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

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

SUPPORTING INFORMATION APPENDIX A3

Fitting species distribution and abundance models with eBird data

In this section we describe some of the more complex aspects of fitting the species distribution and relative abundance models with eBird data; class imbalance, model validation, and defining the repeated visits for occupancy models. The R code used for data processing and model fitting is available in supporting information A4 (best practice bookdown document with descriptions of the models) and supporting information A5 (code used to produce the figures in this paper, including the deficient versions of models). This appendix is designed to accompany the main paper, so does not repeat information in the main manuscript and is not a stand-alone document.

Class imbalance

Class imbalance in citizen science data can be substantial, with a high proportion of checklists not reporting a species. This is particularly marked for species that are difficult to detect. Thus, looking for, and ameliorating the effects of class imbalance, should be standard practice when preparing to fit models to eBird data. The steps in this process are firstly to determine the ratio of detection checklists to non-detection checklists for your species of interest. The general rule of thumb used for eBird data is that a data set with fewer than 25% of its checklists containing a detection for the species of interest should be treated as an imbalanced data set.

There are many methods for dealing with unbalanced data. When using a random forest (as we do in the 'encounter probability' model), it is possible to use balanced random forest, where the model draws its bootstrap samples from each class separately, drawing first from the minority class, then drawing an equal number from the majority class (Chen et al. 2004). Another method is weighted random forest, where the model assigns a harsher penalty to misclassification of the minority class than it does to the majority class. It is also possible to address class imbalance via a sampling routine before the modelling. It is possible to use oversampling (by creating duplicated examples of the minority class), or undersampling (by randomly removing observations from the majority class) to create a more balanced dataset. Both techniques are often used together. One specific oversampling technique, synthetic minority oversampling technique (SMOTE) creates synthetic examples from the minority class. This creates synthetic examples that occupy the parameter space between randomly chosen observations and their nearest neighbors so that the added observations are not direct copies (Chawla, Bowyer, Hall, & Kegelmeyer, 2002).

When using eBird data it is also important to consider the spatial bias when over- or undersampling to address class imbalance. Undersampling eBird checklists may exacerbate the spatial bias in the dataset. However, spatial sampling can be combined with undersampling, and this method was shown to be very useful for eBird data (Robinson, Ruiz-Gutierrez, & Fink, 2018). This method targets a particular species and splits the data into two subsets; 1) checklists with species detections; and 2) checklists with non-detections for a given species. The non-detection dataset 2 is then spatially filtered to address the spatial bias (see Supporting Information A2), and the filtered dataset is combined with the complete detection dataset 1. The result is a spatially undersampled data set where the degree of class imbalance is mitigated. This method works well for species with a very low proportion of detections, as the spatial bias maintained by not spatially filtering the detection observations is minimal. It is also possible to spatially filter both datasets before recombining the

filtered data sets. This method will likely work best for species where the class imbalance is less drastic as the example in Robinson et al. (2018). This approach of spatially subsampling both the detections and non-detections was the method used in this study.

Model Validation

In order to assess the impacts of different practices on the accuracy of encounter-rate and relative-abundance models we used cross validation. We fitted our models the 80% training data, and then validated the models with the remaining 20% test data. We will describe the rationale for two aspects of our cross-validation procedure below.

Firstly, for our assessments we wanted to evaluate models based on their ability to produce accurate predictions across the entire spatial extent of our study region. For this purpose, it is appropriate to create a spatially balanced set of test data that evenly represents all study region. We approximated this type of even spatial representation by randomly selecting our test data after filtering the input data in order to correct for spatial bias. Other processes for creating the test data set would be appropriate if our objectives were different (Valavi, Elith, Lahoz-Monfort, & Guillera-Arroita, 2018). Given the uneven spatial distribution of eBird data, we recommends that a purely random selection of data for the test set in cross validation would be unwise.

Secondly, we do not have a single criterion (e.g., accurate prediction of non-zero values, accurate prediction of rank order of observed values) with which we want to judge the predictive performance of models. Instead, we want to assess the all-round predictive performance of models; i.e. we want to assess whether the extent to which any and all types of weaknesses exist. Given this objective, we have used multiple performance metrics, each of which emphasizes a different facet of model accuracy. We used a set of performance metrics available in the R packages *PresenceAbsence* (Freeman & Moisen, 2008) and *verification* (NCAR Research Applications Laboratory, 2015).

Occupancy models, due to their nature as two linked logistic regressions (describing detection and occupancy processes), are problematic for cross validation, and no generally accepted method currently exists. As such, we have not assessed the predictive performance of occupancy models in this paper.

Estimating species occupancy

The initial stages of preparing data for occupancy modelling were the same as those described for modelling of encounter rates in the main paper. The data from each checklist were converted into binary detection/non-detection data during the process of creating zero-filled data, using the R package *auk*.

The next step in formatting data was to group our data into series of repeated count, with each series representing the data from a single ‘site’ that was repeatedly visited. We defined each site as being a combination of geographic point and observer (locality_id and observer_id; see Table A1). One other option would have been to allow multiple observers’ data from the same geographic point to be combined into a single time set; however, in most cases all data from a locality still would have been from the same observer, so we opted for the most conservative approach of defining sites of combinations of geographic point and observer. The inclusion of the checklist_calibration_index (Table A1) as a site-level predictor in our models enabled much of the observer-linked variation in detection rates to be statistically removed, so that effectively the remaining site-level variation represented variation in the environments among geographic points. A second alternative definition for ‘site’ would have been to create a spatial grid across study region, and treat all checklists within a grid cell as replicate counts. However, this space-for-time substitution is not universally appropriate (Kendall & White, 2009), and thus we chose not to have to deal with this added interpretational complexity.

The definition of which checklists can be used to form series of repeated visits to a site requires an explicit definition of a period of closure during which

occurrence at a site does not change (i.e. the species is either always present or always absent). Prior knowledge indicates that the month of June represents a period of time after spring migration, and during the period of time during which Wood Thrushes are nesting. Thus, we defined the period of closure to be the entire month of June.

A final aspect of defining the time period of closure is the decision of how to deal with the existence of multiple years' observations within our data. We treated each calendar year as a separate temporal entity, because we could not assume that sites occupied in one calendar year would also be occupied in other years. Within each site and calendar year we only retained data if a site-year combination had data from at least two repeated visits. Where there were more than 10 visits, we randomly selected 10 of these. We also allowed separate years' data from the same site to appear in the data set (a process sometimes referred to as "stacking"). The R package *auk* was used to convert the input data from their original format into a form in which each series of observations from a site and year are treated as a separate row of data, using the *auk* function *format_unmarked_occu*.

Next we spatially subsampled the data that had been formatted for occupancy modelling as described in the previous paragraph, in order to create a final data set in which the density of data was distributed relatively evenly across the study region. We retained 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, with grid cells created as described in Estimating species encounter rate.

References

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Freeman, E. A., & Moisen, G. (2008). PresenceAbsence: An R package for presence-absence model analysis. *Journal of Statistical Software*, *23*(11), 1–31.

Kendall, W. L., & White, G. C. (2009). A cautionary note on substituting spatial subunits for repeated temporal sampling in studies of site occupancy. *The Journal of Applied Ecology*, *139*, 657.

NCAR Research Applications Laboratory. (2015). verification: Weather Forecast Verification Utilities. Retrieved from https://CRAN.R-project.org/package=verification

Robinson, O. J., Ruiz-Gutierrez, V., & Fink, D. (2018). Correcting for bias in distribution modelling for rare species using citizen science data. *Diversity & Distributions*, *24*(4), 460–472.

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.