# Automated Data Verification in a Large-scale Citizen Science Project: a Case Study

Jun Yu[1], Steve Kelling[2], Jeff Gerbracht[2], Weng-Keen Wong[1]

[1]School of EECS
Oregon State University
Corvallis, OR USA
{yuju,wong}@eecs.oregonstate.edu

[2]Cornell Lab of Ornithology
Cornell University
Ithaca, NY USA
{stk2,jag73}@cornell.edu

*Abstract*— **Although citizen science projects can engage a very large number of volunteers to collect volumes of data, they are susceptible to issues with data quality. Our experience with eBird, which is a broad-scale citizen science project to collect bird observations, has shown that a massive effort by volunteer experts is needed to screen data, identify outliers and flag them in the database. The increasing volume of data being collected by eBird places a huge burden on these volunteer experts and other automated approaches to improve data quality are needed. In this work, we describe a case study in which we evaluate an automated data quality filter that improves data quality by identifying outliers and categorizing these outliers as either unusual valid observations or mis-identified (invalid) observations. This automated data filter involves a two-step process: first, a data-driven method detects outliers (ie. observations that are unusual for a given region and date). Next, we use a data quality model based on an observer's predicted expertise to decide if an outlier should be flagged for review. We applied this automated data filter retrospectively to eBird data from Tompkins Co., NY and found that that this automated process significantly reduced the workload of reviewers by as much as 43% and identifies 52% more potentially invalid observations.**

***Keywords-Applications, Citizen Science, Crowdsourcing, Data Quality, Data Filters, Species Distribution Modeling***

## I. INTRODUCTION

Citizen science enlists the help of volunteers from the general public (citizen scientists) in scientific research [1]. Data collected by citizen scientists can be achieved at little cost, enabling scientific research to gather data over longer periods of time and across broader spatial extents [5]. Engaging citizen scientists in meaningful projects also broadens the public understanding of the scientific process, which in turn can lead to better-informed decision making at all levels of society [23].

Recruiting volunteers in broad-scale citizen science projects to gather biodiversity information can generate enormous quantities of data across broad spatial and temporal domains [22]. However, maximizing the information gathered from citizen-science data depends on finding the proper balance between data quantity and quality [7]. Data quantity is essential because obtaining sufficient volumes of data of low per-datum information can contain as much information as data with high information content but gathered in smaller amounts [16]. Due to the importance of data quality, citizen science projects must take into consideration the ease of the data gathering process, the ability to limit data entry errors and identify questionable observations, and the offering of incentives to contributors to submit high-quality observations [26].

The most significant data quality issue in broad-scale citizen-science is individual variability in detection and classification of organisms to species. While large citizen-science projects can engage a broad network of tens of thousands of individuals contributing their observations, each participant has different identification skills. Data collected by inexperienced citizen scientists is often of lower quality due to their lack of expertise in accurately detecting and identifying organisms. On the other hand, data collected by experts is much more accurate though not completely free of mistakes.

Our strategy for addressing data quality in broad-scale citizen science is to examine contributions from many citizen scientists in a given geographic region. We can then identify general observational patterns for a specific geographic region, which allows for quality filters to emerge from the data. Furthermore, the expertise level of a citizen scientist can be used to screen that individual's contributions to a broad-scale citizen-science project. For instance, if an individual is a novice and he is frequently reporting rare, difficult-to-detect birds, then his records may be flagged more frequently for review than an expert's records. However, the expertise level of a participant needs to be estimated, which can be accomplished by using data mining techniques based on their historical observations. Thus, the challenge is to identify outliers (i.e. observations of a species that is unusual for a location or date) and categorize these outliers as either unusual valid observations, or mis-identified (invalid) observations.

In this work, we describe a case study in which we evaluate the effectiveness of an automated data filter using eBird (http://www.ebird.org) as our exemplar broad-scale citizen science project [22]. Our automated data filter combines two parts -- an emergent filter, derived from observed frequency patterns in the data, and a data quality model from our prior work [27] that predicts the observer's expertise. This observer expertise model was previously only evaluated on its ability to predict observer expertise and species observations. Our current work applies the observer expertise model from [27] for a different purpose – namely that of data quality control. The main contribution of our current work is the evaluation of the effectiveness of the overall automated data filter in a retrospective study using data from Tompkins Co., NY. This evaluation of data quality control, which was not performed in [27], is particularly challenging for eBird because there is no real ground truth (i.e. it is never known whether all species at a given location were detected and identified by the observer).

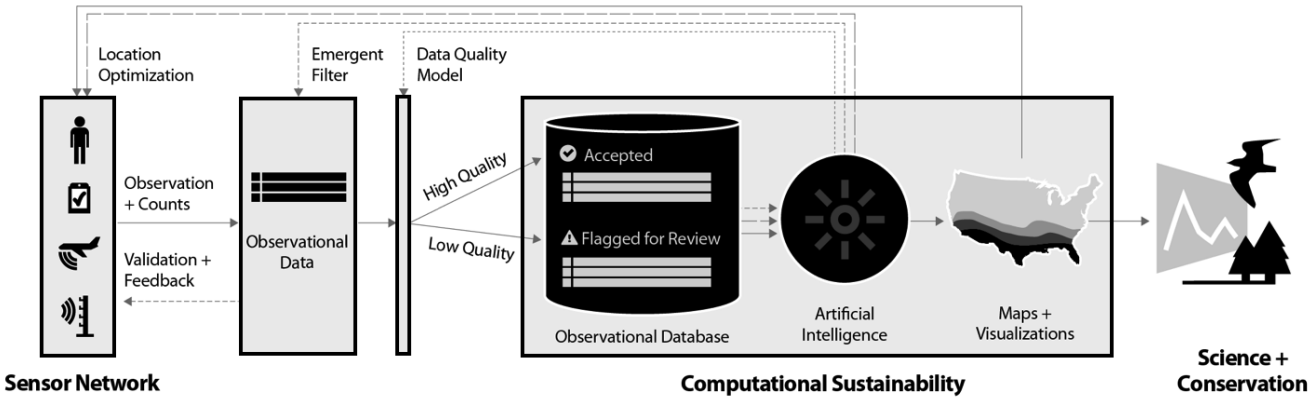


Figure 1. An overview of the eBird system.

Nevertheless, we will show that the observer expertise model combined with the emergent filter produces encouraging results for improving the data quality for a broad-scale citizen science project.

### *A. The eBird Project*

The eBird project [22, 26] is a citizen-science project that engages a global network of bird watchers to identify birds to species and report their observations to a centralized database. Anyone can submit their observations of birds to eBird via the web, and more than 50,000 individuals have volunteered over 4 million hours to collect more than 70 million bird observations for eBird from more than 200 countries; it is arguably the largest biodiversity data collection project in existence. Figure 1 depicts an overview of the eBird system.

eBird data contain information on the observer, location, visit, and species observed. Observer information such as name, ID and contact information allow every bird observation to be attributed to a specific person. Location data such as the site name, the coordinates where the observations were made and the geographic information, are stored with every visit to that location. Information about a specific visit consists of the date and time of the visit, the amount of effort expended and whether all the species observed were reported. Species observations consist of a checklist of birds observed and how many individuals of each species were counted.

eBird data reveal patterns of bird occurrence across space and through time, providing a data-rich foundation for understanding the broad-scale dynamic patterns of bird populations [7]. Recently, the United States Department of the Interior used eBird data as the basis for the 2011 State of the Birds Report, which estimated the occupancies of bird populations on public lands [17]. Figure 2 shows an example of a species distribution estimate of Western Meadowlark across the western US in June 2009 based on eBird observations.

Data quality is a major issue for the eBird project, particularly regarding an observer's ability to detect and correctly identify birds to the taxonomic level of a species. A network of bird distribution experts volunteer their time to create *expert-defined filters* to provide the basis of generating regional-specific checklists of birds for data submission. These experts have a thorough knowledge of the seasonal patterns of bird occurrence for a specific region. Based on this knowledge, a regional checklist filter delineates, when and how many of each species are expected in that region. If a contributor wants to submit a species that is not on the checklist they must take an active additional step to report a species that would not normally be expected and this record is flagged for review. Expert-defined filters can be for an area as large as a country or as small as a nature preserve. Presently eBird project employs more than 1200 expert-defined filters.

A network of more than 550 volunteers review flagged records in eBird. Areas covered by an editor range from an entire country (Central and South America), to a single county in parts of the US and Canada. The reviewers contact those individuals who submitted flagged records to obtain additional information, such as field notes or photographs, in order to confirm unusual records. In 2010, 4% (720k observations) of the 18 million observations submitted to eBird were flagged for review, and 1.5% (11k observations) were marked as invalid following review. All records, their flags and their review history are retained in the eBird database [9].

Depending on the region and time of year, an editor will review 15-1500 records per week (average of 200). About 80% of these records can be reviewed fairly quickly in 5-10 seconds. Other records require following up with the observer and asking for more details. While the process is semi-automated, it usually takes 2-5 minutes to review each record. About 1% of these records will require even more time, as the editor will follow up with the observer in a series of emails, explaining what species was more likely and the review process, and answering questions they may have. These communications between the reviewers and participants are the most frequent discussions eBird participants have with people from eBird.

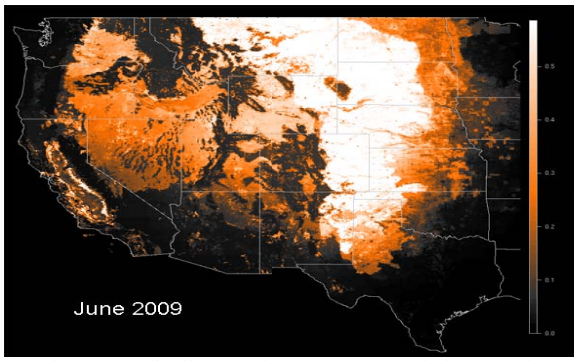


Figure 2. The species distribution of Western Meadowlark across the western US in June 2009. Western Meadowlark is a bird that prefers grasslands. Grassland birds are among the most consistently declining species in the US.

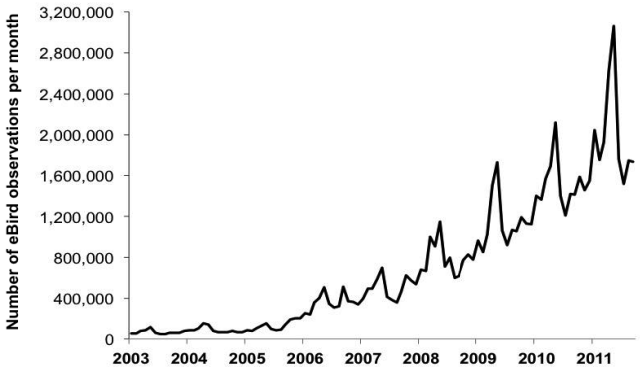


Figure 3. The total number of eBird observations per month since 2003.

Thus, there are three major data quality challenges faced by the eBird project. First, given the large variability of participants' expertise, invalid observations must be accurately identified to improve data quality. Second, the current expert-defined filters have generated an enormous volume of observations for review, which overwhelms the network of volunteer editors. This problem becomes even more severe with the tremendous growth of observations submitted to eBird as illustrated in Figure 3. Finally, due to the expansion of eBird, data filters need to be created for new regions without existing regional experts.

## II. RELATED WORK

The eBird project is an example of a crowdsourcing system [8] that engages a large number of people to perform tasks that automated sensors and computers cannot readily accomplish. Crowdsourcing systems have been successfully applied in computer vision [4] and natural language processing [21]. The Galaxy Zoo project [12] uses volunteers to classify images from the Sloan Digital Sky Survey.

One way to improve data quality in crowdsourcing systems is to have multiple observers annotate the same object. Dawid and Skeene [2] first addressed this problem in diagnostic tests and developed a general model by treating the ground truth as a latent variable. Smyth et al. [20] used a similar approach in labeling Venus images. Unlike previous work that focused on estimating the label of the objects, Raykar et al. [18] proposed an EM algorithm to estimate the label and to learn a classifier jointly. Whitehill et al. [25] took one step further by taking the difficulty of labeling an image into account. Welinder et al. [24] proposed a Bayesian generative model that can model an annotator's attributes, including expertise, competence and bias. In a different line of work, Sheng et al. [19] proposed to reduce uncertainty of the labels by repeated-labeling.

The eBird project has characteristics that make it distinct from other citizen science projects and thus require novel approaches for improving data quality. First, citizen scientists in the eBird project collect the actual data, rather than annotating data that has already been collected (eg. unlike in [12]). Since observers vary greatly in skill and effort expended, their observations cannot be simply accepted as ground truth. Second, eBird is a high-volume and low per-datum information system, meaning that a wider geographic area may have multiple observations from many observers.

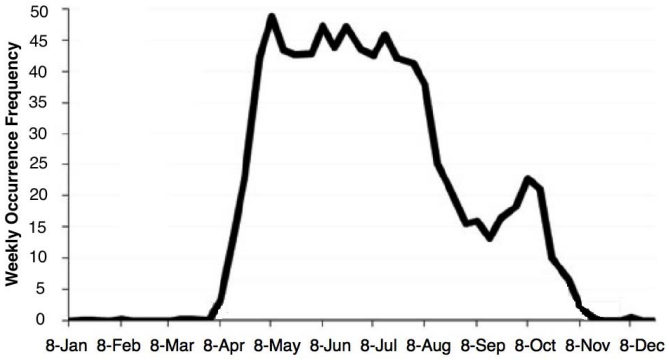


Figure 4. The weekly occurrence frequency for Chipping Sparrow in Tompkins Co., NY.

Although these observations are rarely from the exact same location and time (making the multiple observers strategy for improving data quality not viable for eBird), the broad spatial and temporal scale of eBird, allows the emergence of general observational patterns which can be effective for detecting outliers and assessing the reliability of the data collected.

The approach in this work is more broadly applicable to other citizen science projects that rely on citizen scientists, with a wide range of expertise levels, to collect a high volume of low per-datum information.

## III. METHODS

The automated data filter consists of the *emergent data filter* and the *data quality model*, both of which we will describe in the following subsections.

### A. Emergent data filter

We can use the large volume of historical observations from eBird as the basis for automatically generating regional checklists. We replace the expert-defined data filters with data-driven filters that emerge from the historical data to generate regional checklists and identify unusual observations.

The emergent data filter in eBird project is based on the frequency of reporting a species. The frequency is calculated as the number of checklists that reported the species divided by the total number of checklists submitted for a specific region. These frequencies are easily updated as new data is reported, and thus the emergent filters are constantly updated. The result is a measure of the *likelihood* of observing a specific species within that region. Since each observation contains details of where and when a bird was detected, we can calculate the frequencies of bird occurrence at any spatial level and for any date of year. Figure 4 shows an example of Chipping Sparrow's weekly occurrence frequency across a year window in Tompkins Co., NY.

For any specific region (e.g. county) and date, the emergent data filter automatically identifies unusual observations as follows: the frequency of occurrence estimates is made for all species that have been reported to eBird for that region and for that day of year. The frequency is then used to generate an online checklist by including all species whose occurrence frequency is past a threshold. The emergent data filter flags the observations falling below the threshold and processes the observations with the data quality model.

### B. Data quality model

The eBird data are provided by tens of thousands of observers with a wide range of expertise in identifying birds and with variable effort made in contributing to eBird. For example, at one extreme, several thousand observers with high identification skill levels contribute "professional grade" observations to eBird, whereas at the other extreme tens of thousands of participants contribute data of more variable quality. While there is much variability in the number of checklists that eBird volunteers submit, the top third of eBird contributors submit more than 90% of all data. Although the identification skills of this subset of contributors are unknown, it is probably skewed to the more skilled because individuals who regularly contribute tend to become better observers [6].

This inter-observer variation must be taken into account during analysis because valid outlier observations (i.e. those observations that are unusual but valid) could provide potentially important information on unique or changing patterns of occurrence. Since eBird engages a significant number of skilled observers who are motivated to detect rare species or are skilled in detecting elusive and cryptic species, being able to accurately distinguish their observations from those of less-skilled observers is crucial. The challenge is to obtain an objective measure of observer expertise that can be used to classify unusual observations.

In this case study, we investigate using a data quality model based on an observer's predicted expertise. In order to predict expertise, we use the Occupancy-Detection-Expertise (ODE) model from our prior work [27] because it was successful at accurately predicting an observer's expertise. The ODE model is a probabilistic graphical model [11] that extends a well-known model in Ecology called the Occupancy-Detection (OD) model [13]. We will provide an overview of the ODE model by first describing the OD model and then showing how to extend it with observer expertise in the ODE model. More details about the ODE model can be found in [27].

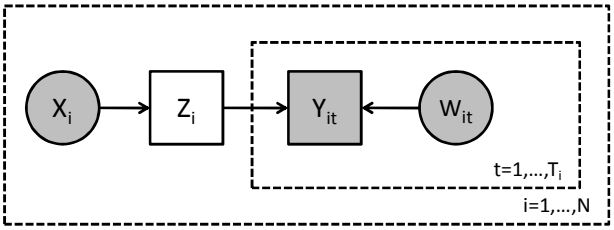

Figure 5. The graphical model representation of the OD model.

#### 1) Occupancy-Detection model

Ecologists often build species distribution models (SDM), which predict the occupancy of a site by a species. Occupancy determines if a geographic site is viable habitat for a species and it depends on environmental covariates such as temperature, precipitation and land use. A general form for SDMs is shown as the occupancy function $f^{occ}$ in Equation 1, in which the Boolean random variable $Z_i$ represents the occupancy of site $i$, $\boldsymbol{X_i}$ are the set of environmental covariates (i.e. features), and $\boldsymbol{\alpha}$ are the parameters of the model. Many different machine learning approaches have been used to model $f^{occ}(\boldsymbol{X_i}; \boldsymbol{\alpha})$. In our work, we use logistic regression.

$$P(Z_i = 1) = f^{occ}(\boldsymbol{X_i}; \boldsymbol{\alpha}) \quad (1)$$

Data for SDMs are often collected by observers who record occurrences of species in the field. Detecting the species in the field can pose a major problem, particularly if the members of the species are camouflaged, only appear at night, or are evasive. If an observer wrongly reports a species to be absent at a site when it was in fact present at that site, species distribution models built from such data will underestimate the true occupancy of that species for that site.

To address this issue, Mackenzie et al. [13] proposed the OD model, which models the relationship between occupancy and detection. In the OD model, the true occupancy of a site $i$ is represented as a latent variable $Z_i$ and each site is visited multiple times. Each visit $t$ results in an observation $Y_{it}$, corresponding to an observation of the species at site $i$ on visit $t$. The detection process is shown in Equation 2, where the detection function $f^{det}$ depends on a set of detection features $\boldsymbol{W_{it}}$, examples of which include the effort expended, the time of day, and the current weather conditions.

$$P(Y_{it} = 1) = Z_i \cdot f^{det}(\boldsymbol{W_{it}}; \boldsymbol{\beta}) \quad (2)$$

In Equation 2, the Boolean random variable $Y_{it}$ is influenced by both the true occupancy of a site and by the detection function $f^{det}(\boldsymbol{W_{it}})$, which we model using a logistic function with parameters $\boldsymbol{\beta}$.

The OD model makes two key assumptions. First, the population closure assumption [14] assumes that the species occupancy status at a site stays constant over the course of the visits. Second, the OD model does not allow the observers to report false detections. False detections occur when observers incorrectly report a species to be present at a site when in fact the species does not occupy that site.

#### 2) Occupancy-Detection-Expertise model

The detection process can also be affected by the expertise level of an observer. The ODE model extends the OD model by adding an expertise component. In this expertise component, as shown in Equation 3, the function $f^{exp}$ predicts $E_j$, which is a binary random variable representing the expertise of birder $j$ (i.e. taking values *expert* or *novice*), from a set of expertise features $\boldsymbol{U_j}$. Examples of expertise features include features derived from the birder's personal information and history of observations, such as number of years in bird watching, number of observations submitted to eBird and number of observations invalidated. We model $f^{exp}(\boldsymbol{U_j}; \boldsymbol{\gamma})$ using a logistic function with parameters $\boldsymbol{\gamma}$.

$$P(E_j = 1) = f^{exp}(\boldsymbol{U_j}; \boldsymbol{\gamma}) \quad (3)$$

The detection function in the ODE model is now influenced by the detection features $\boldsymbol{W_{it}}$ and by the expertise of the birder that supplied observation $Y_{it}$. Equation 4 illustrates the detection process in the ODE model, where the

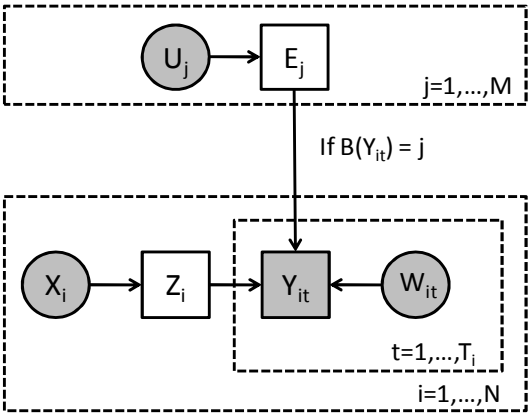

Figure 6. The graphical model representation of the ODE model.

notation $B(Y_{it})$ returns the index of the birder submitting observation $Y_{it}$.

$$P(Y_{it} = 1) = Z_i \cdot f^{det}(\boldsymbol{W_{it}}, E_{B(Y_{it})}, \boldsymbol{\beta}) \qquad (4)$$

By adding the expertise variable to $f^{det}(\boldsymbol{W_{it}}, E_{B(Y_{it})}, \boldsymbol{\beta})$, we distinguish the detection process of expert and novice birders. In effect, we use a mixture model in which one mixture component models the experts' detection process and one models the novices' detection process. Furthermore, in order to improve the model fit on the eBird data, we also modified the assumptions of the OD model to allow false detections by both experts and novices

Given the observations from a set of labeled expert and novice birders, we estimate the ODE model parameters using Expectation Maximization [3]. With the learned ODE model parameters, we can perform inference to predict a birder's expertise based on their checklists. In deployment, we update the ODE model annually.

## IV. Evaluation

### A. Data description

For this case study we analyzed eBird data from Tompkins Co., which is an average sized county (1,270 km$^2$) in the ecologically rich Finger Lakes Region of west-central NY. Participation in eBird is high in this county, with more than 48,000 checklists representing almost 700,000 observations. A regional expert developed a checklist filter for this county, which was the basis for all following comparisons. To evaluate the expert-defined filter and automated data filter, we applied both filters to eBird data collected from January 1, 2003 to June 23, 2011.

### B. Emergent data filter

To generate the emergent data filter, we calculated the frequency of occurrence based on all data reported for that species at the county level and date range, and compared with eBird submissions. This frequency was calculated as follows. First, a day-of-year value was assigned to each checklist ranging from 1 to 365, and then a raw daily frequency was associated with this day. However, there were large variations in the raw daily frequency, which ranged from 3 to 125 checklists. To account for this variation, we replaced each raw daily frequency with a value computed by taking the highest raw daily frequency of a day within a sliding 7-day window (3 days before to 3 days after the current day). In this study, we calculated the day-of-year frequencies for every species observed in Tompkins Co., NY based on eBird data gathered from 2003 to 2011. We used a frequency threshold of 5%.

### C. Estimating an observer's expertise level

For our analysis, we used observations from the original eBird Reference Data [15] from New York State during May-July in years 2009 and 2010. To train the ODE model, we used the observations from a list of birders with their expertise levels labeled. The eBird project leaders manually labeled the expertise of these birders using a variety of criteria including personal knowledge of the birder, the number of misidentified observations, the frequency of poor spatial accuracy in checklist submissions and manual inspection of their eBird submissions. There were a total of 134 expert birders and 229 novice birders used to train the ODE model.

We divided the checklists into training and test sets according to the observers submitting them. Birders that submitted checklists from Tompkins Co. in 2009 and 2010 were placed into an independent test set while labeled birders were placed into a training set. The test set consisted of 176 birders. We trained the ODE model and then used the trained model to predict the probability of a birder from the test set being an expert. To get a more reliable estimation of observer expertise, we applied the ODE model to 18 species (8 common species and 10 uncommon species) and the final expertise prediction was based on the average score over all 18 species.

## V. Results and Discussion

### A. Emergent data filter

In all cases examined, the expert-defined checklist filters for Tompkins Co. accepted observations over a broader temporal window than the emergent data filters. Three general categories for the expert-defined filter were apparent. First, an expert may have had a particular interest or knowledge of certain species and these filters could be very accurate (e.g. Figure 8B American Tree Sparrow Jan. - May). Second, the expert-generated filter may accurately describe the bird's biology, which may be quite different from what eBird contributors report. For example, Chipping Sparrows (see Figure 8A) are a common breeding bird in Tompkins Co., which are often found in close proximity to lawns and gardens, and have a very distinctive plumage and song. However, immediately after the breeding season (end of July) they stop singing, disperse, and begin to molt into a less distinctive plumage. They become more cryptic and harder to detect, which would lower the probability that they are reported to eBird. The final category included expert-defined filters that accepted observations, even when it was very unlikely that the bird would be encountered. For example, although expert filters allowed either Swamp or Savannah Sparrow to be reported for any month of the year in Tompkins Co., observations falling outside the typical pattern of occurrence, especially during winter, should be reviewed.

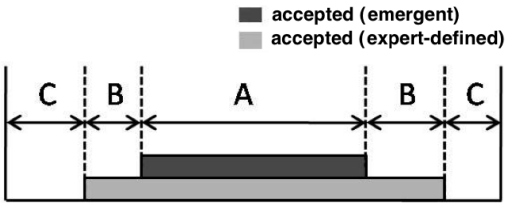


Figure 7. An illustration of the time periods covered by the expert-defined filter (light grey bar) and the emergent filter (dark grey bar). Observations falling within the bars are automatically accepted. Observations falling outside of the bars are flagged for review.

For the emergent data filter, the temporal resolution and the 5% threshold in frequency created a more conservative window of occurrence than that developed by the expert-defined filter. Since the emergent data filters were based on observer submissions, they matched the patterns of when most eBird volunteers reported a particular species for Tompkins Co. However, the emergent data filters significantly increased the number of flagged records. The emergent data filters flagged more than 35425 observations for review, compared to 4006 observations that were flagged by the expert-defined filters. We conclude that the emergent data filters set at a 5% cut-off accurately represented the patterns of reporting to eBird for the majority of observations, and allowed the easy identification of any outliers. However, it was a very conservative filter, which resulted in a significant increase in the number of flagged records that a regional editor must review. If the automated frequency filter alone were employed, it would lead to a greatly increased workload for the regional editors. One alternative for reducing the number of flagged records would be to make the filter less conservative (e.g. set the cutoff to be 3% of detections), but this would increase the possibility of allowing misidentifications to become part of eBird database.

Figure 7 represents a schematic of the expert-defined filter and emergent data filter for a single species. Observations falling outside of the bars were flagged for review. As was mentioned previously, in our data, the emergent filter was always a shorter window of acceptance than the expert-defined filter and was thus more conservative. The emergent and expert-defined filters in Figure 7 created three distinct regions labeled A, B, and C that we will use in our discussion of evaluation metrics.

Records falling in region A were not flagged by both filters and added to the eBird database without review. Since these records were not reviewed, we did not have information about the actual misidentifications in region A. However, the number of actual misidentifications in region A will be identical for both the emergent and expert-defined filters.

Second, records falling in region B were flagged by the emergent filter but not by the expert-defined filter. The region B corresponded to a time period in which misidentifications were common, such as after a particular bird species departed for migration and before their return. Since these records in Region B were also not reviewed in our retrospective analysis, we did not have ground truth about actual misidentifications.

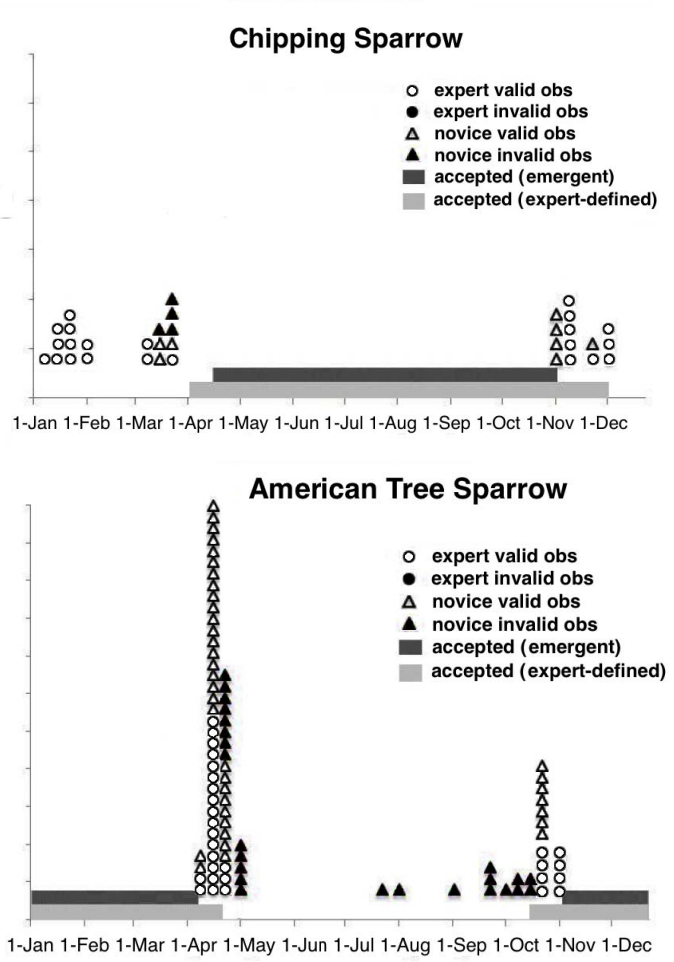


Figure 8. Flagged observations for (A) Chipping Sparrow and (B) American Tree Sparrow in Tompkins Co., NY. The time periods of the emergent filter (dark grey) and the expert-defined filter (light grey) are shown as horizontal bars on the bottom. Any observations falling within the time period indicated by the emergent filter were automatically accepted into eBird and are not shown on this graph. Observations falling outside of the emergent filter were flagged for review and are shown as triangles (from novices) or circles (from experts). Valid observations are shaded black while invalid observations are white.

Finally, records falling in region C were flagged by both filters. Unlike for regions A and B, these records were in fact reviewed by experts and then designated to be either valid and added to eBird or designated invalid and discarded. We used the validity of these records as a measure of the accuracy of a filter in region C.

*B. The automated data filter*

In Figures 8A and 8B, we illustrate the ODE model predictions of expertise in relation to records flagged by the two filters. What is most striking is how individuals with a low level of eBird expertise tended to report both American Tree Sparrow and Chipping Sparrow outside their typical windows of occurrence more frequently. These two species are very similar looking sparrows that are attracted to bird feeders and easily observed. Many inexperienced observers confuse these species, and misidentification is a problem particularly at their first seasonal arrival. The observers of low

predicted expertise reported more American Tree Sparrows earlier in fall than observers of high predicted expertise, and their observations fell outside the general patterns of the frequency graphs. This example shows the significant contribution that the automated filter process could have for identifying outlier reports for birds that are relatively common, and which would normally pass as valid records under the expert-defined filter model.

Table 1 provides examples of a variety of bird species with difference occurrence patterns in Tompkins Co. and the percent of expert/novice observations that were flagged by the emergent filter. These results indicate that expert observers tend to identify more unusual birds than novice observers. Use of the automated data filter would significantly reduce the number of flagged records that must be reviewed since it accepts records from expert observers.

Table 1. Example of 9 bird species with different occurrence patterns in Tompkins Co, NY. The second and third columns are the percentage of expert/novice observations flagged by the emergent filter.

| Bird Species | Number of Observations | % Expert | % Novice |
|---|---|---|---|
| Common Raven[1] | 448 | 96 | 4 |
| Pine Siskin[2] | 128 | 97 | 3 |
| Acadian Flycatcher[3] | 61 | 82 | 18 |
| Savannah Sparrow[3] | 86 | 79 | 21 |
| Black-throated Blue Warbler[3] | 128 | 77 | 23 |
| Cerulean Warbler[4] | 91 | 65 | 35 |
| Wilson's Warbler[5] | 65 | 77 | 23 |
| Ruby-crowned Kinglet[5] | 41 | 83 | 17 |
| Hermit Thrush[6] | 162 | 88 | 12 |

[1]Species that occur year round at frequencies below the emergent filter.
[2]Species that occur periodically in the county.
[3]Species that are locally common breeders in the county.
[4]Species that are locally uncommon breeders in the county.
[5]Migrant species that are locally common when they pass through the county.
[6]Species that are locally common breeders and uncommon throughout the year.

Table 2 shows the number of flagged records from all three filters. The emergent data filter significantly increased the total number of observations for review to 35425, but when the emergent filter was combined with the ODE model in the automated data filter, the number of flagged records decreased by 93% to 2303. When compared to the expert-defined filter, the automated data filter decreased the number of flagged records by as much as 43%, showing the potential of the automated data filter for substantially reducing the workload of reviewers. Under current expert-defined filters, each reviewer spends approximately 5 hours per week reviewing flagged records; this cost reduces to 2.85 hours (i.e. 2.15 hours saving per week) with the automated data filter. These savings become even larger due to the fast growth of eBird project.

Table 2. The number of flagged records from Tompkins Co., NY and the estimated number of hours needed to review them.

| Filter Type | Number of Flagged Records | Estimated number of hours to review |
|---|---|---|
| Expert-defined filter | 4006 | 101 hrs |
| Emergent data filter | 35425 | 890 hrs |
| Automated data filter | 2303 | 58 hrs |

Although the automated data filter can substantially reduce the number of records to be reviewed, it must also not carelessly discard any truly erroneous records that should indeed be reviewed. In order to measure the accuracy of the automated data filter in our retrospective analysis of eBird data from Tompkins Co., we compared how many of the records in region C were designated as valid or invalid after being reviewed. Figure 9 (left) illustrates the fraction of valid and invalid records among all the records in region C, and then the amounts broken down by experts and novices (middle and right pie charts). Only 137 flagged records (5%) from experts were invalid, while 848 records (65%) from novices were invalid. The automated data filter would have allowed the 137 flagged expert records to pass through, but all 848 novice records would have been flagged.

The analysis above only covers region C and gives a partial picture as to the accuracy of the automated data filter as the 2303 records flagged by the automated data filter are in both regions B and C. Records in region B were not reviewed, and as a result, we did not have any ground truth as to their validity. However, we can estimate the number of invalid records by making the assumption that, as in region C, novices submitted invalid records 65% of the time. This assumption is conservative because misidentifications tend to increase in region B as compared to region C. Under this assumption, 65% of the 2303 records flagged by the automated data filter were invalid (i.e. 1497 records). This amount is higher than the 985 invalid observations flagged by the expert-defined filter by as much as 52%, thus showing how effective the automated data filter is at identifying truly erroneous outliers.

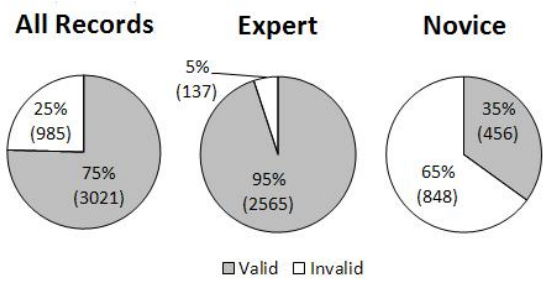


Figure 9. The fraction of valid and invalid records among all records in region C (left), then broken down by expert records (center) and by novice records (right). The number of records in each pie slice is shown in parentheses.

## VI. CONCLUSION

Data quality is a major challenge in any sensor network, especially when the sensor network consists of a massive number of volunteer observers that have differing abilities to accurately identify birds. This paper assessed the performance of a more automated process for addressing a major data quality need in broad-scale citizen-science projects: filtering misidentified organism occurrences. Our automated data filter was based on both the patterns of submissions within a predefined spatial and temporal extent, as well as the contributor's skill level.

We presented the results of applying the automated data filter retrospectively to historical records from Tompkins Co., NY. Our automated data filter allowed us to reduce the workload of reviewers by about 43% as compared to the existing expert-defined filter, which results in about 2.15 hours savings per week for a reviewer. The automated data filter also identified as many as 52% more invalid outliers than the expert-defined filter. These results demonstrate that our automated process has the potential to play a critical role in improving data quality in broad-scale citizen-science projects.

For future work, we will test these results more broadly across the US. In addition, we will model an individual's expertise regionally, as a birder can be an expert observer in their home region, but less so outside of that region.

## ACKNOWLEDGMENT

We thank D. Fink, C. Wood, M. Iliff and B. Sullivan for improving the clarity of our article. This work was funded by the Leon Levy Foundation, Wolf Creek Foundation and the National Science Foundation (Grant Numbers OCI-0830944, CCF-0832782, ITR-0427914, DBI-1049363, DBI-0542868, DUE-0734857, IIS-0748626, IIS-0844546, IIS-0612031, IIS-1050422, IIS-0905385, IIS-0746500, AGS-0835821, CNS-0751152, CNS-0855167).

## REFERENCES

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

[2] A. P. Dawid and A. M. Skene, Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society, Series C*, 28(1): 20-28, 1979.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1): 1-38, 1977.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li, ImageNet: a large-scale hierarchical image database. In *Proceedings of Computer Vision and Pattern Recognition*, pp. 248-255, 2009.

[5] J. L. Dickenson, B. Zuckerberg, and D. N. Bonter, Citizen science as an ecological research tool: challenge and benefits. *Annual Review of Ecology Evolution and Systematics*, 41(1): 149-172, 2010.

[6] K. A. Ericsson and N. Charness, Expert performance: its structure and acquisition. *American Psychologist*, 49: 525-747, 1994.

[7] W. Hochachka, D. Fink, R. Hutchinson, D. Sheldon, W.-K. Wong, and S. Kelling, Data-intensive science applied to broad-scale citizen science. *Trends in Ecology and Evolution*, 27(2): 130-137, 2012.

[8] J. Howe, Crowdsourcing: why the power of the crowd is driving the future of business. Crown Business, New York, 2008.

[9] S. Kelling, Using bioinformatics in citizen science. *Citizen Science: Public Collaboration in Environmental Research*, Cornell University Press. Ithaca, NY, 2011.

[10] S. Kelling, J. Yu, J. Gerbracht, and W.-K. Wong, Emergent filters: automated data verification in a large-scale citizen science project. In *Proceedings of the e-Science 2011 Computing for Citizen Science Workshop*, pp. 20-27, 2011.

[11] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press, Cambridge, MA, 2009.

[12] C. J. Lintott, K. Schawinski, S. Anze, K. Land, S. Bamford, D. Thomas, M. J. Raddick, R. C. Nichol, A. Szalay, D. Andreescu, P. Murray, and J. Vandenberg, Galaxy Zoo: morphologies derived from visual inspection of galaxies from the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society*, 389: 1179-1189, 2008.

[13] D. I. MacKenzie, J. D. Nichols, J. E. Hines, M. G. Knutson, and A. B. Franklin, Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly.*Ecology*, 84(8): 2200-2207, 2003.

[14] D. I. MacKenzie, J. D. Nichols, G. B. Lachman, S. Droege, J. A. Royle, and C. A. Langtimm, Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, 83(8): 2248–2255, 2002.

[15] M. A. Munson, K. Webb, D. Sheldon, D. Fink, W. M. Hochachka, M. Iliff, M. Riedewald, D. Sorokina, B. Sullivan, C. Wood, and S. Kelling, *The eBird reference dataset, version 3.0*. Cornell Lab of Ornithology and National Audubon Society, Ithaca, NY, December 2011.

[16] M. A. Munson, R. Caruana, D. Fink, W. M. Hochachka, M. Iliff, K. V. Rosenberg, D. Sheldon, B. L. Sullivan, C. Wood, and S. Kelling, A method for measuring the relative information content of data from different monitoring protocols. *Methods In Ecology and Evolution*, 1(3): 263-273, 2010.

[17] N. A. B. C. I. U. S. The state of the birds 2011 report on public lands and waters. Washington D.C., 2011.

[18] V. C. Raykar, S. Yu, L. H. Zhao, A. Jerebko, C. Florin, G. H. Valadez, L. Bogoni, and L. Moy, Supervised learning from multiple experts: whom to trust when everyone lies a bit. In *Proceedings of the 26th International Conference on Machine Learning*, pp. 889-896, 2009.

[19] V. S. Sheng, F. Provost, and P. G. Ipeirotis, Get another label? improving data quality and data mining using multiple noisy labelers. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 614-622, 2008.

[20] P. Smyth, U. Fayyad, M. Burl, P. Perona, and P. Baldi, Inferring ground truth from subjective labeling of venus images. In *Advances in Neural Information Processing System*, pp. 1085-1092, 1995.

[21] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng, Cheap and fast - but is it good? : evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 254-263, 2008.

[22] B. L. Sullivan, C. L. Wood, M. J. Iliff, R. Bonney, D. Fink, and S. Kelling, eBird: a citizen-based bird observation network in the biological sciences. *Biological Conservation*, 142(10): 2282-2292, 2009.

[23] D. J. Trumbull, R. Bonney, D. Bascom, and A. Cabral, Think scientifically during participation in a citizen-science project. *Science Education*, 84(2): 265-275, 2000.

[24] P. Welinder, S. Branson, S. Belongie, and P. Perona, The multidimensional wisdom of crowds. In *Advances in Neural Information Processing Systems 23*, pp. 2424-2432, 2010.

[25] J. Whitehill, P. Ruvolo, T. F. Wu, J. Bergsma, and J. Movellan, Whose vote should count more: optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems 22*, pp. 2035-2043, 2009.

[26] C. Wood, B. Sullivan, M. Iliff, D. Fink, and S. Kelling, eBird engaging birders in science and conservation. *PLoS Biol*, 9(12): e1001220, 2011.

[27] J. Yu, W.-K. Wong, and R. A Hutchinson, Modeling experts and novices in citizen science data for species distribution modeling. In *Proceedings of the 2010 IEEE International Conference on Data Mining,* pp. 1157-1162, 2010.