BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 2: e25395
doi: 10.3897/biss.2.25395

OPEN ACCESS

Conference Abstract

An Integrated Data Quality System for Species Observations

Steve Kelling ‡

‡ Cornell Lab of Ornithology, Ithaca, United States of America

Corresponding author: Steve Kelling (stk2@cornell.edu)

Received: 31 Mar 2018 | Published: 18 May 2018

Citation: Kelling S (2018) An Integrated Data Quality System for Species Observations. Biodiversity Information Science and Standards 2: e25395. https://doi.org/10.3897/biss.2.25395

Abstract

Species-level observational data comprise the largest and fastest-growing part of the Global Biodiversity Information Facility (GBIF). The largest single contributor of species observations is eBird, which so far has contributed more than 361 million records to GBIF. eBird engages a vast network of human observers (citizen-scientists) to report bird observations, with the goal of estimating the range, abundance, habitat preferences, and trends of bird species at high spatial and temporal resolutions across each species' entire life-cycle. Since its inception, eBird has focused on improving the data quality of its observations, primarily focused in two areas:

1. ensuring that participants describe how they gathered their observations and,
2. all observations are reviewed for accuracy.

In this presentation I will review how this is done in eBird.

Standardized Data Collection. eBird gathers bird observations based on how bird watchers typically observe birds with units of data collection being "checklists" of zero or more species including a count of individuals for each species observed. Participants choose the location where they made their observations and submit their checklists via Mobile Apps (50% of all submissions) or the website (50% of all submissions). All checklists are submitted in a standard format identifying where, how, and with whom they made their observations. Mobile apps precisely record locations, the track taken, and the

distance they traveled while making the observations. The start time and duration of surveys are also recorded. All observers must report whether they reported all the birds they detected and identified, which allows analysts to infer absence of birds if they were not reported. All data are stored within an Oracle data management framework.

Data Accuracy. The most significant data quality challenge for species observations is detecting and correctly identifying organisms to species. The issue involves how to handle both false positives — the misidentification of an observed organism, and false negatives— failing to report a species that was present. The most egregious false positives can be identified as anomalies that fall outside the norm of occurrence for a species at a particular time or space. However, false positives can also be misidentifications of common species. These challenges are addressed by:

- **Data-driven filters.** eBird's existing data can identify and flag potentially erroneous records at increasingly fine spatial, temporal, and user-specific scales. These filters can identify outliers and likely errors, which are the foundation of the eBird review process. By using the vetted data to identify outliers, data quality checks run against expected occurrence probabilities at very fine scales and identify anomalies during data submission (including on mobile devices).
- **Incorporate observer expertise scores.** Observer differences are the largest source of variability in eBird data. Assessment of observer metrics, and the inclusion of these data in species distribution models, improves analysis output and model performance.
- **Expert reviewer network.** More than 2000 volunteers review records identified by the data-driven filters and contact data submitters to confirm their observations. The existing data quality process functions globally. Currently the approach is focused on misidentified birds, but in the future will also involve collection event issues (e.g., issues with protocol, location, or methodology), sensitive species, exotic species, and better handle widely-observed individual rarities. Additional tools are also to be developed to help editors improve efficiency and better prioritize review.

In 2017, 4,107,757 observations representing 4.6% of all eBird records submitted were flagged for review by the data driven filters. Of these records 57.4% were validated and 42.6% were invalidated.

Keywords

Data quality, observational data, eBird, citizen science

Presenting author

Steve Kelling

Funding program

The National Science Foundation (ABI sustaining: DBI-1356308)