Evaluation of Citizen Science data for biodiversity research in Florida

Matthew Boone
Data scientist
@birderboone
birderboone
m.boone@ufl.edu

Matthew Boone[1], Henry Hochmair[2], and Mathieu Basille[1]

[1]Ft Lauderdale REC, Wildlife Ecology and Conservation, University of Florida/IFAS
[2] Ft Lauderdale REC, School of Forest Resources & Conservation, University of Florida/IFAS

UF IFAS UNIVERSITY of FLORIDA

Introduction

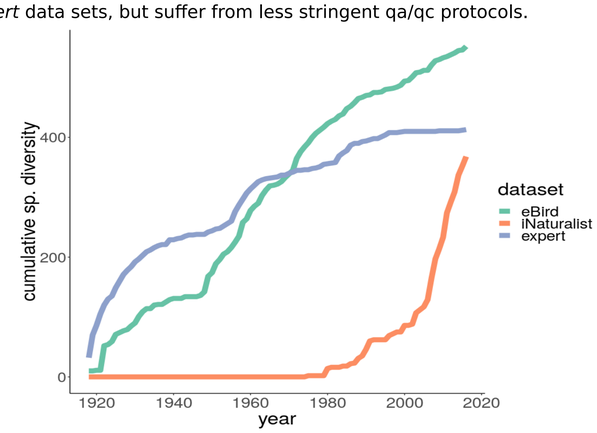
Citizen science data has rapidly expanded in the last 10 years, attributing to over 1 billion records in the Global Information Facility Database (GBIF)[1]. Datasets, like eBird and iNaturalist, vastly increase knowledge of biodiversity, however can suffer from uncertain data quality and uneven sampling effort.[2] Of the remaining data in the GBIF data set, the majority of entries come from musuem collections (termed 'Expert' in this analysis). These specimens are identified and vetted by experts but can suffer from smaller sample sizes and an oppurtunistic collection method.

We took of the three most populous data sources from the GBIF presence data (eBird, iNaturalist, and Museum collections) and assesed their usefulness and biases for species distribution modeling and biodiversity studies in Florida. We then modeled presence of 16 wading birds in Florida using **Boosted Regression Trees** to test the datasets accuracy.[3]

Comparing GBIF datasets

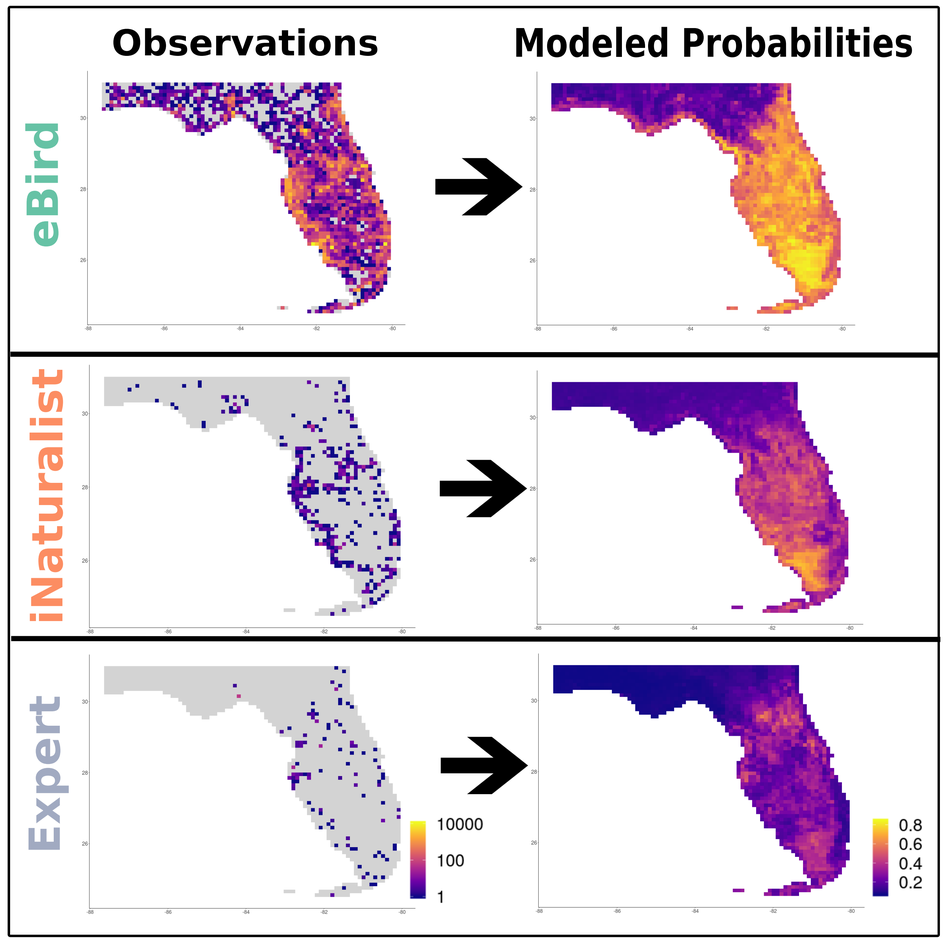
dataset	qa/qc	observations	coverage	legacy data
eBird	moderate	1500000	0.97	moderate
iNaturalist	medium/high	52000	0.67	poor
expert	high	45000	0.52	high

eBird data sets contain over **300x** more data points than both *iNaturalist* and *Expert* data sets, but suffer from less stringent qa/qc protocols.



Cumulative species diversity from data taken from GBIF. Expert data sets contain the largest amount of data prior to 1970, but eBird is likely to surpass expert datasets as more legacy field observations are entered ever year. Despite iNaturalists lack of legacy data, it will soon surpass museum collections in both total avian observations and species diversity.

Predicted occurence of Little Blue Herons (*Egretta caerulea*)

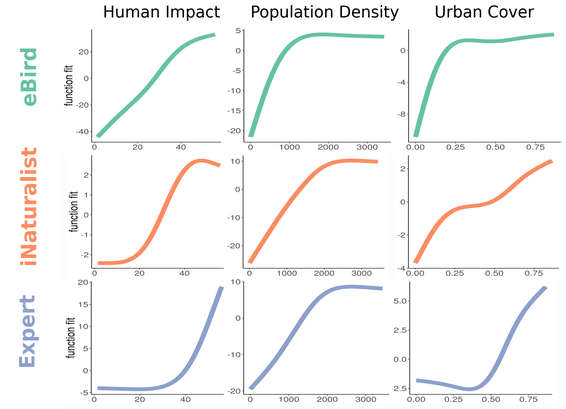


Species distribution models of Little Blue Heron created with Boosted Regression Trees using the dismo package in R and averaged over 50 replicates. Model inputs were controlled for by adjusting the distribution of pseudo-absences to match the sampling variable that affected each data set the most.[4] eBird was modeled by controlling for Human Impact, while iNaturalist and Expert data sets were modeled by controlling for population. Variable inputs included landcover metrics, climate metrics, as well as landscape metrics. Model outputs were qualitatively similar, but differed in their **magnitude** of predictability, likely owing to the sample size and spatial variation of data points.

References

1. GBIF.org (15 Dec 2018) GBIF Occurrence Download https://doi.org/10.15468/dl.yi1jfs
2. Jacobs, C., & Zipf, A. (2017). Completeness of citizen science biodiversity data from a volunteered geographic information perspective. Geo-Spatial Information Science, 20 (1), 3–13. https:// doi.org/10.1080/10095020.2017.1288424
3. Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. Journal of Animal Ecology, 77(4), 802–813. https://doi.org/10.1111/ j.1365-2656.2008.01390.x
4. Bucklin, D. N., Basille, M., Benscoter, A. M., Brandt, L. A., Mazzotti, F. J., Romañach, S. S., ... Watling, J. I. (2015). Comparing species distribution models constructed with different subsets of environmental predictors. Diversity and Distributions, 21(1), 23–35. https://doi.org/10.1111/ddi.12247

Correcting for sampling bias



Relative Influence on model

dataset	Human Impact	Urban Cover	Population Density	Model fit
eBird	50	28	22	0.68
iNaturalist	38	14	48	0.58
expert	18	14	68	0.52

All datasets were influenced to some degree by population density suggesting that oppurtunistic sampling bias was ubiqutous. The eBird was highly influenced by Human Impact. This could be because the mobility of birds assumes a stronger link to site accessibility rather than distance to population centers.

Conclusions

eBird
- Large spatial distribution allowed for finer modeling detail across the range. This allowed more stringent data input limitations for the model, helping to offset noise.
- Sampling bias focused more on access to sites (*human impact* and *roads*) than *population*.

iNaturalists
- Focus on identifiable media creates a more trusted data set, but suffers from much lower data volume than eBird, leading to lower predictive power.
- Sampling bias focused highly on a mix of population and accessibility metrics. Model accuracy was highest by controlling for *population* alone.

Expert
- Legacy data prior to 1970 is of important research value, however modern collection numbers will soon be outpaced in data submissions by iNaturalist.
- Data wass less influenced by sampling bias metrics, and random models were effective in correcting for sampling bias.