

Supporting Information. Daniel Fink, Tom Auer, Alison Johnston, Viviana Ruiz-Gutierrez, Wesley M. Hochachka, and Steve Kelling. 2020. Modeling avian full annual cycle distribution and population trends with citizen science data. *Ecological Applications*.

Appendix S2: Spatiotemporal Sampling

Within each stixel, a spatial case-control sampling strategy was used to address the challenges of highly imbalanced data and site selection bias. Imbalanced data arise when there are a very small number of species detections and a very large number of non-detections. This is a modeling concern because binary regression methods, like the first component of the ZI-BRT model, become overwhelmed by the non-detections and perform poorly (King & Zeng 2001, Robinson *et al.* 2017). The low detection rates of many species, especially along range boundaries, can result in highly imbalanced training data. This makes data imbalance a defining challenge for broad-scale, year-round modeling. By sampling detection and non-detection cases separately, case-control sampling (e.g. Breslow 1996, Fithian & Hastie 2014) improves data balance and model performance. Additionally, to alleviate spatial biases caused by the eBird site selection process, spatiotemporally balanced samples were drawn as part of the case-control sampling.

To generate spatially and temporally balanced samples for the case-control sampling, we drew data from a randomly located regular grid, with one checklist randomly selected per 10km × 10km x 1week grid cell, applied separately for each year and separately for detection and non-detection cases. The 10km spatial grid dimension was selected to reduce the impact of repeated checklists from popular sites.

Additionally, detection data were over-sampled, using the same spatiotemporally balanced procedure, when they represented less than 25% of the balanced data. Oversampling generates ties, or repeated observations, among the training data. Because boosting, used in the ZI-BRT base models, is driven more by the set of distinct data points than the number of tied data points, Mease et al. (2007) suggested breaking ties to force the boosting algorithm to respond to oversampled data. To do this we mimic the effects of imprecisely recorded checklist locations, and jitter all the spatial covariate values for each of the replicated oversampled checklists.

Because over-sampling detections changes the fraction of detections in training data, it biases upwards the average detection rate estimated by the occurrence model. To correct for this bias, we apply the prior correction discussed in King & Zeng (2001), essentially adjusting the intercept term of the boosted model to match the fraction of detections in the training data before oversampling. Moreover, to insure the quality of the *probabilistic* predictions of from the Bernoulli response BRT we calibrate the predictions using a generalized additive model constrained to be monotonically increasing (Natalya 2018).

For the trend base models, we also balanced the per year sample size, after spatiotemporal case control sampling, to control for the strong inter-annual increases in eBird data volume, 20-30% per year since 2005. First, we computed the average sample size per year from 2007-16, the ten-year period over which trends were estimated. Years with less than the average sample sizes, were over-sampled (i.e. randomly sampled with replacement) and years with more than the average sample size were under-sampled (i.e. randomly sampled without replacement). This sampling strategy, what we call a reverse-mullet, resulted in a training data set with the same

per-year sample size.

Literature Cited

- Breslow, N. E. (1996). Statistics in epidemiology: the case-control study. *Journal of the American Statistical Association*, 91(433), 14-28.
- Fithian, W. and Hastie, T., (2014). Local case-control sampling: Efficient subsampling in imbalanced data sets. *Annals of statistics*, 42(5), 1693.
- King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political analysis*, 9(2), 137-163.
- Mease, D., Wyner, A. and Buja, A., (2007). Cost-weighted boosting with jittering and over/under-sampling: JOUS-boost. *J. Machine Learning Research*, 8, 409-439.
- Natalya Pya (2018). scam: Shape Constrained Additive Models. R package version 1.2-3.
<https://CRAN.R-project.org/package=scam>
- Robinson, O.J., Ruiz-Gutierrez, V., Fink, D. (2017). Correcting for bias in distribution modeling for rare species using citizen science data. *Diversity and Distributions*, 24(4), 460-472. DOI: 10.1111/ddi.12698