Darcy Higgins
EE Baja V: Field Methods
June 8, 2015

Optimizing and Correcting Citizen Science Data Quality

Citizen science has been touted as an efficient conservation method that kills two birds with one stone: it can potentially provide both increased public engagement and education around conservation, and gather needed ecological data for research or management. But scientists and managers sometimes perceive data quality and quality of citizen engagement as a trade-off (Reisch et al. 2011). 'Compromise' is not an ideal for data quality. Does using citizens as data collectors actually produce predictable pitfalls in data quality, and if so, are there methods and protocols that can overcome those flaws?

One of the simplest methods to assess the quality of citizen-collected data is to use professionally-collected data as the standard. Studies comparing data collected by amateur citizens to professionally-collected data suggest that it is possible for citizens to collect data that is statistically comparable to professionals' (Aceves-Bueno et al. 2015, Belt & Krausman 2012, Crall et al. 2011, Delaney et al. 2008, Fore et al. 2001). One study of crab identification skills found that even third-graders made the same identification as biologists 80% of the time, and citizens with two years or more of college education made the same identification 95% of the time (Delaney et al. 2008). Fore et al. (2001) found that volunteers performed as well as professionals conducting biological stream monitoring in the field, and categorization of stream quality was the same from both data -- an important finding for 'stream team' programs underway in several states conservation departments.

However, although researchers have been able to draw comparable inferences from citizen-collected data in many studies, that does not mean that professional-collected and citizen-collected data sets are the same. Many comparisons of professional and citizen-collected data include a caveat: volunteers' data can generate the same inferences as professionals' *if* the volunteer data provide more information -- generally, if citizen-collected data provide a larger sample size (Belt & Krausman 2012, Hochachka et al. 2011, Munson et al. 2010, Schmeller et al. 2009). Belt and Krausman, for example, found that although volunteers were less likely than biologists to detect the presence of mountain goats on a given site visit, this was counterbalanced by the greater number of site visits conducted by volunteers (2012). Ultimately, the volunteers' collective estimates of overall mountain goat abundance was less variable than the biologists'. These is evidence that citizen science encourages larger sample sizes in studies. For example, an ambitious survey of biodiversity monitoring programs across Europe found that higher sample size and more days spent sampling were correlated with higher percentages of volunteers amongst data collectors. That is higher percentages of volunteers, not just higher numbers of volunteers, meaning more data was collected even in small monitoring programs if they relied more on volunteers (Schmeller et al. 2009).

This tendency is fortunate, because, as the mountain goat example demonstrates, larger sample sizes may be a tool for overcoming some common flaws that emerge in the literature on citizen science data quality.

One flaw lies in inconsistently-interpreted or loose collection protocols, which result in inconsistent data that are difficult to compare (Munson et al. 2010). For example, although the bird-checklist web platform eBird requests that users list the approximate centerpoint of their observation route as their "location", many list their starting point instead. This inconsistency creates challenges for geographic analysis of data.

Variations in effort by volunteers, or insufficient skill in species identification, species detection or equipment use, are other sources of error and data variability (Aceves-Bueno et al. 2015, Cohn 2008, Dickinson et al. 2010, Gonsamo & D'Odorico 2014). For example, one anuran call survey program found that volunteers only reliably identified certain species' calls correctly, and needed to mitigate this skill problem by requiring its volunteers to pass an online "frog quiz." (Dickinson et al. 2010). Another study found that volunteers newer to identifying birds by song did not hear birdsong that was outside of the frequencies easiest heard by the human ear. Unchecked, that auditory bias would create the illusion in a dataset that these low-singing birds were not present, when they in fact were.

Munson et al. (2010) conducted a detailed comparison of the opportunistic eBird monitoring program to the relatively highly-structured Breeding Bird Survey. eBird lets birders submit checklists of species they saw on their own expeditions. Timing, length of the outing, and location are all dependent on the birder's own choices, and are not strictly codified. The North American Breeding Bird Survey, on the other hand, has specific protocols, including predetermined roadside observation routes, three-minute point surveys along the route, and a starting time half-an-hour before dawn. The investigators consider it the most reliable dataset of its kind. eBird, by comparison, could be anticipated to have many of the common flaws of citizen science project data, including: inconsistent protocol, unknown variability in the time birders spent looking for species, and overrepresentation of species that come out at popular times and places for participants to look for birds. Indeed, Munson et al. (2010) did find that the eBird dataset generated a lot of noise, that is, illusions of patterns that were caused by data collection methods rather than real population trends. eBird has more reports from population centers, for example, so a relatively higher abundance of species in urban areas than rural areas says more about where eBird participants live than where birds live.

However, despite the noise, Munson et al. (2010) ultimately concluded that eBird provided the same quality of population data as the Breeding Bird Survey. How is that possible, when the eBird data possessed such greater variability? First, they excluded surveys that users had marked as casual or left incomplete. Then, rather than using the raw data, they created a calibration model that corrected for some of eBird's biases, and crosschecked it against a smaller subsample of the Breeding Bird Survey to verify its efficacy. Once calibrated, both datasets as complete wholes were able to provide the same quality of inferences; however, each individual datum of the eBird set contained a lesser quantity of useful information. The Breeding Bird Survey had highly efficient data, providing a more bang for the buck than eBird -- able to provide the same information from one survey that eBird needed 3 to 4 surveys to provide reliably. But with 21,175 surveys from eBird and 6,460 from Breeding Bird Survey suitable for inclusion in the analysis, inefficiency did not impede eBird's usefulness (ibid).

In a professionally-conducted study, inefficiency would be a major concern because each datum has a price tag. For certain kinds of data collection, such as monitoring forest habitat, citizen-collected data can be as much as half the cost as a professional; this does not hold true for studies that require complex infrastructure or additional insurance against risks, such as diving off of boats (Aceves-Bueno et al. 2015). The survey of European biodiversity monitoring found that, at least in more developed countries where labor costs are higher than in the developing world, volunteer participation lowered project costs (Schmeller et al. 2009). In the eBird example, the inefficiency of loose protocols may have paid for itself by lowering barriers to volunteer participation, facilitating the larger sample size that makes the data useful.

However, to work the statistical magic that made the large eBird dataset useful, a dataset known to be more controlled was required as a benchmark (Munson et al. 2010). This implies that this loose and large style of citizen science may be best used as a complement with more regulated research, not as a replacement. Return on investment, as well as potential opportunity cost from allocating resources to one project over another, could be considered in relation to the research being done in a field as a whole to evaluate true strategic value.

Many evaluations of citizen-collected data recommend more defined and rigid collection protocols as a best practice for citizen science projects, as well as perhaps limiting those protocols to narrower scopes than professional research or monitoring schemes (Cohn 2008, Dickinson et al. 2010, Fore et al. 2001). While this is undeniably good advice, the eBird example provides a counter-example of what is possible. The field of birding in general has one of the longest-established case histories in citizen science, with the still-running Christmas Bird Count started in 1900, and these long-lasting projects did not have such rigorous controls (Dickinson et al. 2010). Yet citizen science data on birds have provided some of the strongest extant evidence of climate-change-related population range shifts (ibid). This is practical evidence that imperfect data can still provide important information about trends and be relevant to management choices. Numerically, Aceves-Bueno et al. (2015) surveyed citizen science projects, and found that 89% of projects reporting on quality assurance spoke of problems in that data. Fifty-nine percent reported that the imperfections were minor, and 39% said that the problems were "critical, but fixable," meaning that 98% of the projects were able to get useful, meaningful conclusions from even flawed datasets.

Beyond the eBird survey, other investigators have used statistical adjustments as the tool to 'fix' those flawed datasets, and methods vary by the particular flaw in the data. Gonsamo and D'Odorico (2014) examined observer bias in citizen science data, which is any trend in the data that says more about the observers than ecological reality: favoring locations and times when the observers like to sample, errors in species identification, or different perception of what exact stage counts as the bud bursting. They developed a statistical method for analyzing phenological records. Instead of directly averaging all the available records of species and sites, they removed observations from seldom and first-time observers, then calculated anomaly and trend data from a baseline specific to each site and each species. This brought the data much closer to echoing professionally-collected conclusions than the direct aggregation. Kéry et al. (2010) suggest another statistical method for correcting for observer effort, if protocols have not standardized it. Because observer effort tends to increase as citizen science projects get older, this can create an erroneous impression that species populations are improving. Kéry et al. estimated the probability that species 'presence or absence were correctly detected in the data set, and used this probability to correct changes in effort from year to year (2010).

However, statistical methods that increase in complexity tend to include more assumptions about reality, and always carry the risk of moving further and further away from facts on the ground (Feinsinger 2001); using statistical tricks to improve citizen-collected may be useful but worth taking with a grain of salt. Methods for filtering data quality before analysis may be simpler. For example, the Missouri Stream Team programs accepts chemical and biological stream monitoring data from anyone who has taken its workshops, which begin at an introductory level and continue to level 3. However, only data from verified level 2 and 3 reporters are used independently for substantial analysis (Missouri Department of Conservation, 2007). The introductory and level 1 users' data are used to monitor for emerging problems, as baseline data, or as supplements to government-collected data sets, echoing how the bird surveys above supplemented each other. There is also evidence from bird surveys that first-year observers do not perform as well as more experienced observers, an effect that is magnified in early years of a new project when there are no experienced volunteers to guide and balance out the newcomers (Jiguet 2009). However, these kinds of filtering choices should be chosen based on analysis of actual data sets or those of very similar projects; generic demographic data, such as age, education, attitude, scientific literacy, and even experience, were not found to be good predictors of performance in one invasive species citizen science project. Only participants' self-identified comfort levels predicted success (Crall et al. 2011). Training methods are another important contributor to fixing data quality before the analysis step, and deserve their own paper.

One reason that citizen-collected data have been so critical to studies of global climate change, phenology, and population is that use of citizens functions as a method that professionally-collected data schemes lack. National or international citizen science programs have a geographic reach that is difficult to match with staff researchers, even without considering funding. Instead of ecologists drawing inferences that apply only to local ecosystems, inferences can be drawn across species' range; not only is the geographic reach large, the data is also capable of a relatively finer scale, able to zoom in on a location with detail (Kéry et al. 2010, Munson et al. 2010). Dickinson et al. (2010) attribute growth in the spatially-focused fields of macroecology and geographical ecology to the existence of citizen science data. Although, as seen with eBird, using so many data collectors may increase variability in data, professionally-conducted studies across regions or continents also requires coordinating multiple agencies and collectors. That means that professionally-collected data across regions may also struggle with inconsistency of data (Sharpe & Conrad 2006, Aceves-Bueno 2015).

Success of citizen science schemes' data is not uniform across study types. Using public citizens to collect data is perhaps best considered a method choice in and of itself, able to increase or decrease the meaningfulness of data depending on the situation. Like all methods, citizen-collection provides high-quality data for certain objects of study, and poor-quality data for others. In fields that require analysis of large geographic trends, citizen science has been noted as providing some of the most important data sets available. Smaller studies cannot substitute scale for quality control, and so may turn to more tightly controlled designs. Accounting for errors that volunteers are known to make -- including first-time observer effects, more challenging varieties of investigation, and favoring certain times and places over others -- perhaps requires constraining the kinds of data they are asked to contribute. However, if looser protocols encourage greater involvement, citizen science projects may be able to compensate for the data flaws that result by filtering or applying statistical modeling to the larger datasets that result. Designing citizen science projects with thought towards how they complement extant research may provide quality benchmarks, while providing methods that overcome limitations of traditional research, especially resource limitations.

Discussion Questions

1. Think of a data collection project that is relevant to you, such as one at your workplace or that we have interacted with on this trip. What are some likely sources of bias that would be introduced if volunteers collected that data? Is applying data filters, using statistical techniques on a large data set, or some other method the best way to account for those biases?
2. How does the question of data quality control inform your opinion of good citizen science protocols? Is it worth using less strict protocols in pursuit of another goal?
3. How would you evaluate whether a given line of inquiry is a good or poor candidate for citizen data collection?

Works Cited

Aceves-Bueno, E., Adeleye, A., Bradley, D., Tyler Brandt, W., Callery, P., Feraud, M., Garner, K., Gentry, R., Huang, Y., McCullough, I., Pearlman, I., Sutherland, S., Wilkinson, W., Yang, Y., Zink, T., Anderson, S., & Tague, C. (2015). Citizen Science as an Approach for Overcoming Insufficient Monitoring and Inadequate Stakeholder Buy-in in Adaptive Management: Criteria and Evidence. Ecosystems, 18(3), 493-506.

Belt, J. J., & Krausman, P. R. (2012). Evaluating population estimates of mountain goats based on citizen science. Wildlife Society Bulletin, 36(2), 264-276.

Crall, A. W., Newman, G. J., Stohlgren, T. J., Holfelder, K. A., Graham, J., & Waller, D. M. (2011). Assessing citizen science data quality: an invasive species case study. Conservation Letters, 4(6), 433-442.

Cohn, J (2008). Citizen science: Can volunteers do real research? BioScience 58: 30, 192-197.

Delaney, D., Corinne, S., Christiaan, A., & Brian, L. (2008). Marine invasive species: validation of citizen science and implications for national monitoring networks. *Biological Invasions*, *10*(1), 117-128.

Dickinson, J., Zuckerberg, B., & Bonter, D (2010). Citizen science as an ecological research tool: Challenges and benefits. Annual Review of Ecology, Evolution and Systematics 41, 149–172.

Feinsinger, P. (2001). Small Samples and Big Questions: The Role of Statistical Inference. In Feinsinger, P. and The Nature Conservancy, *Designing Field Studies for Biodiversity Conservation* (57-86). Island Press.

Fore, L. S., Paulsen, K., & O'Laughlin, K. (2001). Assessing the performance of volunteers in monitoring streams. *Freshwater Biology*, *46*(1), 109-123.

Gonsamo, A., & D'Odorico, P. (2014). Citizen science: best practices to remove observer bias in trend analysis. *International Journal Of Biometeorology*, *58*(10), 2159-2163.

Jiguet, F. (2009) Method learning caused a first-time observer effect in a newly started breeding bird survey. *Bird Study, 56(253).*

Kéry, M., Royle, J.A., Schmid, H., Schaub, M., Volet, B., Hafliger, G., and Zbinden, N. Site-Occupancy Distribution Modeling to Correct Population-Trend Estimates Derived from Opportunistic Observations. *Conservation Biology, 24(5)*, 1388-1397.

Missouri Department of Conservation Stream Team Program (2007). Volunteer Water Quality Monitoring Validation. Retrieved June 29, 2015, from http://mostreamteam.org/wqval.asp

Munson, M., Caruana, R., Fink, D., Hochachka, W., Iliff, M., Rosenberg, K., & ... Kelling, S. (2010). A method for measuring the relative information content of data from different monitoring protocols. *Methods In Ecology And Evolution*, *1*(3), 263-273.

Riesch, H., Potter, C., & Davies (2013). Combining Public Engagement and Citizen Science: The Open AirLaboratories Programme. *Journal of Science Communication 12:3,* 1-19.

Schemeller, D.S., Henry, P., Julliard, R., Gruber, B., Clobert, J., Dziock, F., & Henle, K. (2009). Advantages of Volunteer-Based Biodiversity Monitoring in Europe. *Conservation Biology*, *23*(2), 307-316.

Sharpe, A., & Conrad, C. (2006). Community based ecological monitoring in Nova Scotia: challenges and opportunities. *Environmental Monitoring And Assessment*, *113*(1-3), 395-409.