ISSN 1948-6596

Broad-scale citizen science data from checklists: prospects and challenges for macroecology

Wesley M. Hochachka* and Daniel Fink

Cornell Lab of Ornithology, Ithaca, NY 14850, USA
*wmh6@cornell.edu

Two recent reviews (Beck et al. 2012, Keith et al. 2012) have outlined potential directions for macroecology to follow, with an expanded suite of questions, methods, and sources of data. One specific recommendation made by Beck et al. (2012) was to expand the use of citizen science data—data collected by a network of volunteers—in macroecological studies. In this commentary we explore the potential for using citizen science data, and feature one specific data resource that we believe will be of interest to macroecologists. We note how such data can be useful for addressing a number of research avenues described in these two reviews, as well as describing some constraints on availability and challenges inherent in using volunteer-collected data. Our own perspectives (and biases) have been shaped largely by our experience using citizen science data for autecological studies of birds. Nevertheless, we see many connections between the process of using citizen science data in studies of individual species and the potential uses of these same data for the interspecific comparisons used in macroecology.

While it is possible to create new citizen science projects to collect required data for answering specific questions (Silvertown et al. 2011), this approach to collecting data for macroecology research suggested by Beck et al. (2012) would require considerable investment in time and resources to implement. For example, the Evolution MegaLab took 2.5 years from receipt of funding until the project was launched (Worthington et al. 2011). The most immediate sources of citizen science data, especially across broad geographic extents, will be from existing projects. While citizen science projects are being used to collect a wide array of types of data[1], one of the most common forms of citizen science data in ecology is observations of species from which one can infer their distributions. These observations are typically collected in the form of 'checklists' of the species seen during an observation period, and often also the numbers of individuals of each species that was observed.

A number of programs exist for the collection of checklist data from birds across large geographical regions, including eBird (Sullivan et al. 2009) that collects data worldwide but concentrates on the western hemisphere, the North American Breeding Bird Survey (USGS Patuxent Wildlife Research Center 2012)[2] and Christmas Bird Count (National Audubon Society 2002)[3] in North America, BirdTrack in the British Isles (Eddowes 2011), and the Ornitho family of schemes in Switzerland[4] and other regions of Europe. Here we focus on describing eBird and the uses of its data specifically, because of our familiarity with the data from this program. Nevertheless, we describe the data from eBird within a broader context because most or all of our observations will be applicable to similar citizen science data from other sources. Checklist data are likely to be easily combined across projects and the regions that they cover, because of the relatively simple and similar protocols across such projects. Hence, while individual programs may collect data from limited geographical regions, we believe that in the future the combination of data from a number of sources will create an invaluable resource with which to describe and understand global distributions of organisms.

1. http://www.birds.cornell.edu/citscitoolkit/projects/find/find accessed 19 November 2012
2. http://www.pwrc.usgs.gov/bbs/ accessed 19 November 2012
3. http://www.audubon.org/bird/cbc accessed 19 November 2012
4. http://www.ornitho.ch/ accessed 19 November 2012

The eBird checklist program and its data

eBird is a web-based system for collecting data on the occurrences and numbers of birds world-wide, although currently concentrating on gathering data from the western hemisphere. Since its inception in 2002, eBird has been gathering an ever-increasing volume of data each year, with a monthly peak of over 110,000 checklists submitted in 2012 (see Sullivan et al. 2009 for more background information on eBird and its operation). All checklists submitted to eBird are reviewed for unusual observations based on the location and date of observation (Kelling et al. in press). The data from eBird are compiled annually into the eBird Reference Dataset, which contains not just the data collected on birds, but additionally for the lower 48 states of the United States associated data describing the habitat around the locations of all observations[5]. Data from eBird have been used for purposes as varied as motivating the development of novel analytical methods (Hochachka et al. 2012) and guiding conservation policy[6].

A source of data, where volunteers exist

The requirement for a large number of volunteers to collect data, particularly at broad geographical scales, leads to unequal distributions of citizen science data across the globe. Similarly, some taxonomic groups are more charismatic and have a greater number of people interested in making observations, leading to a greater availability of data for these taxa. Birds have a large existing constituency of observers recording their presence, and tapping into the motivations of bird watchers has been key to the success of bird checklist programs (Sullivan et al. 2009). However, even these datasets often exhibit biases arising from the uneven distribution of search effort. Without a geographically explicit protocol for data collection, volunteer search effort tends to follow patterns of human population density and roads. The distribution of observers and their decisions of where to make observations are not necessarily motivated by a desire to collect representative data across a landscape (Sullivan et al. 2009, Tulloch and Szabo 2012). Quite simply, data are more likely to be collected where more potential observers live, and in areas that are more easily accessible.

eBird is no exception to this generalization. Within the United States and Canada, densities of data collected in eBird approximate densities of human populations, with the northeastern United States and adjacent Canada having a very high density of available data, and with the lowest densities of data in the Great Plains in the middle of the continent, and in northern Canada. Into the Caribbean, and Central and South America, the density of available data is also lower not because of lower human densities per se, but because of a lower proportion of bird watchers within populations.

As with geographic biases, taxonomic biases in data collection are driven by observers. Data from birds are more readily available than those from other taxa because eBird and similar bird checklist programs tap into existing groups of potential volunteers that are present because bird watching is a well-established hobby. Sometimes expanded interests of bird watchers creates avenues for expanded taxonomic coverage, as is the case with the British BirdTrack system that now also collects observations on dragonflies[7]. Nevertheless, data on birds predominate among data collected as checklists of organisms. An illustration of biases in taxonomic coverage can be obtained by browsing the numbers of observations collected in one all-taxa observation-reporting system, Observado.org[8].

In summary, where appropriate data exist for a taxonomic group and geographic region, we suggest that opportunities to use these data should be explored more fully by macroecologists.

5. http://www.avianknowledge.net/content/download accessed 19 November 2012
6. http://www.stateofthebirds.org/State%20of%20the%20Birds%202011.pdf accessed 19 November 2012
7. http://www.bto.org/volunteer-surveys/birdtrack/taking-part/recording-your-sightings/recording-dragonflies accessed 19 November 2012
8. http://observado.org/statistiek.php accessed 19 November 2012

However, the existing geographic and taxonomic biases in studies in macroecology (see Figures 3–5 in Beck et al. 2012) are unlikely to be erased through the use of citizen science data.

Broad extent with fine resolution

Within the constraints of available data, citizen science checklist projects can provide very useful data with which to address the issues of scale raised in the two reviews (Beck et al. 2012, Keith et al. 2012) by providing data that have both broad extent and small grain. The same loose protocols that result in a sparsity of data from some areas also result in high densities of data in other areas. Using data from eBird, we have found that individual species can exhibit long-range variation in the local-scale associations between prevalence and landcover (Hochachka et al. 2012). Further, when data density is sufficient, distributional patterns can be examined at multiple, nested spatial resolutions and fine-scale patterns may be compared across larger-scale regions. Lovette & Hochachka (2006) demonstrate the utility of this approach using data examined at two spatial extents, in order to define regional pools of species within which to examine constraints on local (location-specific) associations among species.

Temporal information is available

Because observations can be made at any time throughout the year, eBird data also contain temporal information across a range of scales (e.g. time of day, day of year, and year of observation). These data open up the possibility for exploring new questions in macroecology that were not highlighted in either of the two recent reviews. For example, data from eBird contain sufficient information to describe seasonal variation in the distributions of a number of species across the breadth of North America[9], addressing such questions as the flexibility with which the timing of bird migration can be adjusted to climatic conditions (Hurlbert & Liang 2012). The availability of data on within-year temporal variation contrasts markedly with some data sources such as basic species lists for geographical areas. For sessile species such as plants, collapsing time in records may not be critical for many purposes, but for highly mobile taxa such as birds even basic information on species' probabilities of association with each other (in the same place at the very same time) depends on having data of fine temporal grain. Further, the fine temporal grain of checklist data open up the possibility of investigating whether patterns in biodiversity are dependent on not just spatial scale but temporal scale.

Additionally, associations of species with each other or an individual species with features of its environment may change throughout the year, even for species that are non-migratory. To date, macroecology typically searches for patterns that are generalizable across space. We see an intriguing potential for similar investigations of generalities across time.

Ability to filter out the observation process

Data per se do not represent biological truth. Any ecological datum results from the combination of a biological process and an observation process, and assuming that the observation process does not exist can lead to erroneous biological inferences. Even observations of large sessile organisms are imperfect because observers are fallible (Chen et al. 2009). The likelihood of detecting an individual organism can vary for a number of reasons, including the following:

- Observer effort – the longer an observer takes to gather data in a checklist or the farther they travel, the greater the number of organisms they will observer. Variation in observer effort needs to be accounted for in order to compare observations.
- Time of observation – insects are more active and easily observed at warmer times of day away from the morning, whereas many bird species are best detected in the early morning particularly when displaying on their breeding grounds. At the broader time scale of an annual cycle, organisms will change in detectability as well. Thus temporal variation in detect-

9. http://ebird.org/content/ebird/news/ebird-animated-occurrence-maps accessed 19 November 2012

ability needs to be identified and controlled for in the use of checklist data.

- Habitat – the environment in which a checklist is collected not only determines the species present, but can also affect the observers' abilities to detect the organisms that are present. Confounding species' habitat preferences with their detectability in different habitats can lead to erroneous conclusions (Ruiz-Gutiérrez et al. 2010).

A carefully designed survey can control for observation effects, for example by keeping observers' effort constant, or even collecting data in such a way that the biases introduced by the observation process can be eliminated during analyses (Royle and Dorazio 2008). Typically, repeated observations at the same locations are used to estimate detection probabilities. While many checklist data such as those in eBird are not formally collected in a repeated-visits format, analysts may be able to coerce these data into a form in which they can be used in statistical analyses that explicitly model the observation process (Kéry et al. 2010). Alternatively, it is also possible to account for detection probabilities even without repeated data from the same location (Sólymos et al. 2012), if the necessary assumptions can be met by the data. Less ideally, covariates that affect detection probability can be included in models such that estimated probabilities of occurrence are produced that are conditional on selected values of the covariates that affect detection (e.g., Fink et al. 2010). A basic precondition for any of the above methods is that 'presence–absence' data—data for which both detections and non-detections are known—are collected.

Without knowing which species were undetected, or otherwise being able to correct for biases in reporting rates, biased biological inferences likely will result. Data from eBird and some other checklist programs are collected as complete lists of species identified during an observation event, or at least the incomplete nature of some lists can be identified by data analysts. In eBird, observers are explicitly asked whether their lists represent all of the species that were identified during an observation period; data cannot be submitted without this question being answered. Not all checklist programs collect this complete-checklist information. In contrast, museum specimens or data compiled into a museum-like format such those available through GBIF[10] are examples of data that lack the contextual information necessary to assess and, subsequently, correct for non-detection. In such data, each species, or even individual organism, is treated as an entirely separate entity collected in an independent collection event, and stored as a separate record. Statistical methods, notably Maxent (Phillips et al. 2009) have been designed to deal with data for which no information is available about non-observation events ('presence-only' data). Such methods work by creating a set of 'pseudo-absence' data in lieu of having actual information on non-detections.

For studies across broad geographic or long temporal extents, detection rates may vary both temporally (e.g., Hochachka et al. 2009) and spatially (e.g., Ruiz-Gutiérrez et al. 2010). Thus, the lack of detection does not convey the same information about true absence at every geographic location within the region of interest and at every time of data collection. This is a challenge for the analysis of both presence–absence and presence-only data. Our feeling is that such variation in detection probabilities is ubiquitous enough that any analysis and subsequent interpretation of results from broad-scale data should explicitly consider the likelihood that the statistical assumption of stationarity of the observation process is true, and the effects of this assumption being invalid.

In conclusion, we would encourage biogeographers and macroecologists to investigate whether data are available from existing citizen science projects, most likely data collected as checklists of organisms, and consider whether these data are amenable to the investigators' use. We have described the eBird checklist program in detail both to encourage the use of its data and as an example of the types of data available from checklists

10. http://www.gbif.org/ accessed 19 November 2012

of organisms and the considerations that need to go into the analyses of such data. We believe that many existing data can potentially be used to explore temporal dynamics from a macroecological perspective and to explore the consequences of changing spatial and temporal scale—both grain and extent of data (Beck et al. 2012)—on the biological patterns that are observed.

References

Beck, J., Ballesteros-Mejia, L., Buchmann, C.M., Dengler, J., Fritz, S.A., Gruber, B., Hof, C., Jansen, Knapp, F., Kreft, H. Schneider, A.-K., Winter, M. & Dormann, C.F. (2012) What's on the horizon for macroecology? Ecography, 35, 673–683.

Chen, G., Kéry, M., Zhang, J. & Ma, K. (2009) Factors affecting detection probability in plant distribution studies. Journal of Ecology 97,1383–1389.

Eddowes, M.J. (2011) Correlations between long distance migrants in year-to-year fluctuations of arrival timing. Ringing & Migration, 26, 48–55.

Fink, D., Hochachka, W.M., Zuckerberg, B., Winkler, D.W., Shaby, B., Munson, M.A., Hooker, G., Riedewald, M., Sheldon, D. & Kelling, S. (2010) Spatiotemporal exploratory models for broad-scale survey data. Ecological Applications, 20, 2131–2147.

Hochachka, W.M., Fink, D. Hutchinson, R.A., Sheldon, D., Wong, W.-K. & Kelling, S. (2012) Project and analysis design for broad-scale citizen science. Trends in Ecology & Evolution, 27, 130–137.

Hochachka, W.M., Winter, M. & Charif, R.A. (2009) Sources of variation in singing probability of Florida Grasshopper Sparrows, and implications for design and analysis of auditory surveys. Condor, 111, 349–360.

Hurlbert, A. H. & Liang, Z. (2012) Spatiotemporal variation in avian migration phenology: citizen science reveals effects of climate change. PLoS ONE, 7, e31662.

Keith, S.A., Webb, T.J., Böhning-Gaese, K., Connolly, S.R., Dulvy, N.K., Eigenbrod, F., Jones, K.E., Price, T., Redding, D.W., Owens, I.P.F. & Isaac, N.J.B. (2012) What is macroecology? Biology Letters, 8, 902–906.

Kelling, S. Fink, D., Lagoze, C., Wong, W.K., Yu, J., Damoulas, T. & Gomes, C. (in press) eBird: a human/computer learning network for biodiversity conservation and research. AI Magazine.

Kéry, M., Gardner, B. & Monnerat, C. (2010) Predicting species distributions from checklist data using site-occupancy models. Journal of Biogeography, 37, 1851–1862.

Lovette, I.J. & Hochachka, W.M. (2006) Simultaneous effects of phylogenetic niche conservatism and competition on avian community structure. Ecology, 87, S14–S28.

National Audubon Society (2002) The Christmas Bird Count Historical Results.

Phillips, S.J., Dudík, M., Elith, J., Graham, C.H., Lehmann, A., Leathwick, J. & Ferrier, S. (2009) Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecological Applications, 19, 181–197.

Royle, J.A. & Dorazio, R.M. (2008) Hierarchical modeling and inference in ecology: the analysis of data from populations, metapopulations and communities. Academic Press, San Diego, California.

Ruiz-Gutiérrez, V., Zipkin, E.F. & Dhondt, A.A. (2010) Occupancy dynamics in a tropical bird community: unexpectedly high forest use by birds classified as non-forest species. Journal of Applied Ecology, 47, 621–630.

Silvertown, J., Cook, L., Cameron, R., Dodd, M., McConway, K., Worthington, J., Skelton, P., Anton, C., Bossdorf, O., Baur, B., Schilthuizen, M., Fontaine, B., Sattmann, H., Bertorelle, G., Correia, M., Oliveira, C., Pokryszko, B., Ożgo, M., Stalažs, A., Gill, E., Rammul, Ü., Sólymos, P., Féher, Z. & Juan, X. (2011) Citizen science reveals unexpected continental-scale evolutionary change in a model organism. PLoS ONE, 6, e18927.

Sólymos, P., Lele, S. & Bayne, E. (2012) Conditional likelihood approach for analyzing single visit abundance survey data in the presence of zero inflation and detection error. Environmetrics, 23, 197–205.

Sullivan, B.L., Wood, C.L., Iliff, M.J., Bonney, R.E., Fink, D. & Kelling, S. (2009) eBird: A citizen-based bird observation network in the biological sciences. Biological Conservation, 142, 2282–2292.

Tulloch, A.I.T. & Szabo, J.K. (2012) A behavioural ecology approach to understand volunteer surveying for citizen science datasets. Emu, 112, 313–325.

USGS Patuxent Wildlife Research Center (2012) North American Breeding Bird Survey Internet data set.

Worthington, J. P., Silvertown, J., Cook, L., Cameron, R., Dodd, M., Greenwood, R. M., McConway, K. & Skelton, P. (2011) Evolution MegaLab: a case study in citizen science methods. Methods in Ecology and Evolution, 3, 303–309.

Edited by Frank A. La Sorte

