volumes of data covering global areas with year-round coverage and as of January 2019, the database contained nearly 600 million observations from every country in the world and has been widely used in scientific research to study phenology, species distributions, population trends, evolution, behaviour, global change, and conservation (Mayor et al., 2017; Seeholzer, Claramunt, & Brumfield, 2017; Lang, Mann, & Farine, 2018; MacPherson et al., 2018; Mattsson et al., 2018). However, as with many citizen science datasets, robust inference with eBird data requires careful processing and analysis of the data.

The gold standard: complete checklists with effort information

There are two critical aspects to the structure of eBird data that facilitate robust ecological inference. Firstly, data submitted to eBird are structured as 'checklists', where each checklist is a list of bird species recorded during one period of bird-watching. When these lists are 'complete checklists' the participant recorded all birds that they detected and identified. Critically, a complete checklist enables scientists to infer counts of zero individuals for the species that were not reported (i.e. zero-filling). Complete checklists enable distinguishing between a non-detection or a participant not recording a species detection. Complete checklists therefore are advantageous for many analyses, reducing the impact of participants' taxonomic preferences on the data (challenge 1 above) and reducing the impact of imperfect detection (challenge 2 above) while providing a basis for inference of occupancy rates (Guillera-Arroita et al., 2015). Secondly, eBird is a semi-structured citizen science project, which means most eBird checklists have associated metadata describing the 'effort' or observation process (Kelling et al., 2018) . This effort information includes duration searching for birds, start time, distance travelled, etc. which enable the analyst to account for variation in the observation process. These two key aspects of complete checklists and effort information facilitate robust analyses and enable eBird and other citizen science projects to produce robust ecological results (Kelling et al., 2018; La Sorte et al., 2018).

Data access

The eBird Basic Dataset (EBD) is global in extent and updated monthly (www.ebird.org/science/download-ebird-data-products). Data can be freely accessed via an online data portal and processed with the *auk* R package (Strimas-Mackey, Miller, & Hochachka, 2017). eBird has a robust review process, focussed on ensuring correct locations and species identification, that is conducted before data enters the EBD and we provide further details on this in Supporting Information A2.

Considerations for analysing citizen science data

Citizen science data bring a number of challenges to ecological analyses that are not a consideration with more standardised datasets. Semi-structured or unstructured citizen science data generally have spatial bias and temporal bias, and it is also important to consider spatial precision of the data. Citizen science projects are often designed to survey a wide range of species, and this can lead to class imbalance for any given species with many non-detections and few positive detections. In Supporting Information A2 we describe these challenges in greater detail, particularly with respect to eBird data.

With citizen science data it can be particularly important to consider detectability. In this context 'detectability' describes the probability that an individual or species in a given area will be detected, identified, and recorded by the participant. The detectability of birds by citizen scientists varies by ecological factors such as season, habitat, and species, in addition to observation factors such as time of day and observer (Marques, Thomas, Fancy, & Buckland, 2007; Bas, Devictor, Moussus, & Jiguet, 2008; Lehikoinen, 2013; Kelling et al., 2015). There are two aspects of detectability to consider in species distribution models; detection probability and variation in detectability. Detection probability can be estimated with occupancy modelling and under certain assumptions semi-structured citizen science data can be used to fit occupancy models (Kéry, Gardner, & Monnerat, 2010; Johnston, Fink, Hochachka, & Kelling, 2018). To account for variation in detectability, data can be filtered and covariates that describe the variation can be included in models. Projects that collect variables describing the observation process will be able to account for a larger proportion of the heterogeneity in detectability (Kelling et al., 2018). The appropriate modelling framework depends on the goals of the analysis and the available data (Guillera-Arroita et al., 2015), but a greater variety of models are feasible when only estimating and accounting for variation in detectability.

Data Analysis

We explored the impact of various analytical practices when using citizen science data to estimate species distributions. We used eBird data to estimate the encounter rate of wood thrush *Hylocichla mustelina* in the breeding season within a single Bird Conservation Region. Wood thrush is a relatively common passerine that is easily detected by its song and is generally well-monitored by eBird. By using wood thrush as an example, we assess the impact of not following the practices outlined above. Firstly, we describe the general data filtering procedures we used and then we outline three different modelling approaches to estimate different

Table 1. Descriptions of the elements in models 1-6 that include different aspects of the best practice guidelines. Model 1 just uses the presences without checklists and generates pseudo-absences with background points. Models 2-6 use checklists. Models 3-6 use complete checklists. Models 5-6 use effort data.

Data required	Best practice guidelines	Model number					
		1	2	3	4	5	6
Checklists	Infer non-detections		X				
Complete checklists	Include non-detections			X	X	X	X
	Conduct spatial subsampling				X	X	X
Effort data	Filter the data by effort variables					X	X
Effort data	Include effort data as covariates						X
	Number of checklists	2807	60692	48950	12000	9444	9444
	Number of wood thrush positive observations	2807	2807	2648	1544	1028	1028

ecological metrics of wood thrush distribution: encounter rate, occupancy, and relative abundance. Next we describe how we varied the input data and model structure and assessed the impacts. All analyses were conducted with R (R Core Team, 2018). In Supporting information documents A4 and A5 we provide all the R code used in this paper.

General data selection and filtering

The eBird EBD dataset was filtered by region, season, and by a number of effort variables in order to create a more standardised dataset. We selected checklists from the EBD version released in August 2018. We filtered the checklists to the month of June, within Bird Conservation Region 27 "Southeastern coastal plain". We selected complete checklists with either the 'stationary' or 'traveling' protocol. We further filtered to those checklists from 2008 onwards, with a duration no more than 5 hours, a distance up to 5 km, and with up to 10 observers. These filters *post-hoc* create a set of more standardised surveys from the larger dataset.

Estimating species encounter rate

We conducted some data preparation and filtering that was specific to the encounter rate model. We first converted all species counts and presences without counts to detections (see supporting information). As we used only complete checklists, all checklists without a record of wood thrush were defined as non-detections (zero-filled). This produced a dataset of detection/non-detection (or 0/1) data for wood thrush.

We filtered the data to reduce the spatial and temporal bias and to address the class imbalance. There are many different ways to conduct spatial filtering and here we outline one approach. We defined an equal area hexagonal grid across the region, with 5 km between the centres of adjacent hexagons, using the R package *dggridR* (Barnes et al. 2017). We randomly selected one detection and one non-detection checklist from each hexagon from each week (Table 1). We randomly split the recombined dataset into 20% for testing and 80% for training. Other spatially explicit divisions of the training and testing data may be more rigorous and more appropriate for some situations (Valavi, Elith, Lahoz-Monfort, & Guillera-Arroita, 2018). In supporting information A2 we provide further details of addressing class imbalance with eBird data.

We next related the detection/non-detection of wood thrush on checklists to environmental covariates. We estimated the *encounter rate* of wood thrush on eBird checklists accounting for variation in detectability with effort covariates. We fitted a random forest model with detection/non-detection of wood thrush as the response variable. As environmental covariates we used land cover data derived from MODIS product MCD12Q1 v006 (Friedl & Sulla-Menashe, 2015). We estimated the land cover associated with each checklist as the proportion of each land cover category in a 2.5 km x 2.5 km square surrounding the checklist location in the year the checklist was conducted. For data from 2017 or 2018 we associated them with 2016 land cover. We included the proportions of each land cover associated with the checklist as separate covariates in the analysis.

We know from experience that variation in the eBird observation process is the most important source of variation in the likelihood of recording a species. We included the following covariates in our analysis: the time observations started, date, duration of observation process, distance travelled, protocol (stationary/travelling), and the number of observers. Additionally, we included a checklist calibration index, which calibrates observers and checklists against others from similar times and places and essentially accounts for variation in observer behaviour, equipment, and skill at detecting species (Kelling et al., 2015; Johnston et al., 2018). In supporting information A2 we provide further details of the EBD variables that can be used to model the observation process.

We used the 80% training dataset and fit the random forest using the R package *ranger* (Wright & Ziegler, 2017). We grew 1,000 classification trees in the random forest analysis and the number of variables from which each tree could select each split was four (James, Witten, Hastie, & Tibshirani, 2013). In order to calibrate the results (Pearce & Ferrier, 2000) we first predicted encounter rate for each checklist in the

80% training set using the random forest model. We then fitted a binomial Generalized Additive Model (GAM) with the real observations as the response and the predicted encounter rate as the predictor variable. The predictor variable was fitted with a smooth with four degrees of freedom and the shape was constrained to be monotonically increasing with the R package *scam* (Pya, 2013). We validated the fitted model and calibration model with the 20% test dataset. We used a range of performance metrics to compare the estimates to the observations: sensitivity, specificity, AUC, Kappa, and mean squared error (Brier score).

We estimated the encounter rate across the whole region by predicting to the whole of the BCR27. We produced a dataset with the land cover for each 2.5 km x 2.5 km grid cell across the entire region and we set effort variables that were constant across the region. The predictions relate to the hypothetical encounter rate of a single expert eBird participant conducting a 1 hour, 1 km complete checklist on 15 June 2016 at the optimal time of day for species detection. We used the random forest and the calibration GAM to estimate encounter rate for this standardised checklist in each grid cell in BCR27.

Estimating species occupancy

To explore an alternative method of estimating species distributions, we applied single-species occupancy models to estimate occupancy and detection probability of wood thrush. There are many complexities and decisions when using citizen science data for occupancy models and we describe these in detail in the supporting information A3 and A4 with only a brief overview here. We defined a sampling replicate as the same observer, visiting the same location, in a given year, in the month of June. We selected combinations of these ‘sites’ with at least two repeated visits. Where there were more than 10 visits, we randomly selected 10 of these. We then spatially subsampled the data, retaining only a single randomly-chosen ‘site’ (i.e. a set of observations from a single location, observer and calendar year) within each 5 km hexagonal grid cell defined above.

We modeled occupancy probability as a function of MODIS land cover categories (Friedl & Sulla-Menashe, 2015) and selected four categories considered *a priori* to have the most ecological relevance: deciduous broadleaf forest, mixed forest, croplands, and urban. For modeling detection probability, we used the six effort covariates described in the ‘encounter rate model’ above. We used the R package *unmarked* to fit single-season models (Fiske & Chandler, 2011). Predictions were AIC-based model averaged values from the set of models that contained all possible combinations of predictor variables. We predicted occupancy of the species to each 2.5 km x 2.5 km grid cell in BCR27. Further modeling details and R code can be found in the supporting information.

Estimating species relative abundance

The third model we ran estimated relative abundance, and accounted for variation in detectability, but did not estimate detection probability. Therefore the estimated abundance will approximate the relative abundance encountered on a checklist. We conducted the same spatial subsampling with detections and non-detections described above for the ‘encounter rate’ model. We then fitted GAM models with the count as the response using R package *mgcv* (Wood, 2017), which enable non-parametric relationships between predictors and response. We tested three different distributions for the response: zero-inflated Poisson, negative binomial, and Tweedie. In all models we included as covariates the four selected land cover variables described above in the ‘occupancy model’ section and the six effort covariates. All continuous covariates were fitted with a thin plate spline with four degrees of freedom. The ‘time observations started’ covariate was fitted with a cyclic cubic spline with six degrees of freedom.

We selected the negative binomial model based on an assessment of model fit with the test data and then we made predictions with this model to the whole of the BCR27 region. We set the checklist covariates for standardised checklists as described above in the ‘encounter rate’ model. We used the estimated smooth for time of day to select the time of day with the highest estimated abundance on checklists (based on the lower confidence interval to

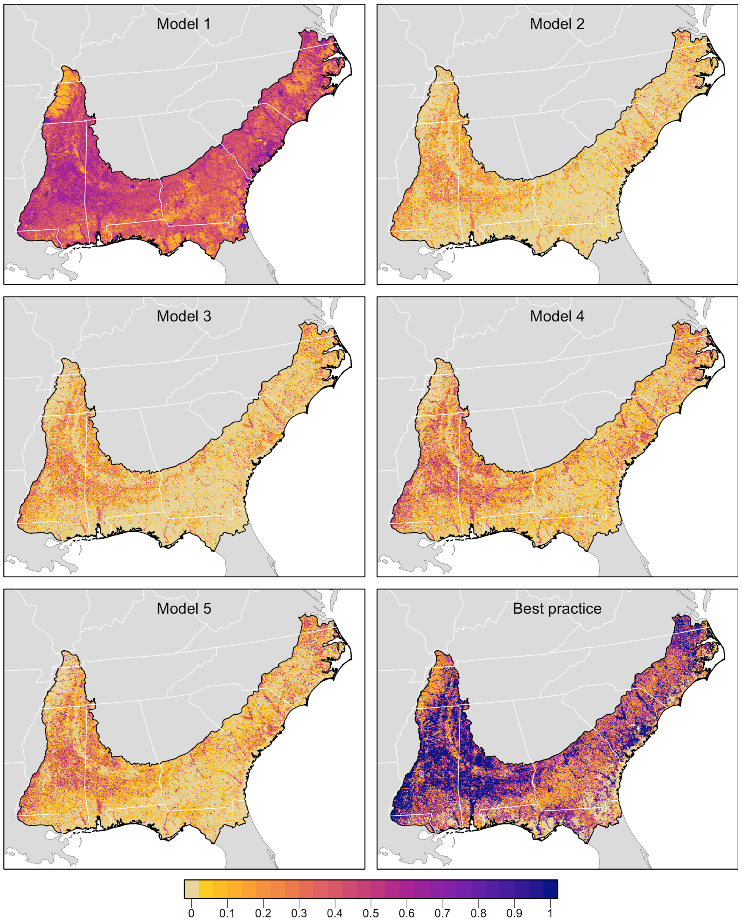

Figure 2. Estimated wood thrush encounter rate across the BCR27 region for models 1 - 6. Estimated encounter rate is the expected proportion of standardised checklists that would record Wood Thrush. These hypothetical standardised checklists are conducted by an expert eBirder, travelling 1km over 1 hour, at the optimal time of day for detecting Wood Thrush.

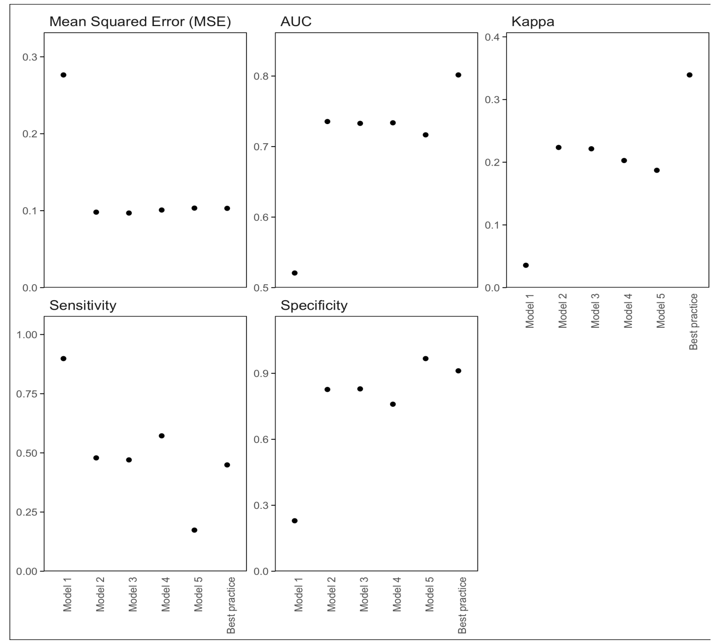

Figure 3. Predictive performance metrics for the encounter rate models 1-6. Model 1 is the Maxent model which uses only presences and produces background psuedo-absences. Model 6 is the random forest encounter rate model using complete checklists, spatial subsampling, effort variable filters, and effort variables as covariates. See Table 1 for further details of models 1-6.

account for uncertainty in the smooth). Predictions across BCR27 therefore estimated the expected number of wood thrush individuals recorded on an eBird checklist by an expert observer focussed on bird watching, travelling 1 km over 1 hour on 15 June, at the time of day when most wood thrush are recorded.

Assessing the impact of best practice guidelines

In order to understand the impact of not following the practices we have outlined above, for each of the three modelling approaches described above we created a set of five additional models with a variety of deficiencies (Table 1). Within each of the three modelling approaches, we compared the model performance and the predictions from the deficient models to those produced from the best practice analysis. For the comparison of model performance metrics we used a fixed test dataset from the best practice model. Models 2-5 used a systematically impaired set of data filters and covariates (Table 1). Model 1, which was only run for the 'encounter rate' and 'relative abundance' approaches, used only checklists with detections. For this model, we produced 10000 random background points across BCR27. For the encounter rate approach, model 1 was fitted with Maxent from the R package *maxnet* (Phillips, 2016). For the relative abundance model we used the background points as zero counts. We provide all R code used to run these models in the supplementary material.

Assessing the impact of varying sample size

For this example we selected a region with relatively high densities of eBird sampling, but the impact of best practice guidelines may vary with sample size. Therefore, we estimated wood thrush encounter rate using model 2 and model 6 for a range of sample sizes. We ran the two models for datasets that were 10-100% (at 10% intervals) of the original dataset size. For each dataset size (e.g., 10%) we produced

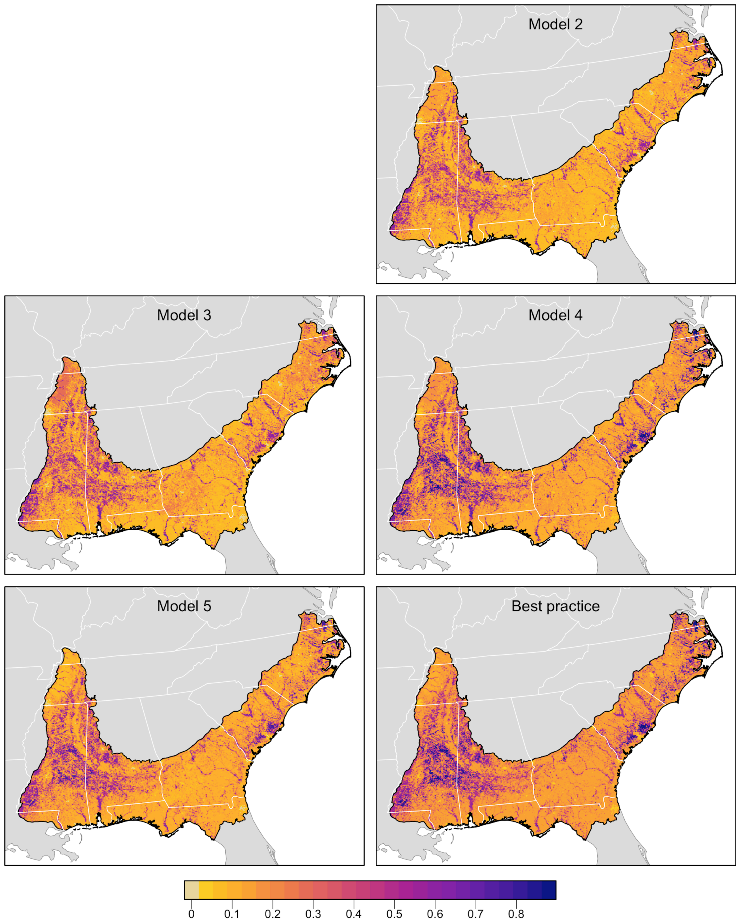

Figure 4. Estimated occupancy of wood thrush across the BCR27 region for occupancy models calculated with data processing steps 2 - 6. The occupancy is the expected probability that cells are occupied by Wood Thrush.

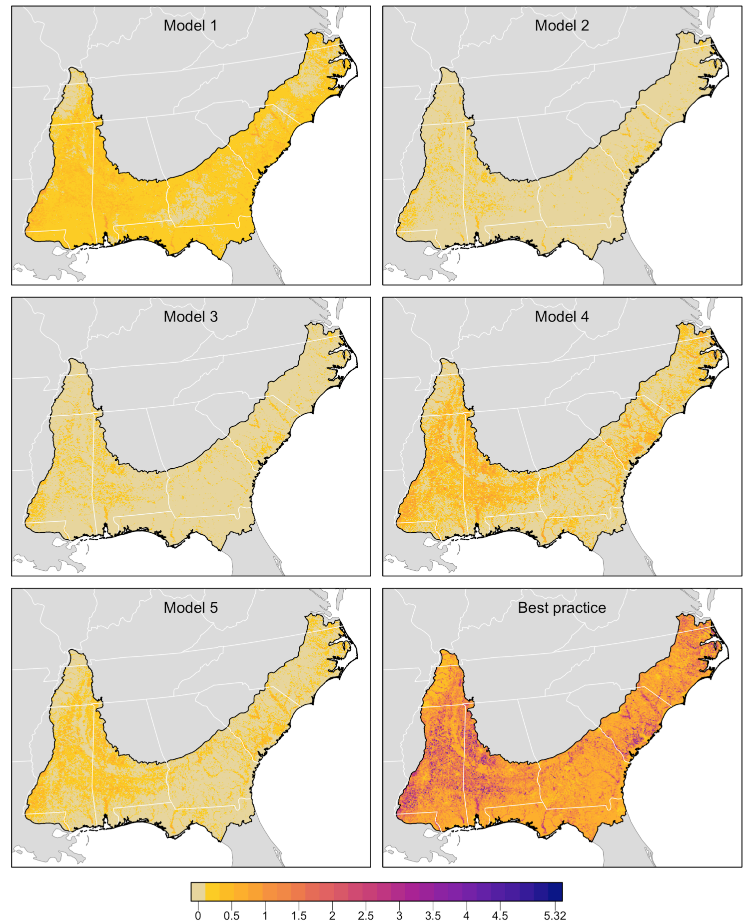

Figure 5. Estimated relative abundance for wood thrush across the BCR27 region for models 1 - 6. Estimated relative abundance is the expected count of Wood Thrush on standardised checklists. These hypothetical standardised checklists are conducted by an expert eBirder, travelling 1km over 1 hour, at the optimal time of day for detecting Wood Thrush.

20 different random subsamples of the original data. We compared the model performance on the fixed test dataset using model 2 and model 6 for 20 subsamples of the 10 different dataset sizes (10%, 20%, … 100%).

Results

Estimating species encounter rate

Estimated predictions of encounter rate varied considerably across the six models. The predictions from model 1 (the Maxent model) are not calibrated and should be treated as relative encounter rate. The models without spatial subsampling (models 2 and 3) had lower estimated encounter rates. Model 6 had the highest estimated encounter rates (Figure 2), likely because predicted encounter rates were for an expert observer at the optimal time of day for species detectability. However differences in the estimated encounter rate may mask similarities in spatial patterns, so it is important to compare the spatial patterns. Models 2-3 all had false negatives, but few false positives (Figure S1), but this effect was mitigated in models 4 and 5.

Model 1 was notably different in model performance from the others. For the wood thrush in BCR27, model 1 had higher mean squared error (MSE), much lower AUC, high sensitivity, but very low specificity (Figure 3). This indicates that it is a model with fewer false negatives, but many false positives. The other five random forest models had relatively similar performance on the test data, with the largest difference notable with the final addition of effort covariates (Figure 3). The best practice model

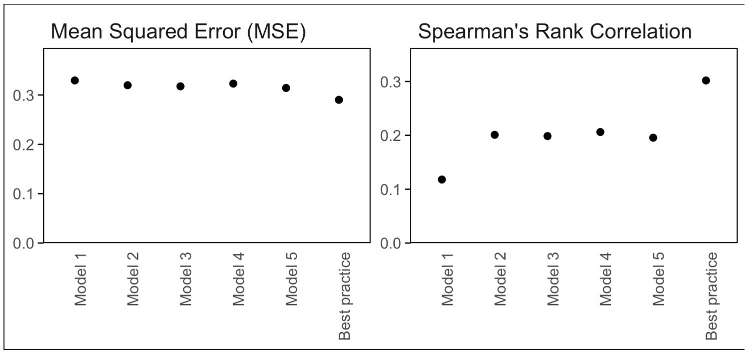

Figure 6. Predictive performance metrics for the relative abundance models 1-6. Model 1 uses only presences and produces background psuedo-absences. The best practice model is a GAM abundance model using complete checklists, spatial subsampling, effort variable filters, and effort variables as covariates. See Table 1 for further details of models 1-6.

(model 6) had the highest AUC and specificity, but slightly lower sensitivity than the other models.

Estimating species occupancy

The spatial patterns in estimated species occupancy were relatively consistent across the different models (Figure 4). However, model 6 still showed areas with the highest estimated occupancy. Therefore species occupancy models were not as negatively impacted by failing to follow our recommended best practices.

Estimating species relative abundance

As with the estimated encounter rate, the addition of effort covariates allowed for estimated predictions that were considerably higher than models 1-5 (Figure 5). Areas of highest estimated abundance appear similar across all models (Figure 5). Models 2-5 were fairly consistent in model performance. Model 1 had notably worse model performance as measured by mean squared error and Spearman's rank correlation and model 6 had notably better model performance (Figure 6).

Assessing the impact of varying sample size

Model performance was better for larger sample sizes across most performance metrics. The improvements gained from using model 6 appeared similar for small and large sample sizes (Figure 7). Therefore these best practice guidelines are appropriate for both data dense and data sparse situations.

Discussion

Citizen science data sets are becoming increasingly valuable research tools due to their increasing prevalence (Pocock et al., 2017) and broad spatiotemporal scope (Chandler et al., 2017). However, citizen science data generally have more errors, assumptions, and biases associated with them, often a result of limited survey design and a highly heterogeneous observation process. Here we demonstrate how thoughtful combinations of data filtering and analysis can leverage the power of citizen science data and help inform ecology and conservation.

Presence-only data are limited in their ability to produce robust ecological inference (Guillera-Arroita et al., 2015). In line with previous findings, our analysis provides strong evidence for including *both* detections and non-detections. The simple approach of generating spatially random pseudo-absences substantially underperformed when estimating encounter rate and relative abundance (Figures 3 and 6). There are multiple approaches to inferring 'completeness' with presence-only data (Hill, 2012; van Strien, van Swaay, & Termaat, 2013); however, where complete checklists are available, it is important these are not degraded to presence-only data without effort covariates.

Previous studies have found that including information on the observation process leads to more accurate and robust results (Isaac, Van Strien, August, de Zeeuw, & Roy, 2014; Johnston et al., 2018). We observed a vast improvement in performance when information on variation in effort and detectability was used for filtering, and especially within models. We found these improvements occurred across all three modeling frameworks, although occupancy models seemed most robust to data and model deficiencies (van Strien et al., 2013).

Our suggestions for best practices are relevant for a range of citizen science datasets and target

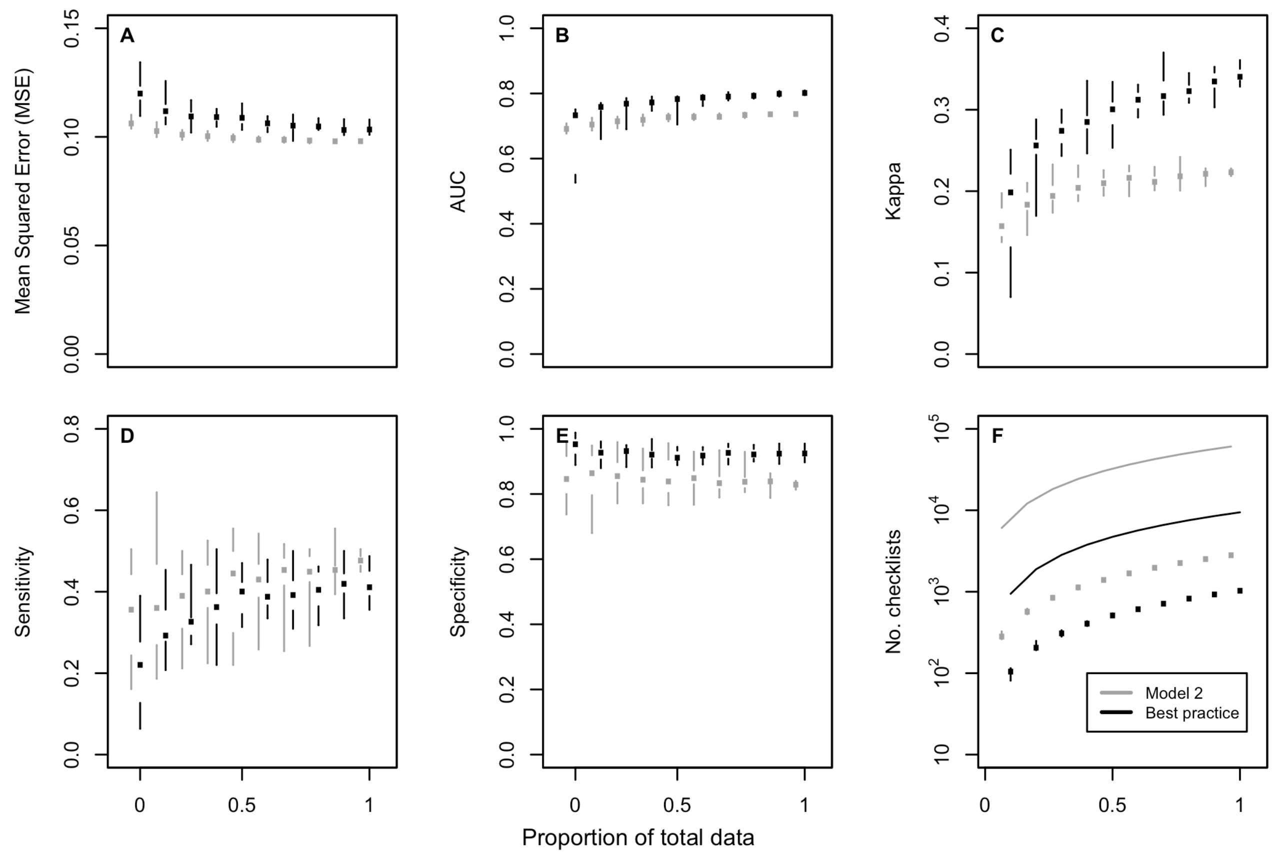

Figure 6. Effect of sample size on the model performance of encounter rate models for wood thrush in BCR 27. The test data was was a fixed test dataset across all models, based on the data processing for Model 6. Trials were run for proportions of 0.1, 0.2, ...1 of the full dataset. Both model 2 and model 6 had the same proportions, but they are shown jittered for visual clarity. Panels show A mean squared error; B AUC; C Kappa; D sensitivity; E specificity; and F the number of checklists with wood thrush observations and (in lines) the total number of checklists (which did not vary across the 20 iterations).

ecological metrics. These practices improved model performance under both data dense and data sparse conditions (Figure 7), even though the volume of data analysed decreased. These results are therefore relevant for data-poor biodiverse regions where information on species distribution is critical and often lacking (Figure 1).

The best practices we propose are most relevant to citizen science projects designed to collect a large quantity of data, with important information describing the observation process (Kelling et al., 2018). There are numerous citizen science programs in the world today, but only a limited number of them collect even the information needed to infer absences (Pocock et al., 2017). The case study using eBird data provides additional evidence that at least for this taxonomic group, this information can be collected without decreasing participation.

Although we only focused on modeling species distribution, many other types of ecological inference will also benefit from these best practices. In combination, our best practices for collecting, processing and modeling citizen science data can inform ways to improve existing and future programs, while increasing our current capacity to conduct robust analyses using growing volumes of citizen science data.

Acknowledgements

We thank the thousands of participants in the eBird project. We are grateful to Frank La Sorte and Marshall Iliff for discussions and comments that improved this manuscript. Funding for this work came from the Wolf Creek Foundation, The Leon Levy Foundation, The Packard Foundation, and the National Science Foundation (grants ITR-0427914, DBI-0542868, IIS-0612031, ABI-1356308, and CCF-1522054).

Author contributions

All authors conceived the ideas and designed methodology; AJ, WMH, VRG, and OR analysed the data; MES led the writing of the best practices bookdown document in supporting information; AJ, WMH, VRG, ETM, and OR wrote the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

Data accessibility

In Supporting Information Appendix A5 we provide the subset of data used in this paper with all the code used for the analyses presented here. We will provide a DOI for the github repository A5 on publication.

References

Altwegg, R., & Nichols, J. D. (2019). Occupancy models for citizen-science data. *Methods in Ecology and Evolution / British Ecological Society*, *10*(1), 8–21.

Bas, Y., Devictor, V., Moussus, J.-P., & Jiguet, F. (2008). Accounting for weather and time-of-day parameters when analysing count data from monitoring programs. *Biodiversity and Conservation*, *17*(14), 3403–3416.

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.

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. *Biodiversity and Conservation*, *20*(1), 119–139.

Casanovas, P., Lynch, H. J., & Fagan, W. F. (2014). Using citizen science to estimate lichen diversity. *Biological Conservation*, *171*, 1–8.

Chandler, M., See, L., Copas, K., Bonde, A. M. Z., López, B. C., Danielsen, F., … Masinde, S. (2017). Contribution of citizen science towards international biodiversity monitoring. *Biological Conservation*, *In press*.

Cohn, J. P. (2008). Citizen Science: Can Volunteers Do Real Research? *Bioscience*, *58*(3), 192–197.

Dennis, R. L. H., & Thomas, C. D. (2000). Bias in Butterfly Distribution Maps: The Influence of Hot Spots and Recorder's Home Range. *Journal of Insect Conservation*, *4*(2), 73–77.

Ellis, M. V., & Taylor, J. E. (2018). Effects of weather, time of day, and survey effort on estimates of species richness in temperate woodlands. *Emu - Austral Ornithology*, *118*(2), 183–192.

Fiske, I., & Chandler, R. (2011). unmarked: An R package for fitting hierarchical models of wildlife occurrence and abundance. *Journal of Statistical Software*, *43*(10), 1–23.

Friedl, M., & Sulla-Menashe, D. (2015). MCD12Q1 MODIS/Terra+ Aqua Land Cover Type Yearly L3 Global 500m SIN Grid V006 [Data set]. *NASA EOSDIS Land Processes DAAC. Doi*, *10*.

Guillera-Arroita, G., Lahoz-Monfort, J. J., Elith, J., Gordon, A., Kuiala, H., Lentini, P. E., … Wintle, B. A. (2015). Is my species distribution model fit for purpose? Matching data and models to applications. *Global Ecology and Biogeography: A Journal of Macroecology*, *24*(3), 276–292.

Hazzi, N. A., Moreno, J. S., Ortiz-Movliav, C., & Palacio, R. D. (2018). Biogeographic regions and events of isolation and diversification of the endemic biota of the tropical Andes. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(31), 7985–7990.

Hijmans, R. J., Garrett, K. A., Huaman, Z., Zhang, D. P., Schreuder, M., & Bonierbale, M. (2000). Assessing the geographic representativeness of genebank collections: the case of Bolivian wild potatoes. *Conservation Biology: The Journal of the Society for Conservation Biology*, *14*(6), 1755–1765.

Hill, M. O. (2012). Local frequency as a key to interpreting species occurrence data when recording effort is not known. *Methods in Ecology and Evolution / British Ecological Society*, *3*(1), 195–205.

Howard, E., Aschen, H., & Davis, A. K. (2010). Citizen Science Observations of Monarch Butterfly Overwintering in the Southern United States. *Psyche; a Journal of Entomology*, *2010*. doi:10.1155/2010/689301

Isaac, N. J. B., Van Strien, A. J., August, T. A., de Zeeuw, M. P., & Roy, D. B. (2014). Statistics for citizen science: extracting signals of change from noisy ecological data. *Methods in Ecology and Evolution / British Ecological Society*, *5*(10), 1052–1060.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*.

Johnston, A., Fink, D., Hochachka, W. M., & Kelling, S. (2018). Estimates of observer expertise improve species distributions from citizen science data. *Methods in Ecology and Evolution / British Ecological Society*, *9*, 88–97.

Kadmon, R., Farber, O., & Danin, A. (2004). EFFECT OF ROADSIDE BIAS ON THE ACCURACY OF PREDICTIVE MAPS PRODUCED BY BIOCLIMATIC MODELS. *Ecological Applications: A Publication of the Ecological Society of America*, *14*(2), 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. *Diversity and Distributions*, *22*(10), 1024–1035.

Kelling, S., Johnston, A., Fink, D., Ruiz-Gutierrez, V., Bonney, R., Bonn, A., … Guralnick, R. (2018). Finding the signal in the noise of Citizen Science Observations. *bioRxiv*. Retrieved from https://www.biorxiv.org/content/early/2018/05/18/326314.abstract

Kelling, S., Johnston, A., Hochachka, W. M., Iliff, M., Fink, D., Gerbracht, J., … Yu, J. (2015). Can Observation Skills of Citizen Scientists Be Estimated Using Species

Accumulation Curves? *PloS One*, *10*(10), e0139600.

Kéry, M., Gardner, B., & Monnerat, C. (2010). Predicting species distributions from checklist data using site-occupancy models. *Journal of Biogeography*, *37*(10), 1851–1862.

Lang, S. D. J., Mann, R. P., & Farine, D. R. (2018). Temporal activity patterns of predators and prey across broad geographic scales. *Behavioral Ecology: Official Journal of the International Society for Behavioral Ecology*. doi:10.1093/beheco/ary133

La Sorte, F. A., Lepczyk, C. A., Burnett, J. L., Hurlbert, A. H., Tingley, M. W., & Zuckerberg, B. (2018). Opportunities and challenges for big data ornithology. *The Condor*, *120*(2), 414–426.

Lehikoinen, A. (2013). Climate change, phenology and species detectability in a monitoring scheme. *Population Ecology*, *55*(2), 315–323.

MacPherson, M. P., Jahn, A. E., Murphy, M. T., Kim, D. H., Cueto, V. R., Tuero, D. T., & Hill, E. D. (2018). Follow the rain? Environmental drivers of Tyrannus migration across the New World. *The Auk*, 881–894.

Mair, L., & Ruete, A. (2016). Explaining Spatial Variation in the Recording Effort of Citizen Science Data across Multiple Taxa. *PloS One*, *11*(1), e0147796.

Marques, T. A., Thomas, L., Fancy, S. G., & Buckland, S. T. (2007). Improving estimates of bird density using multiple-covariate distance sampling. *The Auk*, *124*(4), 1229–1243.

Mattsson, B. J., Dubovsky, J. A., Thogmartin, W. E., Bagstad, K. J., Goldstein, J. H., Loomis, J. B., … López-Hoffman, L. (2018). Recreation economics to inform migratory species conservation: Case study of the northern pintail. *Journal of Environmental Management*, *206*, 971–979.

Mayor, S. J., Guralnick, R. P., Tingley, M. W., Otegui, J., Withey, J. C., Elmendorf, S. C., … Schneider, D. C. (2017). Increasing phenological asynchrony between spring green-up and arrival of migratory birds. *Scientific Reports*, *7*(1), 1902.

Miller, D. A. W., Pacifici, K., Sanderlin, J. S., & Reich, B. J. (2019). The recent past and promising future for data integration methods to estimate species' distributions. *Methods in Ecology and Evolution / British Ecological Society*, *10*(1), 22–37.

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. *Biological Conservation*, *191*, 38–49.

Oliveira, C. V., Olmos, F., dos Santos-Filho, M., & Bernardo, C. S. S. (2018). Observation of Diurnal Soaring Raptors In Northeastern Brazil Depends On Weather Conditions and Time of Day. *The Journal of Raptor Research*, *52*(1), 56–65.

Pacifici, K., Reich, B. J., Miller, D. A. W., Gardner, B., Stauffer, G., Singh, S., … Collazo, J. A. (2017). Integrating multiple data sources in species distribution modeling: a framework for data fusion. *Ecology*, *98*(3), 840–850.

Pearce, J., & Ferrier, S. (2000). Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling*, *133*(3), 225–245.

Phillips, S. (2016). Maxnet: Fitting 'maxent'species distribution models with 'glmnet'.

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*(4), e0172579.

Pya, N. (2013). *scam: Shape constrained additive models*.

R Core Team. (2018). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Sauer, J. R., Pardieck, K. L., Ziolkowski, D. J., Smith, A. C., Hudson, M.-A. R., Rodriguez, V., … Link, W. A. (2017). The first 50 years of the North American Breeding Bird Survey. *The Condor*, *119*(3), 576–593.

Seeholzer, G. F., Claramunt, S., & Brumfield, R. T. (2017). Niche evolution and diversification in a Neotropical radiation of birds (Aves: Furnariidae). *Evolution; International Journal of Organic Evolution*, *71*(3), 702–715.

Strimas-Mackey, M., Miller, E., & Hochachka, W. (2017). auk: eBird Data Extraction and Processing with AWK. *R Package Version*.

Sullivan, B. L., Aycrigg, J. L., Barry, J. H., Bonney, R. E., Bruns, N. E., Cooper, C. B., … Kelling, S. (2014). The eBird enterprise: An integrated approach to development and application of citizen science. *Biological Conservation*, *169*, 31–40.

Troudet, J., Grandcolas, P., Blin, A., Vignes-Lebbe, R., & Legendre, F. (2017). Taxonomic bias in biodiversity data and societal preferences. *Scientific Reports*, *7*(1), 9132.

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

Tulloch, A. I. T., & Szabo, J. K. (2012). A behavioural ecology approach to understand volunteer surveying for citizen science datasets. *Emu - Austral Ornithology*, *112*(4), 313–325.

Valavi, R., Elith, J., Lahoz-Monfort, J. J., & Guillera-Arroita, G. (2018). block CV: An r package for generating spatially or environmentally separated folds for k -fold cross-validation of species distribution models. *Methods in Ecology and Evolution / British Ecological Society*, *67*, 617.

van Strien, A. J., van Swaay, C. A. M., & Termaat, T. (2013). Opportunistic citizen science data of animal species produce reliable estimates of distribution trends if analysed with occupancy models. *The Journal of Applied Ecology*, *50*, 1450–1458.

Vianna, G. M. S., Meekan, M. G., Bornovski, T. H., & Meeuwig, J. J. (2014). Acoustic telemetry validates a citizen science approach for monitoring sharks on coral reefs. *PloS One*, *9*(4), e95565.

Wiggins, A., & Crowston, K. (2011). From Conservation to Crowdsourcing: A Typology of Citizen Science. In *2011 44th Hawaii International Conference on System Sciences* (pp. 1–10).

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R, Second Edition*. CRC Press.

Wright, M., & Ziegler, A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software, Articles*, *77*(1), 1–17.

SUPPORTING INFORMATION

Appendix A1: Template for describing eBird data used in analyses

Appendix A2: eBird data description

Appendix A3: Fitting species distribution and abundance models with eBird data

Appendix A4: Best practice code and descriptions of models

http://strimas.com/ebird-best-practices/

Appendix A5: Code for the analyses in this paper

https://github.com/mstrimas/ebp-paper

Table S1: Descriptions of eBird dataset variables

Figure S1: Comparison of predicted wood thrush encounter rates