Clustering Species Accumulation Curves to Identify Groups of Citizen Scientists with Similar Skill Levels

Jun Yu
Department of EECS
Oregon State University
Corvallis, OR 97331
yuju@eecs.orst.edu

Weng-Keen Wong
Department of EECS
Oregon State University
Corvallis, OR 97331
wong@eecs.orst.edu

Steve Kelling
Cornell Lab of Ornithology
Cornell University
Ithaca, NY 14850
stk2@cornell.edu

Abstract

Although citizen science projects such as eBird can compile large volumes of data over a broad spatial and temporal extent, the quality of this data can be a concern due to differences in the skills of volunteers at identifying bird species. Species accumulation curves, which plot the number of unique species observed over time, are an effective way to quantify the skill level of an eBird participant. Intuitively, the more skilled observers can identify more species per unit time than inexperienced birders, resulting in a steeper curve. We propose a mixture model for clustering species accumulation curves. With these clusters, we can identify distinct skill levels of eBird participants, which can be used to classify birders into skill categories and to develop automated data filters to improve data quality.

1 Introduction

Citizen science is a paradigm in which volunteers from the general public collect scientifically-relevant data. This paradigm is especially useful when the scope of the data collection is too broad to be performed only by trained scientists. Our work is in the context of the eBird project (www.eBird.org) [8, 5], which relies on a global network of citizen scientists to record checklists of bird observations, identified by species, through a protocol-driven process. These checklists are submitted via the web and compiled by the Cornell Lab of Ornithology, forming one of the largest biodiversity datasets in existence, with over 140 million observations reported by 150,000 birders worldwide. This data plays an important role in ecological research [4] and conservation [7].

With such a large volume of data submitted by volunteers, data quality is an ongoing concern. The current eBird system employs a regional filter based on expected occurrences of each species at specific times of the year. This filter flags anomalous observations and any flagged records are reviewed by a large network of volunteer reviewers. Observations are discarded if they do not pass the review stage; otherwise the data is accepted to the database.

A major factor influencing data quality is the variation in observer skill at identifying bird species. An observer's skill level can be characterized by a species accumulation curve (SAC) [3], which plots the number of unique species observed over time. SACs are typically used in the ecological literature to quantify species richness [1] but they are also effective at modeling an observer's skill level. Intuitively, skilled birders rely on both sound and sight to identify bird species and thus are able to identify more species per unit time than inexperienced birders, resulting in a steeper SAC.

Our goal is to identify distinct groups of eBird participants that are at similar skill levels. To accomplish this, we develop a mixture model to cluster the SACs of eBird participants. These clusters can be used to classify birders into different skill levels, which can then be used to develop automated data quality filters [9] and to track how the skills of individual birders evolve over time. We apply our clustering algorithm to eBird data in 2012 and show that the skill levels corresponding to the resulting clusters are meaningful.

2 The mixture of Species Accumulation Curves model

In the mixture of SACs model, we assume that there is a fixed number K of distinct groups of observers and that observers in the same group are at similar skill levels. As eBird is our application domain, we use *observer* and *birder* interchangeably. Figure 1 shows a plate diagram of the mixture of SACs model. The plate on the left represents K groups where group k is parameterized with $\boldsymbol{\beta}_k$. The outer plate on the right represents M birders. The variable $Z_i \in \{1, \cdots, K\}$ denotes the group membership of birder i. The inner plate represents N_i checklists submitted by birder i. The variable X_{ij} represents the amount of effort (e.g. duration) and Y_{ij} specifies the number of unique species reported on checklist j of birder i. Finally, let $\boldsymbol{X}_{ij}$ denote the variable X_{ij} with the intercept term.

The observation variable Y_{ij} depends on the effort X_{ij} and the skill level of birder i, indicated by the group membership Z_i. To model their relationship in a SAC, we use a linear regression model with a square root transformation on X_{ij} (i.e. $Y_{ij} = \beta_0 + \beta_1\sqrt{X_{ij}}$) because it produces the best fit to the data, where the fit is measured in terms of mean squared error on a holdout set.

The structure of the mixture model corresponds to the following generative process. For each birder i, we first generate its group membership Z_i by drawing from a multinomial distribution with parameter $\boldsymbol{\pi}$. Next, birder i produces N_i checklists. On each checklist j, the expected number of species detected is $\boldsymbol{\beta}_{Z_i} \cdot \boldsymbol{X}_{ij}$. Finally, the number of species actually reported (Y_{it}) is generated by drawing from a Gaussian distribution with mean $\boldsymbol{\beta}_{Z_i} \cdot \boldsymbol{X}_{ij}$ and variance σ^2. Here we assume SACs in different groups share the same variance σ^2. The log-likelihood for this mixture model is given in Equation 1.

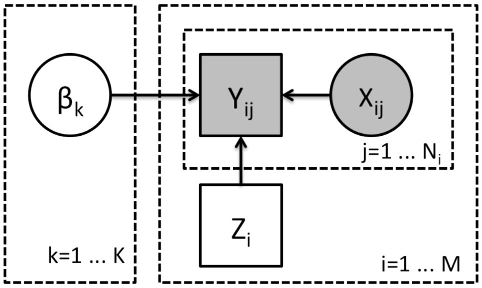

Figure 1: The mixture of SACs model.

$$\log P(\boldsymbol{Y}|\boldsymbol{X};\boldsymbol{\pi},\boldsymbol{\beta},\sigma^2) = \sum_{i=1}^{M} \log \left(\sum_{k=1}^{K} P(Z_i = k;\boldsymbol{\pi}) \prod_{j=1}^{N_i} P(Y_{ij}|\boldsymbol{X}_{ij}, Z_i = k;\boldsymbol{\beta},\sigma^2) \right) \tag{1}$$

2.1 Parameter estimation

During learning, we estimate the model parameters $\{\boldsymbol{\pi},\boldsymbol{\beta},\sigma^2\}$ and the latent group membership $\boldsymbol{Z}$ for each birder using Expectation Maximization [2]. In the E-step, EM computes the expected group membership for every birder i. In the M-step, we re-estimate the model parameters $\{\boldsymbol{\pi},\boldsymbol{\beta},\sigma^2\}$ that maximize the expected complete log-likelihood in Equation 2.

$$\begin{aligned}\mathcal{Q} &= E_{\boldsymbol{Z}|\boldsymbol{Y},\boldsymbol{X}}[\log(P(\boldsymbol{Y},\boldsymbol{Z}|\boldsymbol{X};\boldsymbol{\pi},\boldsymbol{\beta},\sigma^2))] \\ &= \sum_{i=1}^{M}\sum_{k=1}^{K} E_{\boldsymbol{Z}|\boldsymbol{Y},\boldsymbol{X}}[\mathbb{I}(Z_i = k)] \log \left(P(Z_i = k;\boldsymbol{\pi}) \prod_{j=1}^{N_i} P(Y_{ij}|\boldsymbol{X}_{ij}, Z_i = k;\boldsymbol{\beta},\sigma^2) \right)\end{aligned} \tag{2}$$

In the E-step, the expected group membership of birder i belonging to group k is computed as the posterior probability r_{ik} in Equation 3.

$$r_{ik} = P(Z_i = k|\boldsymbol{X}_{i\cdot}, \boldsymbol{Y}_{i\cdot};\boldsymbol{\pi},\boldsymbol{\beta},\sigma^2) = \frac{P(Z_i = k;\boldsymbol{\pi}) \prod_{j=1}^{N_i} P(Y_{ij}|\boldsymbol{X}_{ij}, Z_i = k;\boldsymbol{\beta},\sigma^2)}{\sum_{k'=1}^{K} P(Z_i = k';\boldsymbol{\pi}) \prod_{j=1}^{N_i} P(Y_{ij}|\boldsymbol{X}_{ij}, Z_i = k';\boldsymbol{\beta},\sigma^2)} \tag{3}$$

In the M-step, we re-estimate $\{\boldsymbol{\pi}, \boldsymbol{\beta}, \sigma^2\}$ using the expected membership computed in the E-step. To estimate π_k, we introduce a Lagrange multiplier to ensure that the constraint $\sum_{k=1}^{K} \pi_k = 1$ is satisfied. i.e. $\sum_{i=1}^{M} r_{ik} - \lambda\pi_k = 0$. Summing over all $k \in \{1, \cdots, K\}$, we get the updating equation for π_k in Equation 4. The gradient of $\boldsymbol{\beta}_k$ in Equation 5 has the same form as a linear regression model, except that each instance is associated with a weight of r_{ik}. Thus we can use the method of least squares to update $\boldsymbol{\beta}_k$ efficiently. Finally, the parameter σ^2 can be estimated using the closed-form solution in Equation 6. Given the expected memberships $r_{i\cdot}$ of birder i, we then assign birder i to the group of the largest expected membership.

$$\pi_k = \frac{1}{M}\sum_{i=1}^{M} r_{ik} \tag{4}$$

$$\frac{\partial \mathcal{Q}}{\partial \boldsymbol{\beta}_k} = \frac{1}{\sigma^2}\sum_{i=1}^{M} r_{ik}\sum_{j=1}^{N_i}(Y_{ij} - \boldsymbol{\beta}_k \boldsymbol{X}_{ij})\boldsymbol{X}_{ij} \tag{5}$$

$$\sigma^2 = \frac{\sum_{i=1}^{M}\sum_{k=1}^{K} r_{ik}\sum_{j=1}^{N_i}(Y_{ij} - \boldsymbol{\beta}_k \boldsymbol{X}_{ij})^2}{\sum_{i=1}^{M} N_i} \tag{6}$$

3 Results and discussion

We evaluate the mixture of SACs model in four different states (NY, FL,TX,CA) using the eBird Reference Data [6]. First, we remove the birders who submitted fewer than 20 checklists in 2012 because their data is too sparse to fit our model. In addition, we limit our analysis to only include checklists with duration less than 2 hours. To find the value of K, we randomly split birders into training and validation sets and learn the mixture model using data submitted by birders in the training set with different values of $K \in \{1, \cdots, 6\}$. Then we calculate the average log-likelihood[1] on the holdout data (data submitted by birders in the validation set) and choose the value of K when increasing K does not improve the average log-likelihood. In Table 1, we show the average log-likelihood on the holdout data in four states. It clearly shows that there are 3 distinct groups in all four states.

State	K=1	K=2	K=3	K=4	K=5	K=6
New York	-3.456	-3.407	**-3.396**	-3.400	-3.406	-3.417
Florida	-3.398	–3.389	**-3.387**	-3.393	-3.405	-3.419
Texas	-3.543	-3.496	**-3.491**	-3.495	-3.501	-3.511
California	-3.507	-3.483	**-3.481**	-3.489	-3.493	-3.501

Table 1: The average log-likelihood of the holdout data in four states. The numbers in bold indicate the number of distinct groups found in that state.

Once we determine the best value of K for each state, we retrain the model using the entire eBird data from 2012 and show the SACs of different groups learned from the mixture model. In Figure 2, we sort the SACs by their slope coefficient β_1 in decreasing order so that the top group corresponds to the most skilled observers. For example, in New York there are 7% observers falling into the top group as they are able to observe more species per unit of time.

A good partition of birders leads to distinct differences in the skill levels of different groups. Since we do not have ground truth on the skill level of birders, we characterize their skill levels in terms of their ability to identify hard-to-detect bird species. We use 6 hard-to-detect species in New York and Florida suggested by experts at the Cornell Lab of Ornithology and calculate the average detection rate of a species within each group in 2012. A birder's detection rate of a species s is the percent of their checklists reporting species s. In Table 2, the top group has the highest detection rate across all 6 species, showing that a steeper SAC does in fact correspond to a better skill level. As we go from group 1 to group 3, the detection rate of reporting these species keeps decreasing and shows large differences even between two adjacent groups. These differences show that birders in different groups vary greatly in their skill levels and the mixture model is able to cluster birders of similar skills into the same group.

In addition, we sent a list of birder IDs in the top group for New York to the eBird project leaders and asked them to verify if these birders are top-notch birders in the community. Out of 30 birders in the top group, 25 are experts from the Cornell Lab of Ornithology or known regional experts in

[1] Calculated by computing the data likelihood of a birder and dividing by the number of checklists submitted by that birder.

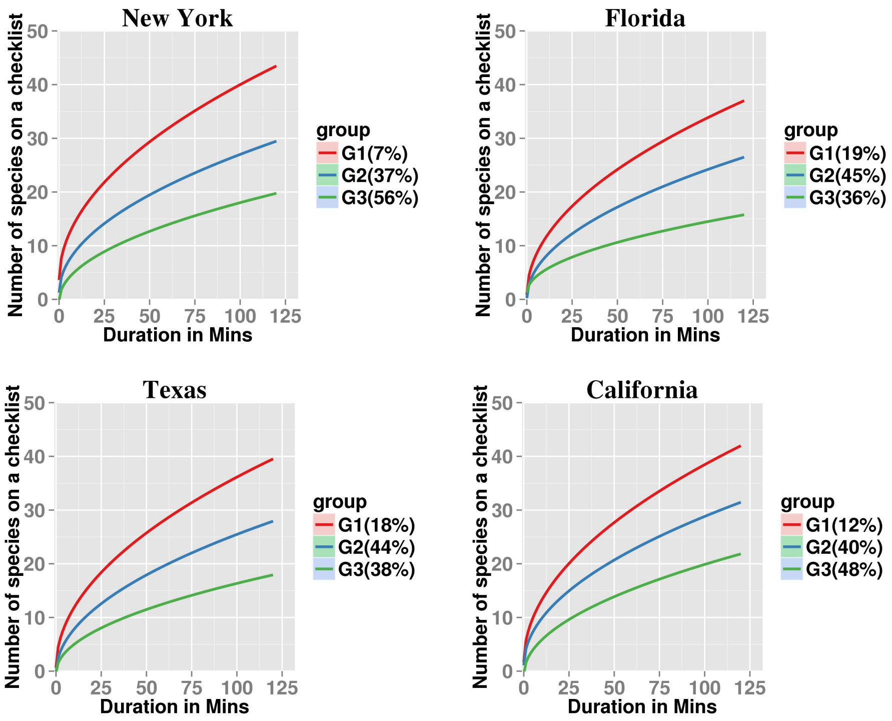

Figure 2: Species accumulation curves learned from the mixture of SACs model in four states. The number in the legend indicates the proportion of birders in each group.

New York	Bank Swallow	Least Flycatcher	Marsh Wren	Savannah Sparrow	Swamp Sparrow	Wood Thrush
Group 1	6.55 ± 1.17	5.56 ± 1.02	5.68 ± 1.22	9.45 ± 1.31	15.1 ± 1.38	9.23 ± 1.41
Group 2	2.48 ± 0.30	2.64 ± 0.26	2.91 ± 0.37	6.06 ± 0.61	10.2 ± 0.63	5.97 ± 0.52
Group 3	0.59 ± 0.08	1.09 ± 0.12	1.03 ± 0.13	2.21 ± 0.20	4.58 ± 0.37	3.92 ± 0.32

Florida	Chimney Swift	Hermit Thrush	House Wren	Marsh Wren	Savannah Sparrow	Yellow-rumped Warbler
Group 1	9.69 ± 1.36	2.31 ± 0.42	15.0 ± 1.61	4.97 ± 0.90	11.7 ± 1.40	24.9 ± 1.74
Group 2	4.50 ± 0.51	1.11 ± 0.17	5.45 ± 0.54	1.86 ± 0.32	5.58 ± 0.40	16.2 ± 1.01
Group 3	2.99 ± 0.55	0.60 ± 0.13	1.80 ± 0.34	0.81 ± 0.15	2.78 ± 0.36	10.5 ± 0.96

Table 2: The average detection rate (with standard error) on 6 hard-to-detect species in NY and FL.

New York while the other 5 observers are known to be reputable birders submitting high quality checklists to eBird.

4 Conclusion

We proposed a mixture model for Species Accumulation Curves that was successful at identifying distinct groups of citizen scientists with similar skill levels in the eBird project. In addition, the clusters discovered from NY data do in fact correspond to groups that vary in their ability to observe hard-to-detect bird species. In future work, we plan to account for other factors in the model such as location and extend this model to capture the evolution of an observer's skill level over time.

Acknowledgements

We would like to thank Marshall Iliff, Brian Sullivan and Chris Wood for their bird expertise. This material is based upon work supported by the National Science Foundation under Grants No. 0941748 and 1209714.

References

[1] A. Chao and L. Jost. Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size. *Ecology*, 93:2533–2547, 2012.

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

[3] N. Gotelli and R. Colwell. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters*, pages 379–391, 2001.

[4] 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.

[5] S. Kelling, C. Lagoze, W.-K. Wong, J. Yu, T. Damoulas, J. Gerbracht, D. Fink, and C. P. Gomes. ebird: A human/computer learning network to improve conservation and research. *AI Magazine*, 34(1):10–20, 2013.

[6] 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 1.0. Cornell Lab of Ornithology and National Audubon Society, Ithaca, NY, June 2009.

[7] North American Bird Conservation Initiative, U.S. Committee. The state of the birds 2013 report on private lands, 2013.

[8] 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.

[9] J. Yu, S. Kelling, J. Gerbracht, and W.-K. Wong. Automated data verification in a large-scale citizen science project: a case study. In *Proceedings of the 8th IEEE International Conference on E-Science*, pages 1–8, 2012.