Emergent Filters: Automated Data Verification in a Large-scale Citizen Science Project

Steve Kelling
Cornell Lab of Ornithology, Ithaca, NY USA
stk2@cornell.edu

Jun Yu
School of EECS Oregon State University Corvallis, OR USA
yuju@eecs.oregonstate.edu

Jeff Gerbracht
Cornell Lab of Ornithology, Ithaca, NY USA
jag73@cornell.edu

Weng-Keen Wong
School of EECS Oregon State University Corvallis, OR USA
wong@eecs.oregonstate.edu

***Abstract*—Research projects that use the efforts of volunteers ("citizen scientists") to collect data on organism occurrence must address issues of observer variability and species misidentification. While citizen science projects can engage a very large number of volunteers to collect volumes of data, they are prone to contain reporting errors. Our experience with eBird, a citizen science project that engages tens of thousands of volunteers 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. But the increasing volume of data being collected by eBird places a huge burden on these volunteer experts. In order to minimize this human effort, we explored whether previously collected eBird data can be used to create automated quality filters that emerge from the data. We do this through a two-step process. First a data-based method detects outliers (i.e., observations that are unusual for a given region and week of the year). Next, a novel machine learning method that estimates observer expertise is used to decide if the unusual observation should be flagged or not. Our preliminary findings indicate that this automated process reliably identifies outliers and accurately classifies them as either an error or represents a potentially valuable observation.**

Keywords- citizen-science; data quality; database filters; species occurrence; machine learning.

I. INTRODUCTION

Citizen-science projects enlist the public in scientific endeavors [1], and often provide data with more intensive sampling through longer periods of time and across broader spatial extents. Citizen scientists can be motivated to collect large amounts of data at relatively little cost to the scientific research enterprise [2]. Furthermore, engaging citizen scientists in meaningful projects has the added benefit of broadening the public understanding of the scientific process, which in turn can lead to better-informed decision making at all levels of society [3].

To realize the great potential of citizen science, researchers must address many challenges in the interpretation and use of broad-scale, volunteer-collected data. For example, volunteers have a wide range of expertise in their ability to identify organisms. Additionally, observers may fail to detect some organisms altogether, and hence report "false absences". These, as well as other characteristics of citizen-science data, make their accurate use challenging, and must be considered when designing citizen-science projects or when analyzing their data.

In this paper we address the challenge of identifying organism misidentifications in broad-scale citizen-science projects. Traditionally, if data are screened in citizen-science projects, experts are engaged. However, when large-scale citizen science projects gather massive quantities of data, the workload on these experts can be overwhelming. In this paper we describe an automated system in which spatial and temporal data filters emerge from existing data. First, previously collected data is used to identify outliers. Next, an estimate of the observer's ability based on their previous submissions determines if an unusual observation is flagged for further review. We demonstrate the use of this emergent filter data vetting system using bird observations from eBird (http://www.ebird.org).

II. eBIRD

eBird is an online checklist program that gathers tens of millions of bird observations annually from a global network of volunteers [4]. eBird is amassing one of the largest data resources on species distributions; more than 60 million species observations from more than 500 thousand locations globally and is growing at a rate of 25% annually. In 2010 over 22 thousand participants volunteered more than 1 million hours collecting almost 18 million bird observations. All eBird data are openly available (http://www.avianknowledge.net) and shared with numerous global biodiversity initiatives.

eBird data reveal patterns of bird occurrence across space and through time, and are providing a data-rich foundation for understanding the broad-scale dynamic patterns of bird populations [5-14]. Recently, the United States Department of the Interi-

or used eBird data as the basis for the 2011 State of the Birds Report, which estimated the occupancies of bird populations on public lands [15].

eBird's data contain the following information: observer, location, visit, species, and number observed. Basic information about the observer such as name 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 area represented by the location are stored with every visit to that location. Information about a specific visit consists of date and time of visit, amount of effort expended, such as distance traveled, time spent and area covered, and whether or not all the species observed were reported. Species observations consist of a checklist of birds observed and how many individuals of each species were counted. These data form the core of the eBird relational database [16].

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 bird distribution experts volunteer their time to create *expert-defined checklist filters*. These experts have a thorough knowledge of the seasonal patterns of bird occurrence for a specific region. Based on this knowledge, each expert creates a regional checklist filter that delineates, by month, when and how many of each species are expected in that region. The checklist of birds available on the eBird data entry form is based on these filters, and only those species expected in a specific geographical region at a specific time of year, are available. 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. *Expert-defined checklist filters* can be for an area as large as a country, or as small as a nature preserve. Presently eBird employs more than 1200 filters.

A network of more than 450 volunteers review flagged records in eBird. The reviewers are knowledgeable about bird occurrence for a region, and many also established the *expert-defined checklist filters*. 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% (720,000 observations) of the 18 million observations submitted to eBird were flagged for review, and 1.5% (10,800 observations) were marked as invalid following review. All records, their flags and their review history are retained in the eBird database.

The challenge that eBird now faces is that he project's success has generated an enormous volume of new observations to be reviewed (e.g., more than 3 million observations in May 2011) and the network of volunteer regional reviewers is being overwhelmed by the number of records needing to be reviewed. In order to address this issue we tested whether a data quality filter and screening process could be automated, and consider three questions:

1. Emergent Filters: Can historic data submissions to eBird be used to automatically generate accurate checklist filters?
2. ODE Model: Can we use an occupancy-detection model to rank observer ability and use these scores to improve data quality?
3. Do the Emergent Filters and ODE model improve data quality processes?

To answer these questions we describe an approach where existing eBird data establish count limits and generate daily regional checklist filters, and a machine-learning algorithm categorizes a contributor's ability. We then test this approach within a limited eBird region to compare the automated with the expert data quality process.

III. METHODS

For this preliminary study we analyzed eBird data from one county in New York State. Jefferson County, located at the border of Lake Ontario and the Saint Lawrence River, is a large (4,810 km²) ecologically rich and diverse county. It was selected for this study because it has reasonably good year-round coverage from a wide spectrum of users. More than 6,000 checklists representing over 73,000 observations were submitted from Jefferson County as of July 1, 2011.

A regional expert developed a checklist filter for Jefferson and the surrounding counties, which was the basis for all following comparisons.

A. *Emergent Data Filter*

One of the most common and powerful calculations performed on eBird data is a measure of how frequently a species is reported (Figure 1), which is calculated using the number of checklists that reported the species divided by the total number of checklists submitted for a specific region. The result is a measure of the "likelihood" of observing a specific bird 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.

For this study we compared eBird submissions with the calculated frequency of occurrence based on all data reported for that species at the county level and date range. We calculated day of year frequencies for every species observed in Jefferson County, New York based on eBird data gathered between

2003 and 2011. First a day of year value was assigned to each checklist ranging from 1 to 365. Each day had as many as 125 checklist submissions, or as few as 3 checklist submissions. To account for this variation in the number of checklists per day the frequencies were calculated based on a sliding 7-day window. The frequency for day X was calculated using a total number of checklists from 3 days prior through 3 days after day X. We then assigned the highest initial frequency within that same sliding 7-day window to day X. The resulting frequency is an estimate of the likelihood of observing a species on each of the 365 days of the year.

B. *Ranking observer ability*

eBird data are provided by thousands of observers with a wide range of expertise in identifying birds, and 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. While the identification skills of this subset of contributors is unknown, it is probably skewed to the more skilled because individuals who regularly contribute tend to become better observers [17].

This inter-observer variation must be taken into account during analysis because outlier observations (i.e., those observations that are unusual) 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.

To do this we developed a model to estimate observer expertise. Ecologists are frequently interested in the viability of a site as suitable habitat for a species, and have developed models to predict the occupancy of a site by a particular species given a set of environmental covariates describing that site. A general form for these models is shown in Equation 1, where $\boldsymbol{v}_l$ is a set of environmental covariates for location l, $\boldsymbol{z}_l$ represents the occupancy of location l and $f^{occ}(v_l)$ is the function capturing the occupancy model (see Table 1 for notation description).

$$\Pr(z_l = 1) = f^{\text{occ}}(\boldsymbol{v}_l) \qquad (1)$$

Many different approaches have been used to model $f^{occ}(v_l)$ including GLMs/GAMs [18], Maximum Entropy models [19], and boosted regression trees [20].

Many species are difficult to detect for a variety of reasons such as camouflage, nocturnal behavior, and evasiveness. If a species is erroneously reported to be absent at a site when it was in fact present at that site, then SDMs built from such data will underestimate the true occupancy of that species for that site. To address this issue, Mackenzie et al. [21] proposed an Occupancy-Detection (OD) model where true occupancy of a site l is represented as a latent variable $\boldsymbol{z}_l$. Under the OD model, a site is visited multiple times. Each visit i results in an observation $\boldsymbol{y}_i$, where the observation process is influenced by the true occupancy of the site and by a function $f^{\text{det}}(\boldsymbol{w}_i)$, where $\boldsymbol{w}_i$ are detection covariates (under the notation of Mackenzie et al. [21], $\psi = f^{\text{occ}}(\boldsymbol{v}_l)$ and $p = f^{\text{obs}}(\boldsymbol{w}_i)$). Equation 2 summarizes the process:

$$\Pr(y_i = 1) = z_{l(i)} \cdot f^{\text{obs}}(\boldsymbol{w}_i) \qquad (2)$$

$z_l \in (0, 1)$	The occupancy of location l by the species of interest.
$y_i \in (0, 1)$	The detection/non-detection of the species of interest in observation i.
$e_b \in (nov., exp.)$	The expertise of the observer b
$\boldsymbol{v}_l$	A vector of environmental covariates for location l.
$\boldsymbol{w}_i$	A vector of covariates describing the observation process for observation i.
$\boldsymbol{u}_o$	A vector of expertise covariates for observer o.
$l(i)$	The location of observation i.
$o(i)$	The observer that recorded observation i

Table 1: Terms and notation used for the Occupancy-Detection and Occupancy-Detection-Experience models.

The OD model makes two key assumptions. First, it assumes population closure in which the true occupancy of a site $\boldsymbol{z}_l$ remains unchanged over the multiple visits to that site. Second, the OD model assumes that observers do not report false positives

(i.e., an observer does not mistakenly report a species to be present when it is in fact absent).

The eBird experience level of an observer, which is the combination of their ability in identifying birds and their level of participation in eBird, can also influence the observation process. As a result we extended the OD model with an eBird experience component resulting in the Occupancy-Detection-Experience (ODE) model. In this extension, we add a new latent variable $E_{o(i)}$ and associated function $f^{\exp}(\boldsymbol{u}_{o(i)})$ which capture the experience level (ie. eBird experience rated as *high* or *low*) of the observer $o(i)$ that recorded observation i.

As shown in Equation 3, this experience variable is a function of a set of covariates $\boldsymbol{u}_{o(i)}$ that include characteristics of the observer such as the total number of checklists submitted and relative to the total number of species reported, and the total number of flagged records rejected. As shown in Equation 4, the observation process is now influenced by the true occupancy of a site and by the function $f^{obs}(\boldsymbol{w}_i, e_{o(i)})$, which is now a function of the observation covariates.

$$\Pr(e_{o(i)} = 1) = f^{\exp}(\boldsymbol{u}_{o(i)}) \quad (3)$$

$$\Pr(y_i = 1) = z_{l(i)} \cdot f^{\text{obs}}(\boldsymbol{w}_i, e_{b(i)}) \quad (4)$$

The ODE model relaxes the assumptions of the OD model by allowing false positives by the observers, for both levels of expertise. More details about the ODE model can be found in [22].

The automatic prediction of a contributor's eBird experience level can provide additional information that can be used to classify submissions. To test this, we used the ODE model to identify the eBird experience level based on the observer's checklist history, which included the total number of checklists they submitted, the total number of birds identified, and the total number of rejected records. We choose the breeding season because many bird species are more easily detected during breeding.

For this experiment we considered each bird species as a different prediction problem, and made our observer evaluations based on identification of five common bird species easily detected and identified by novices and experts alike, and five bird species that are difficult to detect and identify (Table 2).

To train and test our models, we divided all checklists according to the observers submitting them into either training or test data sets. eBird project staff manually labeled the eBird experience level of birders in our training set using a variety of criterion including personal knowledge of birder reputation, number of checklists rejected during data verification, and manual inspection of eBird checklists. Data from 100 expert birders and 208 novices birders were used to train the ODE model. Birders that submitted checklists from Jefferson Co. in 2009 and 2010 were placed into an independent test set while all other birders were placed into a training set. The Jefferson Co. test set consisted of 36 birders. We trained the ODE model and then used the trained ODE model to predict birder expertise from the Jefferson Co.

Common Bird Species	*Uncommon Bird Species*
Wild Turkey (*Meleagris gallopavo*)	Willow Flycatcher (*Empidonax traillii*)
Pileated Woodpecker (*Dryocopus pileatus*)	Red-eyed Vireo (*Vireo olivaceus*)
Blue Jay (*Cyanocitta cristata*)	Veery (*Catharus fuscescens*)
Black-capped Chickadee (*Poecile atricapillus*)	Savannah Sparrow (*Passerculus sandwichensis*)
American Robin (*Turdus migratorius*)	Swamp Sparrow (*Melospiza georgiana*)

TABLE 2. Bird species used to categorize identification eBird participants. Common bird species were those that were easy to observe and identify, attracted to humans, or abundant. Uncommon bird species were those that were best identified by vocalizations, or hard to observe, and present numerous identification challenges.

C. *Automated Data Quality Process*

To test the accuracy of the automated data quality filter, a 2-step process was developed. First, we set the *emergent data filters* at 5% of total frequency as a threshold level to identify all outlier observations. Next, the ODE model score was used to identify whether the observer had a high level of eBird expertise. If so, their records were accepted. However, if the model identifies them with a low level of eBird expertise, their records would be marked for review.

IV. RESULTS

Figure 1 shows the results of the two-step data quality experiment for 6 exemplar species.

A. Expert-defined versus Emergent data filters

In all cases examined, the *expert-defined checklist filters* for Jefferson Co. accepted observations over a broader temporal window than *emergent data filters*. Three general filter categories were apparent. First, an expert may have a particular interest or knowledge of certain species and the data filters can be very accurate (e.g., Fig. 1.A 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 (e.g., Fig. 1.B Chipping Sparrow). Chipping Sparrows are a common breeding bird in Jefferson Co., 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 get reported to eBird. The final category of filters included *expert-defined filters* that accepted observations, even when it was very unlikely that the bird would be encountered. For example, the expert filters allowed either Swamp or Savannah Sparrow (Fig. 3 C and D) to be reported for any month of the year in Jefferson Co. While it is certainly possible for either to occur in Jefferson Co. year-round, observations falling outside the typical pattern of occurrence (e.g., breeding season), and especially in winter, should be reviewed. In fact, for both sparrow species there have been no reports to eBird prior to mid-April.

For *emergent data filters* the temporal resolution and the 5% limit in total frequency of reports created a more conservative window of occurrence than that developed by the expert. Since the *emergent data filters* are based on observer submissions they, by definition, match the patterns of when most eBird volunteers reported a particular species for Jefferson Co. (Fig. 1). However, the *emergent data filters* significantly increased the number of flagged records; the *emergent data filters* flagged more than 3,000 observations for review, compared to 750 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 allow the easy identification of any outliers (Fig. 1 light and dark circles). However, it is a very conservative filter, which results 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. While one solution would be to set the cut-off for the filter (e.g., 2% or 3% of detection) could cut back on the amount of review, but would increase the possibility of missing misidentifications to become part of the eBird database.

B. ODE Model Results

The ODE Model ranked all individuals who submitted observations to eBird from Jefferson Co. Of the total of 36 individuals reported observations to eBird from Jefferson Co. 10 individuals were ranked as having a high level of eBird expertise, and the rest with a low level. We next plotted all records that were flagged by the *emergent data filters* as light circles (high level of expertise) and dark circles (low level of expertise) on the plots (Fig. 1).

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 (Figure 1 A and B). This example identifies the potential significance of a 2-level automated data quality filter. The American Tree Sparrow and Chipping Sparrow 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. Those observers who had low ODE scores reported American Tree Sparrows earlier in fall than observers with high ODE scores, and their observations fell outside the general patterns of the frequency graphs. This example shows the significant contribution that the emergent 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.

In several instances the ODE model did not categorize an individual's eBird expertise level similar to how their peers perceive their bird identification skills. For example, several of the individuals who had low ODE scores were known to be very good at bird identification. However, the ODE model did a very good job in identifying individuals who were known to have modest bird identification skills. The ODE model also incorporates the number of checklist submissions that a user contributes to a region, and in so doing does not rank observers solely on their skills in bird identification, but also in their level of participation in eBird in a given region. Weighting eBird expertise over simple identification expertise has distinct advantages for our models, which depend upon users understanding and correctly reporting all data required for an eBird observation (e.g., location, effort etc).

C. A two-step automated filter

The *emergent data filters* combined with the ODE model results an automated data quality filter for eBird appears promising. When this 2-step process was compared to the results of the *expert-defined* filters for Jefferson Co. a comparable number of records requiring review were obtained. Thus, while the total number of records requiring review did not decline, those records that were flagged by the automated filter identified more subtle potential misidentifications.

V. CONCLUSION

Data quality is a major challenge in any sensor network. This is especially true when the sensor network consists of a massive number of volunteer observers that have differing abilities to accurately identify birds. To address one aspect of the data quality challenge we created a two-step automated process that first would reliably *identify* outliers, and then *classify* those outliers either as an error or something

real. What is unique about our approach is that we do not remove outliers, but instead classify them as either unusual or erroneous. We can do this because of sufficient domain knowledge (e.g., understanding of the patterns of bird distributions) to distinguish between these 2 classes of outliers, and sufficient data so that a quality filter can emerge from the data.

The classification of outliers is a major data quality issue in all citizen-science projects, and is seldom addressed and often ignored. One project that has addressed this issue is Galaxy Zoo (http://zool.galaxyzoo.org), which used volunteers to classify objects from images made during the Sloan Digital Sky Survey. Individual variation was addressed by having many individuals classify the same object. While it may appear that a Galaxy Zoo approach to data quality control would not work when volunteers are actively collecting data across a broad spatial and temporal landscape. However, by classifying observations across specific geographic regions (in this case a county in New York State) we can identify general observation patterns, which allows for quality filters to emerge from the data. While Galaxy Zoo harnesses multiple observers to categorize a single object, eBird is able to use multiple checklists from many observers in a given geographic region to categorize outliers. Our findings show that this process could significantly improve the ability to identify outlier observations and categorize them as either true identifications, or false misidentifications.

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. In this paper we specifically addressed three questions.

A. Can previous data submissions be used as an to automatically generate accurate data filters?

We found that using historic data provided a higher level of selectivity than expert-defined filters. However, *data-defined checklist filters* set at a 5% acceptance level is more conservative, and generated a significantly higher number of flagged records for editors to review.

B. Can we rank observer ability and use these scores in the data validation process?

The ODE model resulted in the accurate classification of contributors on their general level of eBird experience.

C. Does this automated data quality functionality improve data quality processes?

The combination of the *emergent checklist filters* with the ODE model provided the best strategy for analyzing species reports in eBird. This is particularly true for common or moderately common species of birds (Fig. 1). The two step approach, where the emergent data filters are used to identify outliers, and the ODE model allowed us to identify valid outliers, allowed very accurate estimates of dynamic patterns of bird occurrence that provide detailed estimates of species migration timing; when they arrive, when they become more regular, and when they depart. Our approach established occurrence probabilities based on when submissions occur and allowed the quick identification and classification of outliers.

An automated approach to checking the validity of identifications made by citizen scientists that is based on both the patterns of submissions within a predefined spatial and temporal extent, as well as the contributor's skill level has the potential to play a critical role in improving data quality in broad-scale citizen-science projects. The results we present, however, are from a very small region, and must be tested more broadly, and within areas where both more data and fewer data are submitted. In addition, the current approach does not consider data entry errors, and any automated filtering process must be extremely careful when blanket-accepting all records from a particular class of users. Finally, a birder can be an expert observer in their home region, but less so outside of that region, which would require that an individual's expertise would need to be established regionally.

ACKNOWLEDGMENT

We thank D. Fink, C. Wood, M. Illiff 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).

[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,* vol. 59, pp. 977-984, 2009.

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

[3] D. J. Trumbull, R. Bonney, D. Bascom, and A. Cabral, "Thinking scientifically during participation in a citizen-science project," *Science Education,* vol. 84, pp. 265-275, 2000.

[4] B. L. Sullivan, C. L. Wood, M. J. Iliff, R. E. Bonney, D. Fink, and S. Kelling, "eBird: A citizen-based bird observation network in the biological sciences," *Biological Conservation,* vol. 142, pp. 2282-2292, 2009.

[5] R. Caruana, M. Elhawary, A. Munson, M. Riedewald, D. Sorokina, D. Fink, W. Hochachka, and S. Kelling., "Mining Citizen Science Data to Predict Prevalence of Wild Bird Species," *Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining,* pp. 909-915, 2006.

[6] D. Fink and W. M. Hochachka, "Gaussian semiparametric analysis using hierarchical predictive models," in *Environmental and Ecological Statistics special monograph on "Modeling Demographic Processes in Marked Populations* D. Thomson, E. Cooch, and M. Conroy, Eds., ed, 2009.

[7] D. Fink and W. M. Hochachka, "Mining information from citizen science data: The use of new analytical techniques for exploratory analysis of broad‐scale observational data.," in *CITIZEN SCIENCE: PUBLIC COLLABORATION IN ENVIRONMENTAL RESEARCH*, J. a. B. Dickinson, R., Ed., ed: Cornell University Press, 2011 (In Press).

[8] W. M. Hochachka, R. Caruana, D. Fink, A. Munson, M. Riedewald, D. Sorokina, and S. Kelling, "Data-mining discovery of pattern and process in ecological systems," *Journal of Wildlife Management,* vol. 71, pp. 2427-2437, 2007.

[9] W. M. Hochachka, R. Caruana, D. Fink, M. A. Munson, M. Riedewald, D. Sorokina, and S. Kelling, "Data-Mining Discovery of Pattern and Process in Ecological Systems," *Journal of Wildlife Management,* vol. 71, pp. 2427-2437, 09/02/ 2007.

[10] W. M. Hochachka, D. Fink, and B. Zuckerberg, "Use of citizen science monitoring for pattern discovery and biological inference," in *Design and Analysis of Long-term Ecological Monitoring Studies*, R. A. Gitzen, J. J. Millspaugh, A. B. Cooper, and D. S. Licht, Eds., ed: Cambridge University Press, 2011 (In Press).

[11] W. M. Hochachka, D. Fink, R. A. Hutchinson, D. Sheldon, W. K. Wong, and S. Kelling, "Project and Analysis Design for Broad-Scale Citizen Science," *Trends in Ecology & Evolution,* 2011 (in Press).

[12] S. Kelling, D. Fink, W. M. Hochachka, K. Rosenberg, R. Cook, T. Damoulas, C. Silva, and W. K. Michener, "Estimating Species Distributions, Across Space through Time and with Features of the Environment," in *Data Intensive Science*, Malcolm Atkinson and P. Brezany, Eds., ed, 2011 (in Press).

[13] M. A. Munson, R. Caruana, D. F. Fink, W. M. Hochachka, M. I. Iliff, K. V. Rosenberg, D. R. Sheldon, B. L. Sullivan, C. L. Wood, and S. Kelling, "A Method For Measuring the Relative Information Content of Data Freom Different Monitoring Protocols," *Methods In Ecology and Evolution,* vol. 1, pp. 263-273, 2010.

[14] D. Sorokina, R. Caruana, M. Riedewald, W. M. Hochachka, and S. Kelling, "Detecting and Interpreting Variable Interactions in Observational Ornithology Data," *Proc. IEEE Int. Workshop on Domain Driven Data Mining (DDDM),* 2009.

[15] N. A. B. C. I. U.S., "The State of the Birds 2011 Report on Public Lands and Waters," Washington, D.C.2011.

[16] S. Kelling, "Using Bioinformatics In Citizen Science," in *Citizen Science: Public Collaboration in Enviromental Research*, J. a. B. Dickinson, R. , Ed., ed: Cornell University Press, 2011

[17] K. A. Ericsson and N. Charness, "Expert performance: Its structure and acquisition," *American Psychologist,* vol. 49, pp. 525-747, 1994.

[18] S. N. Wood, *Generalized additive models: an introduction with R*: Chapman & Hall/CRC, 2006.

[19] S. Phillips, D. Miroslav, and R. Schapire, "A maximum entropy approach to species distribution modeling," presented at the Proceedings of the twenty-first international conference on Machine learning, Banff, Alberta, Canada, 2004.

[20] J. Elith and J. Leathwick, "Species distribution models: Ecological explanation and prediction across space and time. ," *Annual Review of Ecology, Evolution and Systematics 40, 677-69,* vol. 40, pp. 677-690, 2009.

[21] D. I. MacKenzie, J. D. Nichols, J. A. Royle, K. H. Pollock, L. L. Bailey, and J. E. Hines, *Occupancy Estimation and Modeling*. Amsterdam: Elsevier, 2006.

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

[23] D. Fink, W. M. Hochachka, D. Winkler, B. Shaby, G. Hooker, B. Zuckerberg, M. A. Munson, D. Sheldon, M. Riedewald, and S. Kelling, "Spatiotemporal Exploratory models for Large-scale Survey Data," *Ecological Applications,* vol. 20, pp. 2131-2147, 2010.

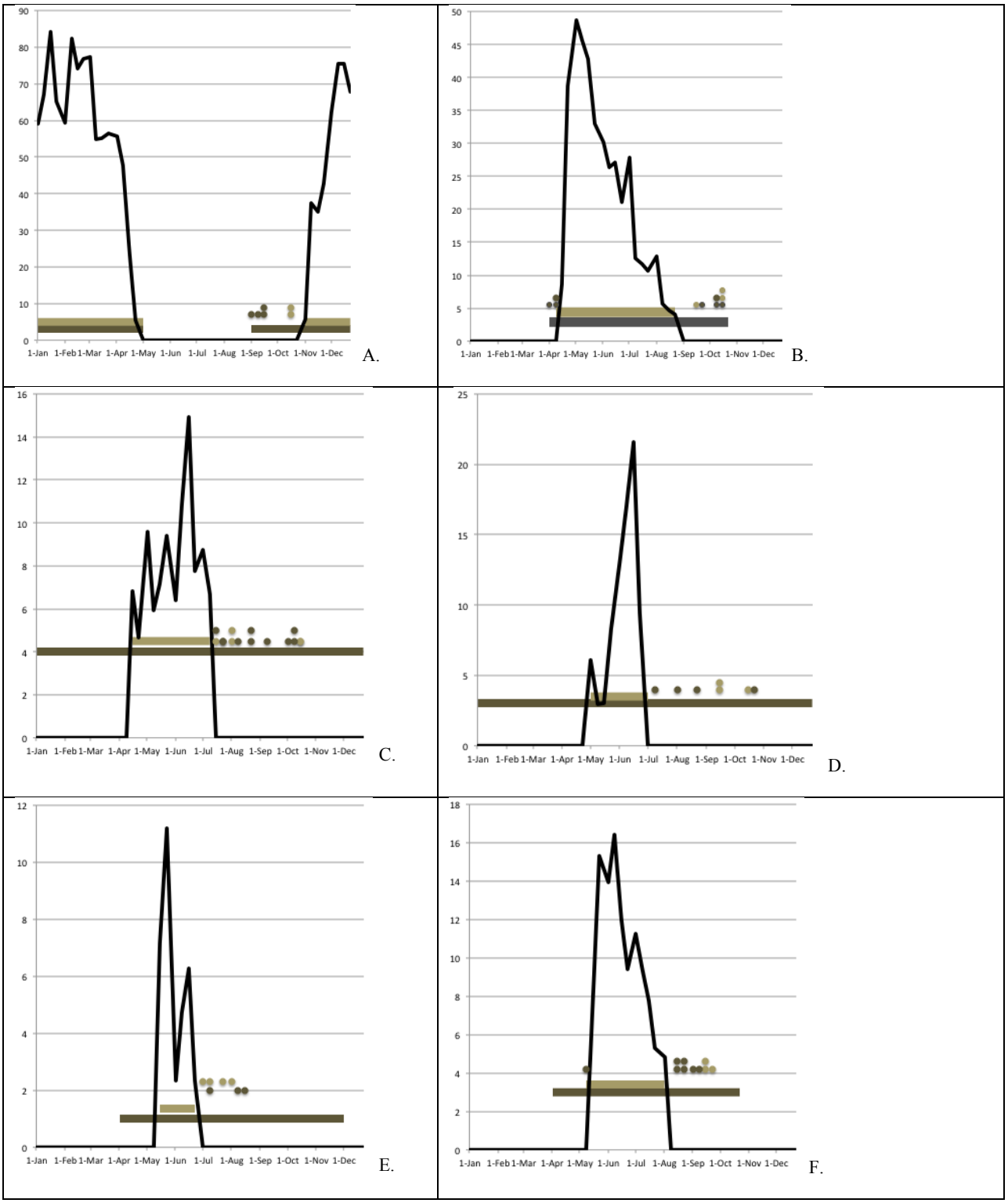

Figure 1. The graphs represent the frequency of checklists reporting a selection of birds in Jefferson Co. New York. They include: American Tree Sparrow (*Spizella arborea*) (A) and Chipping Sparrow (*Spizella passerina*) (B), Swamp Sparrow (*Melospize georgiana*) (C), Savannah Sparrow (*Passerculus sandwichensis*) (D), Veery (*Catharus fuscescens*) (E), and Red-eyed Vireo (*Vireo olivaceus*) (F), The Y-axis is the proportion of checklists that reported the species, and the X-axis is the date. The solid black line is the frequency of checklists that reported that species during a 1-week period. The dark bar is the window where the *expert-defined filter* permits the species to occur, and the light bar shows the *emergent data filter* limits. The solid dark circles are observations made by observers the ODE model identified as birders with low eBird expertise, and solid light circles with high eBird expertise. What is obvious is that species patterns of occurrence vary in the county.