Crowdsourcing Citizen Science Data Quality with a Human-Computer Learning Network

Andrea Wiggins **Jeff Gerbracht** **Carl Lagoze**
University of New Mexico Cornell Lab of Ornithology University of Michigan
Cornell University Ithaca, NY 14850 Ann Arbor, MI 48109
Ithaca, NY 14850 *jeff.gerbracht@cornell.edu clagoze@umich.edu*
andrea.wiggins@cornell.edu

Jun Yu **Weng-Keen Wong** **Steve Kelling**
Oregon State University Oregon State University Cornell Lab of Ornithology
Corvallis, OR 97331 Corvallis, OR 97331 Ithaca, NY 14850
juyu@eecs.oregonstate.edu wong@eecs.oregonstate.edu stk2@cornell.edu

Abstract

Citizen science presents opportunities for crowdsourcing to produce new sources of data that were previously unavailable and even unimaginable. While engaging distributed observer networks is a well-established method for collecting spatiotemporally diverse observational data, ensuring data quality remains a key concern. Related problems of observer reliability, scalability of data verification processes, geospatial biases of observations, and motivating participation are central challenges for citizen science. This paper describes a multidisciplinary strategy for addressing these concerns in eBird through the development of a human-computer learning network.

1 Introduction

The transformational power of modern computing, together with information and communication technologies, creates new opportunities to engage the public in participating in and contributing to myriad scientific, business, and technical challenges. For example, large-scale citizen science projects such as Galaxy Zoo, eBird, and FoldIt demonstrate the power of crowdsourcing for investigating complex scientific problems.

In this paper, we describe citizen science in the context of crowdsourcing scientific data collection and verification. The eBird project provides an example of a human-computer learning network (HCLN) that will leverage the aggregated data provided by volunteers to improve the quality of crowdsourced data.

1.1 Crowdsourced Data and Quality

Crowdsourcing is an ill-defined but increasingly common term that refers to a set of distributed production models. Initially used to describe an outsourcing strategy that makes an open call for contributions from a large, undefined network of people [1], it was introduced as a novel alternative business model.

Early definitions of crowdsourcing focused on corporate entities in drawing on the "wisdom of crowds" [2], but more recent popular use has applied the term to any form of collective intelligence that draws on large numbers of participants through the Internet. In many scientific contexts, doubts arise as to the value of crowdsourcing, primarily regarding veracity and accuracy of crowdsourced research product, or data quality, used here to mean the fitness of the data for its intended purpose.

1.2 Citizen Science for Crowdsourced Data

Citizen science projects enlist the public in scientific endeavors [3], and can provide data with more intensive sampling through longer periods of time and across broader spatial extents. Volunteers can collect large amounts of data at comparatively little cost to the scientific research enterprise [4]. Furthermore, engaging citizen scientists in meaningful research projects can enrich the public understanding of the scientific process, which in turn can lead to better-informed decision making at all levels of society [5].

Data quality is an important issue in citizen science because data contributors often have little or no scientific training. Distributed and large-scale participation eliminates the options of supervision or ground-truthing to ensure data quality. Consequently, the quality of the data collected by volunteers is often questioned. While most citizen science projects employ multiple strategies to improve data quality [6], observational data that rely on species detection and identification skills remain particularly challenging. While volunteers can provide accurate data for easily detected organisms [7], Fitzpatrick et al. [8] found differences between volunteers and professionals for difficult-to-detect organisms led to biases in the data [9]. These issues affect many citizen science projects, such as eBird.

1.2 eBird

The eBird project, launched in 2002 by the Cornell Lab of Ornithology and National Audubon Society [10], is one of the largest citizen science programs in existence. eBird maintains an online database that allows bird watchers (known as *birders*) to record the bird species they have seen or heard. eBird's goal is to maximize the utility and accessibility of the millions of bird observations made each year by recreational and professional birders.

Data quality is a major issue for eBird, particularly regarding an observer's ability to correctly identify birds to the taxonomic level of a species. A network of 450 expert volunteers create checklist filters for outliers and review the resulting flagged records. Reviewers draw upon local bird expertise to create filters delineating when and how many of each species are expected in a specific region. Reviewers also contact individuals for more information to confirm unusual records. eBird's success now generates millions of new observations per month, overwhelming the reviewer network with about one million records to review in 2012. This also stymies further growth due to scalability constraints.

New methods for improving eBird data quality further leverage the aggregated data generated by participation. These include emergent filters and modeling participant expertise, both of which are dependent upon the volume of existing data. More work is needed to incentivize further contribution, both to generate adequate volumes of data to support data quality automation and to reduce geospatial bias in the data.

2 The eBird Human-Computer Learning Network

To address these inter-related issues, the eBird Human-Computer Learning Network (HCLN) will combine emerging techniques that integrate the speed and scalability of *mechanical computation*, using advances in Artificial Intelligence (AI), with the real intelligence of *human computation* to solve computational problems that are beyond the scope of existing algorithms [11]. In addition to developing emergent filters and new models of contributor expertise, this work is novel in that it makes extensive use of the semantic links between observations and observers to mine additional information from the existing data in order to strategically address data quality issues.

eBird's data contain the following information: observer, location, visit, species, and number observed. Information about the observer, such as name and contact information, allow every bird observation to be attributed to a specific person. Location data include site name, coordinates, and the geographic area it represents for every visit to that location. Information about a specific visit includes date and time, amount of effort expended, and whether all observed species were reported. Species observations consist of a bird checklist and counts of individuals of each species. These data form the core of the eBird database, enabling advanced computational methods to improve data quality while reducing human review.

Three core components of the eBird HCLN are discussed briefly below; each uses existing data stores to generate additional value from the original crowdsourced data.

2.1 Modeling Contributor Expertise

In order to incorporate birder expertise into a species distribution model, we need to distinguish between two processes that affect observations: *occupancy* and *detection*. Occupancy determines if a geographic site is viable habitat for a species. Detection describes the observer's ability to detect the species and depends on the difficulty of identifying the species, effort put in by the birder, current weather conditions, and birder expertise.

To evaluate birder expertise, we modified existing Occupancy-Detection models to include detection parameters for novices and for experts. These features are useful if the detection of a species is different for experts versus novices. It also allows for false detections by both experts and novices to improve predictive ability because experts can be over-enthusiastic about reporting bird species under certain circumstances. As a result, the detection probabilities for novices and experts in the resulting Occupancy-Detection-Expertise (ODE) model include both true and false detection probabilities for experts and for novices [9].

In preliminary testing, the ODE model outperformed multiple comparison models. In addition, experts and novices appear to have very similar true detection probabilities for common bird species. For hard-to-detect bird species the differences are much larger, however, making this a promising approach to identifying differences in how experts and novices report bird species.

2.2 Emergent filters

eBird currently relies on expert-defined checklist filters, but could improve data quality through more fine-grained checklist filters based on historical data for each location. Automating filter development would also reduce demands on the already overtaxed volunteer reviewer network.

The emergent filters are calculated from existing data for the observation location, based on daily frequencies for every species reported there [12]. These filters can be tuned to threshold occurrence frequencies to identify outlier observation at any level of specificity. The emergent filters also have potential to identify outlier reports for common birds that would otherwise be overlooked with expert-defined filters, further improving data quality.

In combination with the automatic ranking of expertise discussed above, emergent filters have potential to substantially improve data quality and lighten the load on reviewers, whose time is a limited resource. Combined into a two-step process, these tools would first reliably *identify* outliers, and then *classify* those outliers. Uniquely, outliers are not removed but instead classified as either unusual or erroneous. This process requires sufficient domain knowledge (e.g., understanding of the patterns of bird distributions) to distinguish between types of outliers, and sufficient data so that a quality filter can emerge from the data [12]. One challenge in fully implementing this approach is the current biases with the geospatial coverage of data submitted by volunteers.

2.3 Improving spatial coverage

Another substantive problem in citizen science data quality is the spatial bias in favor of locations near population centers [13]; our work in this area is still in progress. It will integrate Active Learning feedback loops to guide volunteers' sampling process toward improving current models by filling in gaps.

We plan to employ a variation on the Traveling Salesman Problem with Active Learning processes to improve predictive models by providing a context to advise participants where to sample next, in the form of proposed birding routes for any starting location and level of effort planned. Such sampling paths could be incorporated into games and optimized to enhance both the machine learning elements of the system and the overall birding experience. These dynamic, adaptive paths would serve the mutual interests of volunteers and researchers by proposing personalized routes to hone individual detection capabilities or increase the probability of recording a species never previously observed, while simultaneously increasing geospatial coverage.

3 Conclusion

Our work on the eBird HLCN involves translating the preliminary studies described above into an integrated system that fully maximizes the return on investment for data contributions from citizen scientists [13]. Additional evaluation will focus on questions such as whether displaying expertise rankings has motivational effects and whether the system encourages expanded participation or increased learning, and on better understanding volunteers' detection capabilities to further improve expertise measurements and rankings.

In a project like eBird, with a broad and active citizen science volunteer constituency, well managed cyberinfrastructure provides a unique opportunity to test and deploy new techniques that tackle some of the most pressing issues of data quality applicable to any citizen science project. While our focus is on techniques to address errors in data submission, variability in observer skills, and spatial coverage biases in eBird, these strategies have wide applicability across the broader citizen science field of practice.

Acknowledgments

This work was funded by the Leon Levy Foundation, Wolf Creek Foundation and the National Science Foundation (Grants OCI-0830944, CCF-0832782, ITR-0427914, DBI-1049363, DBI-0542868, DUE- 0734857, IIS-0748626, IIS-0844546, IIS-0612031, IIS-1050422, IIS-0905385, IIS-0746500, IIS-1209589, AGS-0835821, CNS-0751152, CNS-0855167).

References

[1] Howe, J. (2006). The Rise of Crowdsourcing. *Wired Magazine*, 14 (6), 1–4.

[2] Surowiecki, J. (2004). *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations.* Doubleday Books.

[3] Bonney, R., Cooper, C., Dickinson, J., Kelling, S., Phillips, T., Rosenberg, K., & Shirk, J. (2009). Citizen science: A developing tool for expanding science knowledge and scientific literacy. *BioScience*, 59(11), 977–984.

[4] Dickenson, J. L., Zuckerberg, B., Bonter, & D. N., *Citizen Science as an Ecological Research Tool: Challenges and Benefits* vol. 41. Palo Alto, CA, ETATS-UNIS: Annual Reviews, 2010.

[5] Trumbull, D., Bonney, R., Bascom, D., & Cabral, A. (2000). Thinking scientifically during participation in a citizen-science project. *Science Education*, 84(2), 265–275.

[6] Wiggins, A., Newman, G., Stevenson, R., & Crowston, K. (2011). Mechanisms for data quality and validation in citizen science. Presented at the IEEE eScience 2011 Workshop on Citizen Science.

[7] Delaney, D., Sperling, C., Adams, C., & Leung, B. (2007). Marine invasive species: Validation of citizen science and implications for national monitoring networks. *Biological Invasions*, 1573–1464.

[8] M. C. Fitzpatrick, E. L. Preisser, A. M. Ellison, & J. S. Elkinton. Observer bias and the detection of low-density populations. *Ecological Applications*, 19(7):1673–1679, 2009.

[9] J. Yu, W. K. Wong, and R. Hutchinson. (2010). Modeling experts and novices in citizen science data for species distribution modeling. Presented at the IEEE International Conference on Data Mining, Sydney, Australia.

[10] S. Kelling. (2011). Using Bioinformatics In Citizen Science. In *Citizen Science: Public Collaboration in Enviromental Research*, J. B. Dickinson, R. Bonney, Eds., ed: Cornell University Press.

[11] Law, E. & L. von Ahn. (2011). "Human Computation." *Synthesis Lectures on Artificial Intelligence and Machine Learning.* **5**(3): 1-121.

[12] Kelling, S., Yu, J., Gerbracht, J., &Wong, W. (2011). Emergent Filters: Automated Data Verification in a Large-scale Citizen Science Project. Presented at the IEEE eScience 2011 Workshop on Citizen Science.

[13] Kelling, S., Gerbracht, J., Fink, D., Lagoze, C., Wong, W., Yu, J., Damoulas, T. & Gomes, C. (2012). eBird: A Human/Computer Learning Network to Improve Biodiversity Conservation and Research. Presented at the annual meeting of the Association for the Advancement of Artificial Intelligence.