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 TABLE S1

Table S1: Variables from the eBird Basic Dataset (EBD) and the *auk* post-processed dataset. Here we list variables that are particularly relevant to scientific analysis of eBird. Variables are listed in groups based on functional categories of the information that they provide. We describe the variables and also note which of the challenges in the main paper are assisted by the information in each variable. Information about all EBD variables is provided on the eBird website.

Variable name in EBD	Variable name after *auk* processing	Description
Basic checklist and species information		
SCIENTIFIC NAME	scientific_name	The scientific name for a species, using the eBird taxonomy. These names as changed retroactively with changes in the eBird taxonomy.
COMMON NAME		The primary English-language name for a species in the eBird database. These names are in wide use in North America, but are not universally recognized. As such, COMMON NAME is not carried through processing by *auk*.
APPROVED		A binary flag indicating whether a record has been flagged potentially erroneous. Non-approved data is not permitted through by *auk*.
OBSERVATION COUNT	observation_count	The total number of individuals of the species reported on a checklist. When zero-filled using *auk*, values of zero are added for species not recorded.
	species_observed	A binary indicator whether a species was detected (1) or not detected (0) on the checklist.
ALL SPECIES REPORTED	all_species_reported	A binary indicator of whether the checklist was a "complete" list of all species detected and identified.
Where did the checklist observations occur?		
LATITUDE, LONGITUDE	latitude, longitude	Decimal latitude and decimal longitude associated with each checklist. Even travelling or area checklists will have a single latitude and longitude.
LOCALITY ID	locality_id	The unique alphanumeric identifier for a location. Many checklists can be associated with a single LOCALITY ID value. It is possible (although extremely rare) for the same LATITUDE and LONGITUDE to be present for multiple LOCALITY ID values. We recommend the use of LOCALITY ID and not LATITUDE/LONGITUDE for connecting together multiple checklists whose data were collected at the same physical location.
COUNTRY CODE	country_code	The two-letter, ISO 3166 code for the country in which a checklist's observations were made. Note that country codes exist for political units that are not autonomous nation-states (e.g., Svalbard and Jan Mayen). While the EBD also contains the fully written out COUNTRY name, these names can be written in extensions to the basic Latin alphabet, or in non-Latin lettering, which both can be mishandled by software (including R). Thus, to guarantee that the identifier for a country will be handled correctly we recommend using COUNTRY CODE rather than COUNTRY. For this reason output from *auk* does not include COUNTRY.
STATE CODE	state_code	The multi-character ISO 3166 code for the political unit immediately below COUNTRY CODE (often termed a "state" or "province") in which a

		checklist's observations were made. These codes are composed of the two-letter COUNTRY CODE, a hyphen, and two or three additional characters. For reasons noted for COUNTRY CODE, the full STATE names are present in the EBD but we do not recommend their use and *auk* does not output the full STATE names.
When did the checklist observations occur?		
OBSERVATION DATE	observation_date	The day of the calendar year for an observation event. This is represented in the EBD in the format YYYY-MM-DD. The output from *auk* converts OBSERVATION DATE into an R date variable of class "Date".
TIME OBSERVATIONS STARTED	time_observation_started	The time at which an observation period was started. This is represented in the EBD as HH:MM:SS. Following processing with *auk*, times are converted into decimal hours (again using a 24-hour clock). Note that all times are reported in the current local time (i.e. official time zone, and local use of standard/daylight savings time).
How did the checklist observations occur?		
DURATION MINUTES	duration_minutes	The total duration of time over which observations were made for the checklist (recorded in minutes).
EFFORT DISTANCE KM	effort_distance_km	The distance travelled during the observation event (in kilometers).
NUMBER OBSERVERS	number_observers	The number of people bird watching together, whose combined observations are reflected in the species and numbers of individual birds recorded on the checklist.
PROTOCOL TYPE	protocol_type	A categorical variable that indicates how the data were collected. The two most common protocol types for complete checklists are "stationary" and "traveling". Unless time is spent understanding the meaning of other PROTOCOL TYPEs, we suggest that users consider using data from only these two PROTOCOL TYPEs. Note that in our experience the PROTOCOL TYPE "stationary" count can be different to a "traveling" count with very low distance travelled. We therefore suggest that both PROTOCOL TYPE and EFFORT DISTANCE KM be used as predictor variables in order to describe variation in the detection process.
SAMPLING EVENT IDENTIFIER	sampling_event_identifier	A unique alpha-numeric identifier for a checklist.
GROUP IDENTIFIER	group_id	An alpha-numeric identifier that connects together multiple checklists that are part of a "shared" group. "Shared checklists" with the same GROUP IDENTIFIER are identical or near-identical checklists that each is associated with the same observation event, but a different OBSERVER ID. This mechanism was required so that each participant in eBird could have all of their bird observations associated with them, even when they differ slightly from those who were bird watching with them. The consequence of the process is that the database contains replicates or near replicates checklists (i.e. statistically non-independent data), which must be collapsed down into a single checklist per observation event. *auk*, by default, collapses the combined observations from all those with the same group_id into a single checklist.
	checklist_calibration_index	This variable is not part of the EBD, but is available separately. checklist_calibration_index provides an index of some combinations of observer behaviour and experience, which vary geographically and through time, that can affect the detectability of species. It can be an important source of variation in detectability.
OBSERVER ID	observer_id	This alpha-numeric identifier has a unique value for each eBird observer. We suggest using checklist_calibration_index as the appropriate method for accounting for inter-observer variation in rates of detection of birds.