AN ABSTRACT OF THE DISSERTATION OF

Jun Yu for the degree of Doctor of Philosophy in Computer Science presented on December 3, 2013.

Title: Machine Learning for Improving the Quality of Citizen Science Data

Abstract approved: ______________________________

Weng-Keen Wong

Citizen Science is a paradigm in which volunteers from the general public participate in scientific studies, often by performing data collection. This paradigm is especially useful if the scope of the study is too broad to be performed by a limited number of trained scientists. Although citizen scientists can contribute large quantities of data, data quality is often a concern due to variability in the skills of volunteers. In my thesis, I investigate applying machine learning techniques to improve the quality of data submitted to citizen science projects. The context of my work is eBird, which is one of the largest citizen science projects in existence. In the eBird project, citizen scientists act as a large global network of human sensors, recording observations of bird species and submitting these observations to a centralized database where they are used for ecological research such as species distribution modeling and reserve design. Machine learning can be used to improve data quality by modeling an observer's skill level, developing an automated data verification model and discovering groups of misidentified species.

Machine Learning for Improving the Quality of Citizen Science Data

by

Jun Yu

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented December 3, 2013
Commencement June 2014

Doctor of Philosophy dissertation of Jun Yu presented on December 3, 2013.

APPROVED:

Major Professor, representing Computer Science

Director of the School of Electrical Engineering and Computer Science

Dean of the Graduate School

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request.

Jun Yu, Author

ACKNOWLEDGEMENTS

I would like to thank the following individuals for their helps during my Ph.D. study at Oregon State University.

- I would like to thank my friends: Wei Xu, Yan Hu, Peng Wang, Shu Chen, Zhiqing Zhang, Yun Fan, Xiaoyu Chai, Meidan Bu and Chao Wang. Coming to OSU for my Ph.D. study in 2006 is my first time away from my parents, it is you who made me feel Corvallis was home even I was away from my family.

- I would like to thank my committee members: Alan Fern, Xiaoli Fern, Prasad Tadepalli and Jimmy Yang. They are very patient and helpful during my Ph.D. exams and meetings.

- I would like to thank my collaborators at Oregon State University: Rebecca Hutchinson, Janardhan Rao Doppa, Shubhomoy Das, Tom Dietterich, Susan Shirley, Sarah Frey, Matt Betts and Julia Jones. It was a great pleasure to work with them on different research problems and I have learned a lot from them.

- I would like to thank my collaborators at Cornell Lab of Ornithology: Marshall Iliff, Brian Sullivan, Chris Wood, Jeff Gerbracht and Steve Kelling. Thanks to their supports in my research on the eBird project. They are great people to work with as they always provide useful suggestions and feedback.

- I would like to thank my Ph.D. advisor Weng-Keen Wong. He is a great great advisor and has always been very patient and supportive during my entire Ph.D. study. Most importantly, he is a role model for me to learn and to follow in every aspect of my research.

- Finally, I would like to thank my family: my wife, Juan He, my parents, and my parents-in-law. Thanks for being there for me during my ups and downs. I could not have archived what I have now without your supports and love.

TABLE OF CONTENTS

Page

1 Introduction 1

2 eBird: A Human/Computer Learning Network for Biodiversity Conservation and Research 4

3 Modeling Experts and Novices in Citizen Science data for Species Distribution Modeling 7

3.1 Introduction 7

3.2 Methodology 8

3.2.1 The Occupancy-Detection model 9

3.2.2 The Occupancy-Detection-Expertise model 11

3.2.3 Parameter estimation 14

3.2.4 Inference 16

3.3 Evaluation and Discussion 18

3.3.1 Data description 18

3.3.2 Task 1: Prediction of observations on a checklist 20

3.3.3 Task 2: Prediction of birder's expertise 23

3.3.4 Task 3: Contrast mining 25

3.4 Conclusion 26

4 Automated Data Verification in a Large-scale Citizen Science Project: a Case Study 27

4.1 Introduction 27

4.2 Methodology 29

4.2.1 Expert-defined filter 29

4.2.2 Automated data filter 30

4.3 Evaluation 32

4.3.1 Data description 32

4.3.2 Emergent data filter 33

4.3.3 Estimating an observer's expertise level 33

4.4 Results and Discussion 34

4.4.1 Emergent data filter 34

4.4.2 The automated data filter 36

4.5 Conclusion 39

TABLE OF CONTENTS (Continued)

Page

5 Clustering Species Accumulation Curves to Identify Groups of Citizen Scientists with Similar Skill Levels 41

5.1 Introduction 41

5.2 Methodology 43

5.2.1 The mixture of Species Accumulation Curves model. 43

5.2.2 Parameter estimation 45

5.2.3 Determining the number of groups 47

5.2.4 Inference 47

5.3 Evaluation and Discussion 47

5.3.1 Species Accumulation Curves 49

5.3.2 Grouping eBird participants 50

5.3.3 Detection of hard-to-detect bird species 53

5.3.4 eBird Hotspots 55

5.4 Conclusion 58

6 Modeling Misidentification of Bird Species by Citizen Scientists 59

6.1 Introduction 59

6.2 Methodology 61

6.2.1 The Multi-Species Occupancy-Detection model 61

6.2.2 Parameterization 63

6.2.3 Structure learning and parameter estimation 64

6.2.4 Inference 67

6.3 Evaluation and Discussion 68

6.3.1 Synthetic dataset 69

6.3.2 eBird dataset 72

6.4 Conclusion 74

7 Improving Predictions of Rare Species through Multi-species Distribution Modeling with an Ensemble of Classifier Chains 78

7.1 Introduction 78

7.2 Methods 81

7.2.1 Study sites 81

7.2.2 Data 82

7.2.3 Model development 83

TABLE OF CONTENTS (Continued)

Page

7.2.4 Temporal and spatial autocorrelation . . . 84
7.2.5 Experiments and evaluation . . . 85
7.2.6 Hypothesis Testing . . . 87

7.3 Results . . . 88

7.4 Discussion . . . 96
7.4.1 Performance of multi-species vs. single-species models . . . 96
7.4.2 Mechanisms for improved performance of multi-species distribution models . . . 98
7.4.3 Future research directions for modeling multi-species distributions 100

7.5 Conclusion . . . 101

8 Conclusions and Future Work 103

8.1 Contributions . . . 103

8.2 Future Work . . . 104

8.3 Acknowledgements . . . 105

Bibliography 106

LIST OF FIGURES

Figure Page

2.1 The number of eBird observations per month since 2003. 4

2.2 An overview of the eBird system. 5

2.3 The distribution of Western Meadowlark across the western United States in June 2009. We thank Daniel Fink for this species distribution map as produced by the STEM model [29]. 6

3.1 Graphical model representation of the Occupancy-Detection model for a single bird species. 10

3.2 Graphical model representation of Occupancy-Detection-Expertise model for a single bird species. Note that the link from E_j to Y_{it} only exists if birder j submits the checklist corresponding to Y_{it}. 13

4.1 The weekly occurrence frequency for Chipping Sparrow in Tompkins Co., NY. 31

4.2 An illustration of the time periods covered by the expert-defined filter (light grey bar) and the emergent filter (dark grey bar). Observations falling with the bars are automatically accepted. Observations falling outside of the bars are flagged for review. 35

4.3 Flagged observations for (A) American Tree Sparrow and (B) Chipping 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 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. 37

4.4 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. . 39

5.1 The mixture of Species Accumulation Curves model. 44

Figure Page

5.2 The species accumulation curves of the active and occasional eBird users in four states. The shaded area of a curve shows the 95% confidence interval. 48

5.3 The species accumulation curves showing the evolution of eBird participants' skills over time. The shaded area of a curve shows the 95% confidence interval. 49

5.4 The average log-likelihood on the holdout data for different values of K in four states. The highest number indicates the best number of distinct groups found in that state. 51

5.5 The species accumulation curves learned from the mixture of SACs model in four states. The number in the parenthesis indicates the proportion of birders in each group. The shaded area of a curve shows the 95% confidence interval. 53

5.6 The species accumulation curves and the scatter plots of eBird participants from three groups in New York. The shaded area of a curve shows the 95% confidence interval. Each point represents a checklist submitted by the birder. The color of a point in the scatterplot specifies the number of checklists overlapped in the location. A darker color indicates more checklists overlapped at that point. 54

5.7 (a): The average detection rate of three groups on 8 hard-to-detect species in New York. (b): the average detection rate of three groups on 8 hard-to-detect species in Florida. The error bars represent the standard error of detection rate within a group. 56

5.8 (a): the average detection rate of three groups on 8 hard-to-detect species in Texas. (b): the average detection rate of three groups on 8 hard-to-detect species in California. The error bars represent the standard error of detection rate within a group. 57

5.9 The species accumulation curves in eBird hotspots *Stewart Park* and *Hammond Hill*. The number in the parenthesis indicates the proportion of birders in that group. The shaded area of a curve shows the 95% confidence interval. 58

LIST OF FIGURES (Continued)

Figure Page

6.1 Graphical model representation of the Multi-Species Occupancy-Detection model. 62

6.2 The flowchart of learning the MSOD model. 66

6.3 The arrows specify the species confusions recovered by the MSOD model. An arrow from species A to species B indicates that the presence of species A may result the detection of species B. We thank Chris Wood from the Cornell Lab of Ornithology for the images of each bird species. 75

7.1 Schematic diagram of experiment to test how the inclusion of information on the presence of con-specifics (A, B) affects the performance of multi-species models (GLM+ECC, BRT+ECC) compared to single-species models (GLM, BRT). Single-species models were generalized linear models (GLMs) and boosted regression trees (BRTs); multi-species versions of these models were constructed using the ensemble of classifier chains (ECC). 80

7.2 The difference in predictive success between multi-species and single-species models using BRT and GLM as base classifiers. A positive Δ species-based AUC indicates that a multi-species model has greater predictive success than single-species models. Standard errors for Δ species-based AUC are computed over S species. 89

7.3 Difference in predictive success of multi-species vs. single-species models for rare vs. common species, based on differences in species-based AUC for boosted regression trees (a) and general linear models (b) of five species datasets. The standard errors for Δ species-based AUC are computed over the species in each group. 90

7.4 Difference in predictive success of multi-species vs. single-species models for four subsets of species (RMC, RC, RM, and R, where R = rare, M = moderately common, and C = common), based on differences in species-based AUC for boosted regression tree models (ECC-BRT vs. BRT) for five species datasets. For example, ECC-R refers to the difference in predictive success of multi-species models compared to single-species models, when only rare species are included. Error bars show the standard error of Δ species-based AUCs across all the rare species in the dataset. 92

LIST OF FIGURES (Continued)

Figure Page

7.5 Drop-one analysis on AND Bird 2009 and 2010 data. (top) Change in species-based AUC for the rare species, olive-sided flycatcher in 2009, associated with dropping all species one by one. The changes for the 14 most influential species are shown. (bottom) Change in species-based AUC for the rare species, American robin in 2010, associated with dropping all species one by one. The changes for the 14 most influential species are shown. 93

7.6 Drop-one analysis on HBR Bird 1999 data. (top) Change in species-based AUC for the rare species, wood thrush and (second from top) rose grosbeak, in 1999, associated with dropping all species one by one. The changes for the 19 most influential species are shown. 94

7.7 Drop-one analysis on HBR Bird 2008 data. Change in species-based AUC for the rare species, black-and-white warbler (top) and scarlet tanager (bottom) in 2008, associated with dropping all species one by one. The changes for the 11 most influential species are shown. 95

7.9 Three different sets of conditions under which ecological processes result in improved species distribution models when boosted regression trees (BRTs) are combined with ensemble of classifier chains (ECC). (a) Species B cannot be effectively predicted from environmental covariates (e.g. because it is rare), but it appears to share A's environmental covariates, (b) Species A and B are predicted by environmental covariates, but also interact through some biotic mechanism (e.g. mutualism, competition) and this adds confidence to predictions from environmental covariates, and (c) The relationship between species A and B suggests that an unmeasured environmental covariate may influence both species. 99

7.8 Drop-one analysis on AND Moth data for 1986 to 2008. (top) The change in species-based AUC for the rare species, Zanclognatha jacchusalis, associated with dropping thirteen selected species one by one. (bottom) change in species-based AUC for the rare species, Lithophane baileyi, 1986 to 2008, associated with dropping nine selected species one by one. . 102

LIST OF TABLES

Table | Page

3.1 Notation in the Occupancy-Detection model. 9

3.2 Notation in the Occupancy-Detection-Expertise model. 12

3.3 Occupancy, Detection and Expertise features in eBird data set. 19

3.4 Bird species in each group. 20

3.5 Average AUC for predicting detections on test set checklists for bird species. 23

3.6 Average AUC for predicting birder expertise on a test set of birders for bird species. 25

3.7 Average Δ_{TD} for Group A and B. 26

4.1 Example of 9 bird species with different occurrence patterns in Tompkins Co., NY. The second and the third columns are the percentage of expert/novice observations flagged by the emergent filter. 38

4.2 The number of flagged records from Tompkins Co., NY and the estimated number of hours needed to review them. 39

5.1 Notation in the mixture of SACs model 43

5.2 The number of birders and the average number of checklists submitted per birder of each group in four states. 52

6.1 Notation in the Multi-Species Occupancy-Detection model. 61

6.2 The AUC and accuracy (and their standard errors) of occupancy and observation prediction averaged over 30 datasets in four synthetic experiments. The metrics are computed per species and averaged across species. Boldface results indicate the best performing model. $\star$ and † indicate the MSOD model is statistically better than the OD model and the ODLP model respectively. 71

LIST OF TABLES (Continued)

Table Page

6.3 The AUC (and its standard error) for the learned model structure of the MSOD model compared to the true model structure in four synthetic experiments. The AUC values are averaged over 30 datasets in each experiment. 72

6.4 Occupancy and detection features in eBird dataset used for evaluation. . . 73

6.5 The AUC and accuracy of observation prediction for three eBird case studies. Boldface results indicate the winner, ⋆ and † indicate the MSOD model is statistically better than the OD and ODLP model respectively. . 76

7.1 The number of sites in the training, validation and testing subsets. 84

7.2 The results of the Paired AUC Bootstrap Difference test on each dataset for all, rare and common species. 91

LIST OF ALGORITHMS

Algorithm Page

Chapter 1: Introduction

Citizen Science refers to scientific research in which volunteers from the community participate in scientific studies as field assistants [11]. Citizen science shares the same fundamental characteristics as crowdsourcing as they both focus on accumulating a large volume of data or completed tasks. Since citizen scientists can collect data or complete tasks relatively cheaply, it allows research to be performed at much larger spatial and temporal scales. For example, Galaxy Zoo[1] is an online astronomy project which invites people to assist in the morphological classification of large numbers of galaxies, EyeWire[2] is an online human-based computation game of coloring brain images which helps scientists to map the neural connections of the retina, and CoCoRaHS[3] encourages a network of volunteer weather watchers to take daily readings of precipitation and report them online, which can be used by the National Weather Service, meteorologists, hydrologists, emergency managers, etc.

In citizen science projects, participants may play the role of either a *sensor* or a *processor*. This is an important distinction when it comes to ensuring the quality of the resulting contributions. Processing tasks are usually repeatable, while sensing tasks rarely are. Processing tasks are typical of the Zooniverse family of projects [74], in which volunteers classify or extract information from images. The same tasks can be performed on the same artifact repeatedly by multiple volunteers and the validation of the process can be done through consensus of responses or directly from an expert [75]. When citizen scientists participate as sensors, ground truth is rarely available and events can neither be repeated nor independently observed. Observer variability and lack of ground truth is a long-standing point of concern and contention in such citizen science projects, with eBird[4] being an exemplar.

eBird [80, 47] is one of the largest citizen-science projects in existence, relying on a global network of bird-watchers to report their observations of birds, identified by species,

[1] www.galaxyzoo.org

[2] www.eyewire.org

[3] www.cocorahs.org

[4] www.ebird.org

to a centralized database. Since its inception in 2002 by the Cornell Lab of Ornithology, eBird has accumulated over 140 million observations reported by 150 thousand birders worldwide, generating one of the world's fastest growing biodiversity data sets. eBird data reveal patterns of bird occurrence across space and through time, allowing us to build large-scale species distribution models. Rapid global environmental change is expected to reduce significant changes in the distribution of many species and complicate efforts for species conservation. *Species Distribution Models* (SDMs), which estimate the pattern of species occurrence on a landscape based on its environmental features, help ecologists understand the relationship between species occurrence and its habitat, and thus play an import role in predicting biodiversity and designing wildlife reserves [51, 62].

In citizen science projects such as eBird, which rely on a large number of volunteer data collectors, 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 organisms by species. To address the data quality problems in citizen science projects, my research focuses on improving the quality of citizen science data using machine learning techniques. We collaborate with the Cornell Lab of Ornithology and work on the data quality issues in the eBird project. In particular, this thesis focuses on the following problems:

First of all, we study the problem of modeling birders' expertise in the SDMs and develop a probabilistic graphical model called the *Occupancy-Detection-Expertise* (ODE) model. This model incorporates the expertise of observers in building SDMs. We show that modeling the expertise of birders not only improves the accuracy of predicting species observations but it also allows us to predict the expertise level of new birders to eBird.

Next, we study the problem of building a data quality control model to filter out the invalid observations submitted to eBird. Our experience with the eBird project has shown that a massive effort by volunteer experts is needed to screen data, identify outliers and flag them in the database. In this work, we build a new two-step automated data verification process for the eBird project. Through our case study, we show its potential to identify more invalid observations and to reduce the total number of observations that need to be manually reviewed.

Then, we propose a mixture model to cluster eBird participants with similar skill levels. The Species Accumulation Curves (SACs) describe the number of unique species detected as a function of effort spent, which can be used to characterize an observer's skill level. We develop a mixture of SACs model and show that it is able to successfully identify distinct groups of citizen scientists with similar skill levels in eBird. With these clusters, we can classify birders to skill categories and develop automated data filters.

After that, we investigate a more proactive way to improve data quality by identifying groups of misidentified species, which can be used to teach inexperienced observers. We extend the *Occupancy-Detection* model in ecology to the multiple species case, which allows false positives for a species as arising from misidentifications of other species. In our study, we show that explicitly modeling the confusions and detection errors between species not only helps discover groups of confusing species, but also improves the estimates of the occupancy patterns of those species.

Finally, we study the problem of multi-species distribution modeling to improve the predictions for rare species. Simultaneous prediction of multiple species may improve predictive performance because interactions may drive distributions directly and many species co-occur in similar habitats. In particular, multiple species models may produce the greatest improvement of predictions for rare species. We apply a previously published multi-label machine learning algorithm that predicts multiple responses and our results show that multi-species models produce more accurate predictions than single-species models and these improvements are more consistent for rare species.

Chapter 2: eBird: A Human/Computer Learning Network for Biodiversity Conservation and Research

eBird [60], launched in 2002 by the Cornell Lab of Ornithology, is a citizen-science project that takes advantage of the human observational capacity to identify species to birds, and uses these observations to accurately represent patterns of bird occurrences across broad spatial and temporal extents. eBird participants report their observations in the form of checklists to the centralized eBird database. Up to 2013, more than 150,000 individuals have volunteered over 5 million hours to collect more than 140 million bird observations for the eBird project; it is arguably the largest biodiversity data collection project in existence. Figure 2.1 shows the total number of observations reported per month since 2003. These amassed observations provide researchers and scientists with data about bird distribution and abundance across a variety of spatiotemporal extents.

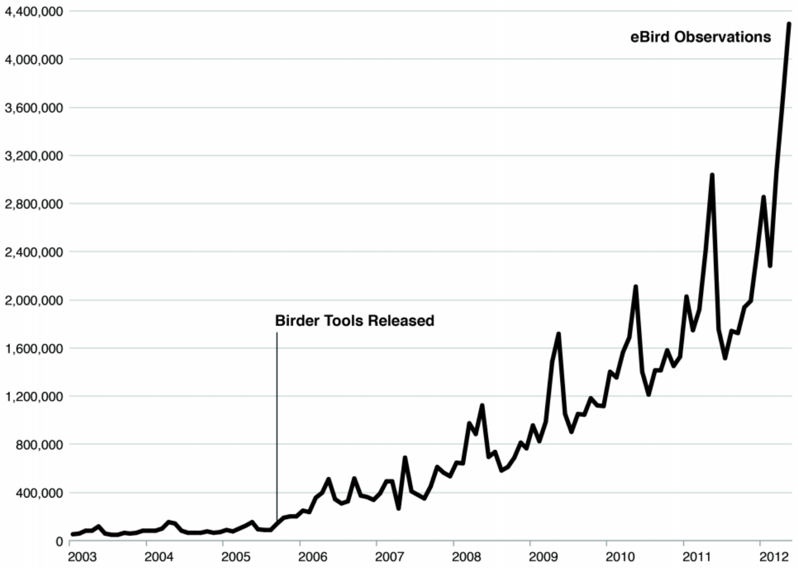


Figure 2.1: The number of eBird observations per month since 2003.

While eBird has been successful in engaging a global community of volunteers to contribute large quantities of observations of birds, it faces the ongoing challenges of data quality and spatial sampling bias. To address these challenges, eBird employs machine learning techniques to improve data quality by taking advantage of the synergies between human computation and mechanical computation. We call this a Human/Computer Learning Network (HCLN) [46], which has at its core an active learning feedback loop between humans and machines that dramatically improves the quality of both and thereby continually improves the effectiveness of the network as a whole. An overview of the eBird system is given in Figure 2.2. In this HCLN, a broad network of volunteers acts as intelligent and trainable sensors to gather observations and the artificial intelligence processes improves the quality of the observational data the volunteers provide by identifying and filtering out the invalid records. With the immediate feedback and the discovered misidentification patterns, the AI processes contribute to advancing observer expertise. Simultaneously, as observational data quality improves, it helps learn more accurate species distribution models and allows for better decisions for conservation planning.

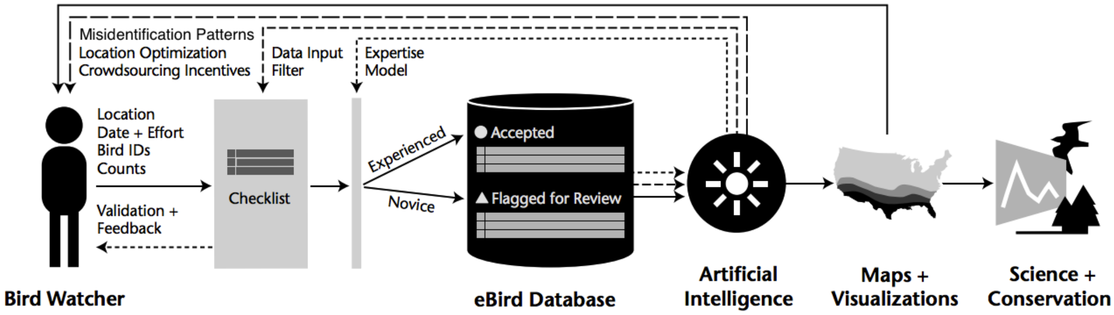


Figure 2.2: An overview of the eBird system.

In the eBird database, a record is a checklist, which corresponds to a single birding event. Every checklist contains four types of information: (a) observer, (b) location, (c) visit, and (d) species. Observer information (e.g. name, ID and contact information) allows every bird observation to be associated with a specific eBird user. Location information includes the site name, the GPS coordinates, and the environmental covariates of the site. Visit information specifies the factors associated with each visit, including the time of day, day of year, amount of effort spent (e.g. effort distance, effort time and area covered), and whether all the species observed were reported during that visit. Species information includes a list of birds identified during the birding event and the estimated number of each species detected.

eBird data reveal patterns of bird occurrence across space and through time, and provide a data-rich foundation for understanding the broad scale dynamic patterns of bird populations [40]. Recently, the United States Department of the Interior used eBird data as the basis for the 2011 State of the Birds Report [72], which estimated the occupancies of bird populations on public lands. Figure 2.3 shows an example of a species's distribution estimate of Western Meadowlark across the western United States in June 2009 based on eBird observations.

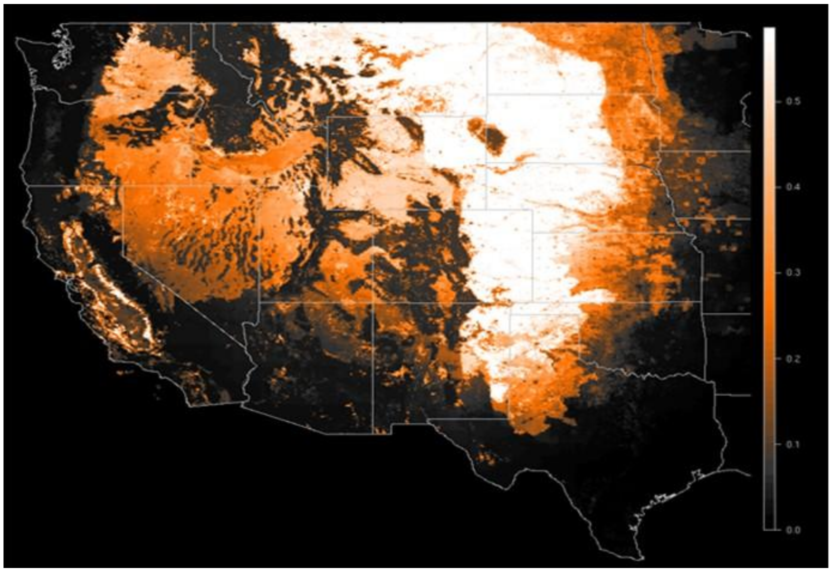


Figure 2.3: The distribution of Western Meadowlark across the western United States in June 2009. We thank Daniel Fink for this species distribution map as produced by the STEM model [29].

Chapter 3: Modeling Experts and Novices in Citizen Science data for Species Distribution Modeling

3.1 Introduction

The term *Citizen Science* refers to scientific research in which volunteers from the community participate in scientific studies as field assistants [11]. Since data collection by *citizen scientists* can be done cheaply, citizen scientists allow research to be performed at much larger spatial and temporal scales than trained scientists can cover. For example, species distribution modeling (SDM) [22] with citizen scientists allows data to be collected from many geographic locations, thus achieving broad spatial coverage. Most citizen scientists, however, have little or no scientific training. Consequently, the quality of the data collected by citizen scientists is often questioned. Recent studies have shown that citizen scientists were able to provide accurate data for easily detected organisms [13]. However, for difficult-to-detect organisms, Fitzpatrick et al. [30] found differences between observations made by volunteers and by experienced scientists leading to biased results.

Since eBird data is contributed by citizen scientists, can accurate species distribution models be built from this data? Checklists submitted to eBird undergo a data verification process which consists of automated data filters which screen out obvious mistakes on checklists. Then, the checklists go through a review process by a network of experienced birders. Nevertheless, biases still exist due to differences in the expertise level of birders who submit the checklists. In our work, we use a hierarchical Bayesian network that incorporates the expertise level of birders. With this model, we show that modeling the expertise level of birders can be beneficial for SDM.

In order to incorporate birder expertise into a species distribution model, we need to distinguish between two processes that affect observations: *occupancy* and *detection.* Occupancy determines if a geographic site is viable habitat for a species. Factors influencing occupancy include environmental features of the site such as temperature, precipitation, elevation and land use. Detection describes the observer's ability to detect the species and depends on the difficulty of identifying the species, the effort put in by the birder, and the current weather conditions. Neglecting to model the detection process can result in misleading models that make poor predictions. Birder expertise fits naturally as an influence on the detection process.

Mackenzie et al. [55] proposed a well-known site occupancy model that separates occupancy from detection. We describe this model, which we refer to as the Occupancy-Detection (OD) model, in detail in Section 3.2.1. In our work, we introduce the Occupancy-Detection-Expertise (ODE) model which extends the OD model by incorporating the expertise of citizen scientists. The benefits of the ODE model are threefold. First, by accounting for birder expertise in the ODE model, we can improve the prediction of observations of a bird species at a site. Second, we can use the ODE model to predict the expertise level of a birder given their history of checklists submitted to eBird. Thirdly, we can use the ODE model to perform a contrast mining task of identifying bird species that novices under/over-report as compared to experts. Ultimately, we would like to account for these sources of bias by novice birders and in doing so, improve the accuracy of species distribution models.

Although the focus of this paper is on SDM, the occupancy / detection problem is a specific instance of a more general problem in which a detection process corrupts a "true" value with noise to produce an observed value. Much of the existing work assumes a simple noise model (e.g. an additive Gaussian noise term), but in some situations the detection process is affected by conditions during detection and requires a more complex model. These situations occur in domains such as object recognition and surveillance. We are interested in applying and extending the graphical models presented in this paper to these domains for future work.

3.2 Methodology

In this section, we first describe the OD model [55, 54]. Then, we describe the extensions made to the OD model to form the ODE model, which incorporates birder expertise. We will use the term *birder* and *observer* interchangeably. In addition, we will use *expert* to denote an experienced citizen scientist and *novice* to denote an inexperienced citizen scientist.

Table 3.1: Notation in the Occupancy-Detection model.

Symbol	Description
N	Number of sites.
T_i	Number of visits at site i.
$\boldsymbol{X}_i$	Occupancy features at site i.
Z_i	Occupancy status (unobserved) of site i.
$\boldsymbol{W}_{it}$	Detection features at site i, visit t.
Y_{it}	Observed presence/absence at site i, visit t.
o_i	Occupancy probability of site i.
d_{it}	True detection probability at site i, visit t.
$\boldsymbol{\alpha}$	Occupancy parameters.
$\boldsymbol{\beta}$	Detection parameters.
λ_o	Occupancy regularization term.
λ_d	Detection regularization term.

3.2.1 The Occupancy-Detection model

In SDMs, the occupancy of a site is the true variable of interest, but this variable is typically only indirectly observed. The OD model separates the concept of occupancy from detection. Models that do not distinctly model these two processes can produce incorrect predictions. For instance, a bird species might be wrongly declared as not occupying a site when in fact, this species is simply difficult to detect because of reclusive behavior during nesting. The detection process allows us to account for two types of errors that cause imperfect observations: false detections, which we have already defined, and false absences. False absences refer to erroneously reporting the absence of the species when the site is in fact occupied by that species. False absences could be due to species that are hard to detect (e.g. due to camouflage), a lack of effort on the part of the observer to detect these species, or simply a lack of experience by the observers in identifying these species.

Figure 3.1 illustrates the OD model for a single species as a graphical model [49], in which nodes represent random variables and directed edges can be interpreted as a direct influence from parent to child. Nodes that are circles are continuous random variables while nodes that are squares are discrete random variables. In addition, shaded nodes denote observed variables and unshaded ones denote latent variables.

As shown in Figure 3.1, the true site occupancy at site i (Z_i) is latent. The dotted boxes in Figure 3.1 represent plate notation used in graphical models in which the contents inside the dotted box are replicated as many times as indicated in the bottom right corner. The outer plate represents N sites. The variable $\boldsymbol{X}_i$ denotes a vector of features that influence the occupancy pattern for the species (e.g. land cover type) and $Z_i \in \{0, 1\}$ denotes the true occupancy status of site i. Site i is surveyed T_i times. The variable $\boldsymbol{W}_{it}$ is a vector of features that affect the detectability of the species (e.g. time of day) and $Y_{it} \in \{0, 1\}$ indicates whether the species was detected ($Y_{it} = 1$) on visit t. A summary of the random variables used in the OD model are given in Table 3.1.

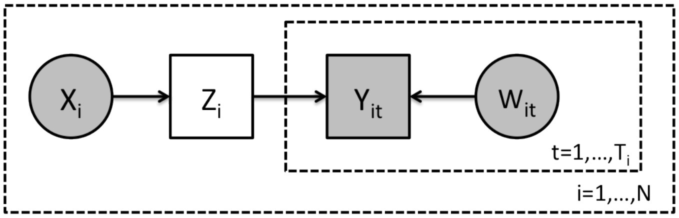


Figure 3.1: Graphical model representation of the Occupancy-Detection model for a single bird species.

The occupancy component of OD model models the occupancy status of the site (i.e. the node Z_i), as a function of the occupancy features associated with the site i. Occupancy features include environmental factors determining the suitability of the site as habitat for the species. Examples of occupancy features include precipitation, temperature, elevation, vegetation and land use. In the OD model, the occupancy probability o_i of site i is related to the occupancy features through a logistic function with parameters $\boldsymbol{\alpha}$. The probability of the occupancy status is modeled as a Bernoulli distribution as shown in Equation 3.1.

$$\begin{aligned} o_i &= \sigma(\boldsymbol{X}_i \cdot \boldsymbol{\alpha}) \\ P(Z_i | \boldsymbol{X}_i; \boldsymbol{\alpha}) &= o_i^{Z_i} (1 - o_i)^{1 - Z_i} \end{aligned} \tag{3.1}$$

The detection component captures the conditional probability of the observer detecting the species (i.e. random variable Y_{it}), during a visit at site i and at time t conditioned on the site being occupied i.e. $Z_i = 1$ and the detection features $\boldsymbol{W}_{it}$. The detection features include factors affecting the observer's detection ability such as weather conditions and factors related to observation effort such as observation duration and route distance. Note that different species have different detection probabilities under the same detection features. For instance, a well-camouflaged, quiet bird requires extra effort to be detected as compared to a loud bird such as a crow. Like the occupancy component, the detection component is parameterized as a logistic function of the detection features as shown in Equation 3.2.

$$\begin{aligned} d_{it} &= \sigma(\boldsymbol{W}_{it} \cdot \boldsymbol{\beta}) \\ P(Y_{it}|Z_i, \boldsymbol{W}_{it}; \boldsymbol{\beta}) &= (Z_i d_{it})^{Y_{it}} (1 - Z_i d_{it})^{1-Y_{it}} \end{aligned} \tag{3.2}$$

Under the OD model, sites are visited multiple times and observations are made during each visit. The site detection history includes the observed presence or absence of the species on each visit at this site. The OD model makes two key assumptions. First, the population closure assumption [55] assumes that the species occupancy status at a site stays constant over the course of the visits. Second, the standard OD model does not allow for false detections. Recall that a false detection occurs when an observer detects a bird of a particular species to be present when in reality, it does not occupy the site. In order to understand the effects of this second assumption, suppose there are 100 visits in which the bird species is not detected. If the bird species is detected on the 101th visit, the site is inferred to be occupied. Hence, reporting the presence of a species at a site indicates the site being occupied. Reporting the absence of a species at a site can be explained by either the site being truly unoccupied or the observer failing to detect the species.

3.2.2 The Occupancy-Detection-Expertise model

The ODE model incorporates birder expertise by extending the OD model in two ways. First, birder expertise strongly influences the detectability of the species. For example, novices are likely to detect bird species by sight and are proficient at identifying common bird species while experts can detect bird species by both sight and sound. As a result, we add to the OD graphical model an expertise component which influences the detection process. The second extension we add to the OD model is to allow false detections by both novices and experts. The occupancy component of the ODE model stays the same as in the OD model because the site occupancy is independent of the observer's expertise.

Table 3.2: Notation in the Occupancy-Detection-Expertise model.

Symbol	Description
M	Number of birders.
$\boldsymbol{U}_j$	Expertise features of birder j.
E_j	Expertise level of birder j.
$B(Y_{it})$	The birder associated with checklist Y_{it}.
v_j	Expertise probability of birder j.
d_{it}^{ex}	True detection probability for expert birders at site i, visit t.
f_{it}^{ex}	False detection probability for expert birders at site i, visit t.
d_{it}^{no}	True detection probability for novice birders at site i, visit t.
f_{it}^{no}	False detection probability for novice birders at site i, visit t.
$\boldsymbol{\gamma}$	Expertise parameters.
$\boldsymbol{\beta}_1^{ex}$	True detection parameters for expert birders.
$\boldsymbol{\beta}_0^{ex}$	False detection parameters for expert birders.
$\boldsymbol{\beta}_1^{no}$	True detection parameters for novice birders.
$\boldsymbol{\beta}_0^{no}$	False detection parameters for novice birders.
$\boldsymbol{\beta}$	The total set of detection parameters (β_1^{ex}, β_0^{ex}, β_1^{no}, β_0^{no}).
λ_e	Expertise regularization term.

A graphical model representation of the ODE model for a single bird species is shown in Figure 3.2. A summary of the random variables and parameters used in the ODE model is given in Table 3.2.

In the expertise component, E_j is a binary random variable capturing the expertise (i.e. 0 for novice, 1 for expert) of the jth birder. The value of E_j is a function of expertise features associated with the birder. Expertise features include features derived from the birder's personal information and history of checklists, such as the total number of checklists submitted by the birder to eBird and the total number of bird species ever identified on these checklists. Once again, we use the logistic function to model the expertise component as follows:

$$\begin{aligned} v_j &= \sigma(\boldsymbol{U}_j \cdot \boldsymbol{\gamma}) \\ P(E_j|\boldsymbol{U}_j;\boldsymbol{\gamma}) &= v_j^{E_j}(1-v_j)^{1-E_j} \end{aligned} \tag{3.3}$$

In order to incorporate birder expertise, we modify the detection process such that it consists of a mixture model in which one mixture component models the detection

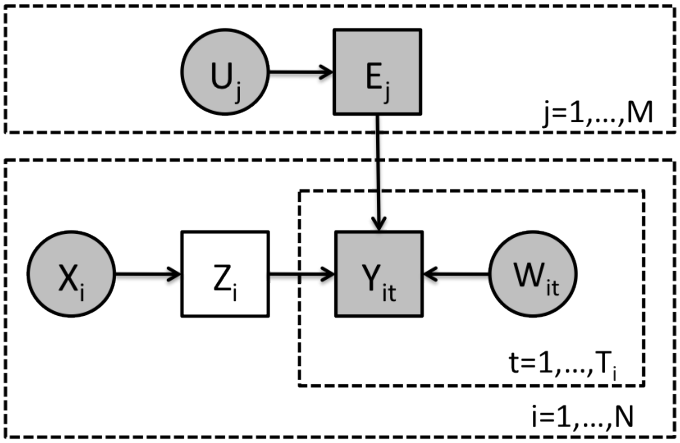


Figure 3.2: Graphical model representation of Occupancy-Detection-Expertise model for a single bird species. Note that the link from E_j to Y_{it} only exists if birder j submits the checklist corresponding to Y_{it}.

probability by experts and the other mixture component models the detection probability by novices. Each detection probability has a separate set of detection parameters for novices and for experts. These two separate feature sets are useful if the detection process is different for experts versus novices. For instance, experts can be very skilled at identifying birds by sound rather than by sight. Let $B(Y_{it})$ be the index of the birder who submits checklist Y_{it}. In Figure 3.2, the links from E_j to Y_{it} only exist if $B(Y_{it}) = j$, i.e. the jth birder is the one submitting the checklist corresponding to Y_{it}.

In addition, we allow for false detections by both experts and novices. This step is necessary because allowing for false detections by experts and novices improves the predictive ability of the model. Experts are in fact often over-enthusiastic about reporting bird species that do not necessarily occupy a site but might occupy a neighboring site. For instance, experts are much more adept at identifying and reporting birds that fly over a site or are seen at a much farther distance from the current site. As a result, the detection probabilities for novices and experts in the ODE model are now separated into a total of 4 parts: a true detection component and a false detection component for experts, and a true detection and a false detection component for novices.

Let $\tilde{P_1^{ex}}(Y_{it})$ be shorthand for the expert true detection probability $P(Y_{it}|Z_i = 1, \boldsymbol{W}_{it}, E_{B(Y_{it})} = 1, \boldsymbol{\beta}_1^{ex})$ and $\tilde{P_0^{ex}}(Y_{it})$ to be shorthand for the expert false detection probability $P(Y_{it}|Z_i = 0, \boldsymbol{W}_{it}, E_{B(Y_{it})} = 1, \boldsymbol{\beta}_0^{no})$. In a similar manner, we use $\tilde{P}_1^{no}(Y_{it})$ and $\tilde{P}_0^{no}(Y_{it})$ for novice true and false detection probabilities. There are now four sets of $\boldsymbol{\beta}$ parameters used in each of the four logistic regressions corresponding to the previous four probabilities: $\boldsymbol{\beta}_1^{ex}$, $\boldsymbol{\beta}_0^{ex}$, $\boldsymbol{\beta}_1^{no}$, and $\boldsymbol{\beta}_0^{no}$. In Equation 3.8, we generically refer to a set of these parameters as $\boldsymbol{\beta}$. The detection probability at site i on visit t conditioned on site occupancy and birder's expertise can be written as follows:

$$\tilde{P}_1^{ex}(Y_{it}) = (d_{it}^{ex})^{Y_{it}}(1 - d_{it}^{ex})^{1-Y_{it}} \text{ where } d_{it}^{ex} = \sigma(\boldsymbol{W}_{it} \cdot \boldsymbol{\beta}_1^{ex}) \tag{3.4}$$

$$\tilde{P}_0^{ex}(Y_{it}) = (f_{it}^{ex})^{Y_{it}}(1 - f_{it}^{ex})^{1-Y_{it}} \text{ where } f_{it}^{ex} = \sigma(\boldsymbol{W}_{it} \cdot \boldsymbol{\beta}_0^{ex}) \tag{3.5}$$

$$\tilde{P}_1^{no}(Y_{it}) = (d_{it}^{no})^{Y_{it}}(1 - d_{it}^{no})^{1-Y_{it}} \text{ where } d_{it}^{no} = \sigma(\boldsymbol{W}_{it} \cdot \boldsymbol{\beta}_1^{no}) \tag{3.6}$$

$$\tilde{P}_0^{no}(Y_{it}) = (f_{it}^{no})^{Y_{it}}(1 - f_{it}^{no})^{1-Y_{it}} \text{ where } f_{it}^{no} = \sigma(\boldsymbol{W}_{it} \cdot \boldsymbol{\beta}_0^{no}) \tag{3.7}$$

$$\begin{aligned} P(Y_{it}|Z_i, \boldsymbol{W}_{it}, E_{B(Y_{it})}; \boldsymbol{\beta}) = & E_{B(Y_{it})}[Z_i \tilde{P}_1^{ex}(Y_{it}) + (1 - Z_i)\tilde{P}_0^{ex}(Y_{it})] + \\ & (1 - E_{B(Y_{it})})[Z_i \tilde{P}_1^{no}(Y_{it}) + (1 - Z_i)\tilde{P}_0^{no}(Y_{it})] \end{aligned} \tag{3.8}$$

3.2.3 Parameter estimation

The ODE model requires a labeled set of expert and novice birders to estimate the model parameters using Expectation Maximization [14]. The EM algorithm maximizes the expected joint log-likelihood shown in Equation 3.9. In the E-step, EM computes the expected occupancies Z_i for each site i using Bayes rule. Let $\boldsymbol{\Theta}^{(t)} = (\boldsymbol{\alpha}^{(t)}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\gamma}^{(t)})$ denote the parameters of the previous iteration. Let $\tilde{P}(Z_i = z_i)$ be shorthand for $P(Z_i = z_i|Y_{i\cdot}, X_i, W_{i\cdot}, E_{B(Y_{i\cdot})}; \boldsymbol{\Theta}^{(t)})$, which is the conditional probability of site i's occupancy. In the previous equation, we use the $i\cdot$ subscript to indicate a random variable affecting all visits to site i. The expected joint log-likelihood is given in Equation 3.9 below.

$$
\begin{aligned}
Q =& \mathbb{E}_{P(\boldsymbol{Z}|\boldsymbol{Y},\boldsymbol{E})}[\log P(\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{E}|\boldsymbol{X},\boldsymbol{U},\boldsymbol{W})] \\
=& \mathbb{E}_{P(\boldsymbol{Z}|\boldsymbol{Y},\boldsymbol{E})}\Bigg[\sum_{j=1}^{M}\log P(E_j|\boldsymbol{U}_j;\boldsymbol{\gamma}) + \sum_{i=1}^{N}\Big[\log P(Z_i|\boldsymbol{X}_i;\boldsymbol{\alpha})\sum_{t=1}^{T_i}\log P(Y_{it}|Z_i,\boldsymbol{W}_{it},E_{B(Y_{it})};\boldsymbol{\beta})\Big]\Bigg] \\
=& \mathbb{E}_{P(\boldsymbol{Z}|\boldsymbol{Y},\boldsymbol{E})}\Bigg\{\sum_{j=1}^{M}[E_j\log(v_j) + (1-E_j)\log(1-v_j)] + \\
& \sum_{i=1}^{N}\Bigg[Z_i\log(o_i) + (1-Z_i)\log(1-o_i) + \sum_{t=1}^{T_i}\log\Big[E_{B(Y_{it})}[Z_i\tilde{P_1^{ex}}(Y_{it}) + (1-Z_i)\tilde{P_0^{ex}}(Y_{it})] + \\
& (1-E_{B(Y_{it})})[Z_i\tilde{P_1^{no}}(Y_{it}) + (1-Z_i)\tilde{P_0^{no}}(Y_{it})]\Big]\Bigg]\Bigg\} \qquad (3.9)
\end{aligned}
$$

There are three regularization parameters $(\lambda_o, \lambda_d, \lambda_e)$ corresponding to the occupancy, detection and expertise components of the ODE model. We regularize these three components using the penalty term in Equation 3.10.

$$
r(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}) = \lambda_o\frac{1}{2}\sum_{i=2}^{|\alpha|}\alpha_i^2 + \lambda_d\frac{1}{2}\sum_{i=2}^{|\beta|}\beta_i^2 + \lambda_e\frac{1}{2}\sum_{i=2}^{|\gamma|}\gamma_i^2 \qquad (3.10)
$$

In the M-step, EM determines the values of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ that maximize Equation 3.9. Using the gradients in Equations 3.11 - 3.13, we apply L-BFGS [53] to perform the optimization. Equation 3.12 is representative of the other parameters $\boldsymbol{\beta}_0^{ex}$, $\boldsymbol{\beta}_1^{no}$, and $\boldsymbol{\beta}_0^{no}$.

$$
\frac{\partial Q}{\partial\boldsymbol{\alpha}} = \sum_{i=1}^{N}\frac{\partial Q}{\partial o_i}\frac{\partial o_i}{\partial\boldsymbol{\alpha}} = \sum_{i=1}^{N}(\tilde{P}(Z_i=1)-o_i)\boldsymbol{X_i} \qquad (3.11)
$$

$$
\begin{aligned}
\frac{\partial Q}{\partial\boldsymbol{\beta}_1^{ex}} =& \sum_{i=1}^{N}\sum_{t=1}^{T_i}\frac{\partial Q}{\partial\boldsymbol{\beta}_{it}^{ex}}\frac{\partial\boldsymbol{\beta}_{it}^{ex}}{\partial\boldsymbol{\beta}_1^{ex}} \\
=& \sum_{i=1}^{N}\sum_{t=1}^{T_i}\tilde{P}(Z_i=1)\frac{E_{B(Y_{it})}\tilde{P_1^{ex}}(Y_{it})(Y_{it}-d_{it}^{ex})\boldsymbol{W}_{it}}{E_{B(Y_{it})}\tilde{P_1^{ex}}(Y_{it}) + (1-E_{B(Y_{it})})\tilde{P_1^{no}}(Y_{it})} \qquad (3.12)
\end{aligned}
$$

$$
\frac{\partial Q}{\partial\boldsymbol{\gamma}} = \sum_{j=1}^{M}\frac{\partial Q}{\partial v_j}\frac{\partial v_j}{\partial\boldsymbol{\gamma}} = \sum_{j=1}^{M}(E_j-v_j)\boldsymbol{U_j} \qquad (3.13)
$$

An *identifiability* problem [71] arises when estimating ODE model parameters. This identifiability issue causes two symmetric but distinct sets of parameter values to be solutions to the EM procedure. While both of these solutions are mathematically valid, one solution yields a model that is more consistent with real world assumptions than the other. We address this issue by adding a constraint during training that biases EM towards the more desirable solution. This constraint encodes the fact that experts always have a higher true detection probability than false detection probability, meaning that experts are more likely to detect a species when the site is truly occupied than falsely detecting the species when the site is unoccupied.

3.2.4 Inference

The ODE model can be used for three main inference tasks: prediction of site occupancy (Z_i), prediction of observations on a checklist (Y_{it}) and prediction of a birder's expertise (E_j). We describe these tasks in more detail below.

3.2.4.1 Prediction of site occupancy

Ecologists are most interested in the true species occupancy at a site. We can use the ODE model to compute the probability that the site is occupied given the site features, the detection history at that site, and the expertise features for each birder submitting checklists at that site. Let $\boldsymbol{E^i}$ be the set of birders submitting checklists at site i and let the expertise of the birders in $\boldsymbol{E^i}$ be unobserved. The occupancy probability of site i can be computed using Equation 3.14.

$$\begin{aligned} P(Z_i = 1|\boldsymbol{X}_i, \boldsymbol{Y}_{i\cdot}, \boldsymbol{W}_{i\cdot}, \boldsymbol{U}) &= \frac{P(\boldsymbol{Y}_{i\cdot}, Z_i = 1|\boldsymbol{X}_i, \boldsymbol{W}_{i\cdot}, \boldsymbol{U})}{\sum_{z_i \in \{0,1\}} P(\boldsymbol{Y}_{i\cdot}, Z_i = z_i|\boldsymbol{X}_i, \boldsymbol{W}_{i\cdot}, \boldsymbol{U})} \\ &= \frac{\sum_{\boldsymbol{e^i}} P(\boldsymbol{Y}_{i\cdot}, Z_i = 1, \boldsymbol{E^i} = \boldsymbol{e^i}|\boldsymbol{X}_i, \boldsymbol{W}_{i\cdot}, \boldsymbol{U})}{\sum_{z_i \in \{0,1\}} \sum_{\boldsymbol{e^i}} P(\boldsymbol{Y}_{i\cdot}, Z_i = z_i, \boldsymbol{E^i} = \boldsymbol{e^i}|\boldsymbol{X}_i, \boldsymbol{W}_{i\cdot}, \boldsymbol{U})} \end{aligned} \tag{3.14}$$

where

$$\begin{aligned} &P(\boldsymbol{Y}_{i\cdot}, Z_i = z_i, \boldsymbol{E^i} = \boldsymbol{e^i} | \boldsymbol{X}_i, \boldsymbol{W}_{i\cdot}, \boldsymbol{U}) \\ &= P(Z_i = z_i | \boldsymbol{X}_i, \boldsymbol{\alpha}) \cdot \prod_{j=1}^{|\boldsymbol{E}^i|} P(E^i_j = e^i_j | \boldsymbol{U}_j; \boldsymbol{\gamma}) \cdot \prod_{t=1}^{T_i} P(Y_{it} | Z_i = z_i, \boldsymbol{W}_{it}, E^i_{B(Y_{it})} = e^i_{B(Y_{it})}; \boldsymbol{\beta}) \end{aligned}$$

Although determining the true site occupancy is the most important inference tasks for ecologists, ground truth on site occupancy is typically unavailable, especially in real-world species distribution data. In order to compare different species distribution models, the observation (i.e. detection) of a species at a site is often used as a substitute for the true site occupancy. Therefore, in order to evaluate our ODE model, we evaluate its performance on predicting Y_{it} in section 3.2.4.2.

3.2.4.2 Predicting observations on a checklist

When predicting Y_{it}, the expertise level of the birders is not recorded in eBird. As a result, we treat the expertise node E_j as a latent variable. We compute the detection probability Y_{it} as shown in Equation 3.15.

$$\begin{aligned} &P(Y_{it} = 1 | \boldsymbol{X}_i, \boldsymbol{W}_{it}, \boldsymbol{U}_{B(Y_{it})}) \\ &= \sum_{z_i \in \{0,1\}} \sum_{e \in \{0,1\}} P(Y_{it} = 1, Z_i = z_i, E_{B(Y_{it})} = e | \boldsymbol{X}_i, \boldsymbol{W}_{it}, \boldsymbol{U}_{B(Y_{it})}) \end{aligned} \tag{3.15}$$

where

$$\begin{aligned} &P(Y_{it} = 1, Z_i = z_i, E_{B(Y_{it})} = e | \boldsymbol{X}_i, \boldsymbol{W}_{it}, \boldsymbol{U}_{B(Y_{it})}) \\ &= P(E_{B(Y_{it})} = e | \boldsymbol{U}_{B(Y_{it})}; \boldsymbol{\gamma}) \cdot P(Z_i = z_i | \boldsymbol{X}_i; \boldsymbol{\alpha}) \cdot P(Y_{it} = 1 | Z_i = z_i, \boldsymbol{W}_{it}, E_{B(Y_{it})} = e; \boldsymbol{\beta}) \end{aligned}$$

3.2.4.3 Predict birder's expertise

In the eBird project, the expertise of the birders is typically unlabeled and prediction of the expertise E_j for birder j can alleviate the burden of manually classifying the new birders into experts and novices. Let $\boldsymbol{Y}^j$ be the set of checklists that belong to birder j (with $\boldsymbol{Y^j_{it}}$ and $\boldsymbol{Y^j_{i\cdot}}$ extending our previous notation), let $\boldsymbol{W^j_{it}}$ be the detection features for $\boldsymbol{Y_{it}^j}$ and let $\boldsymbol{Z}^j$ be the set of sites that birder j submitted checklists at. Since Z_i^j is a latent variable, we predict the expertise of birder j as shown in Equation 3.16.

$$\begin{aligned} P(E_j = 1|\boldsymbol{X}, \boldsymbol{Y}^j, \boldsymbol{W}, \boldsymbol{U}_j) &= \frac{P(E_j = 1, \boldsymbol{Y}^j|\boldsymbol{X}, \boldsymbol{W}, \boldsymbol{U}_j)}{\sum_{e_j \in \{0,1\}} P(E_j = e_j, \boldsymbol{Y}^j|\boldsymbol{X}, \boldsymbol{W}, \boldsymbol{U}_j)} \\ &= \frac{\sum_{\boldsymbol{z^j}} P(E_j = 1, \boldsymbol{Y}^j, \boldsymbol{Z}^j = \boldsymbol{z}^j|\boldsymbol{X}, \boldsymbol{W}, \boldsymbol{U}_j)}{\sum_{e_j \in \{0,1\}} \sum_{\boldsymbol{z^j}} P(E_j = e_j, \boldsymbol{Y}^j, \boldsymbol{Z}^j = \boldsymbol{z}^j|\boldsymbol{X}, \boldsymbol{W}, \boldsymbol{U}_j)} \end{aligned} \tag{3.16}$$

where

$$\begin{aligned} &P(E_j = e_j, \boldsymbol{Y}^j, \boldsymbol{Z}^j = \boldsymbol{z}^j|\boldsymbol{X}, \boldsymbol{W}, \boldsymbol{U}_j) \\ &= P(E_j = e_j|\boldsymbol{U}_j, \boldsymbol{\gamma}) \prod_{i=1}^{|\boldsymbol{Z}^j|} P(Z_i^j = z_i^j|\boldsymbol{X}_i, \boldsymbol{\alpha}) \cdot \prod_{t=1}^{|\boldsymbol{Y_{i\cdot}^j}|} P(Y_{it}^j|Z_i^j = z_i^j, \boldsymbol{W}_{it}^j, E_j = e_j, \boldsymbol{\beta}) \end{aligned}$$

3.3 Evaluation and Discussion

In this section, we evaluate the ODE model over two prediction tasks: predicting observations on a birder's checklist and predicting the birder's expertise level based on the checklists submitted by the birder. In both evaluation tasks, we report the area under the ROC curve (AUC) as the evaluation metric. We also include results from a contrast mining task that illustrates the utility of the ODE model.

3.3.1 Data description

The eBird dataset consists of a database of checklists associated with a geographic site. Each checklist belongs to a specific birder and one checklist is submitted per visit to a site by a birder. In addition, each checklist stores the counts of all the bird species observed at that site by that birder. We convert the counts for each species into a Boolean presence/absence value. A number of other features are also associated with each site-checklist-birder combination: 1) the occupancy features associated with each site, 2) the detection features associated with each observation (which is an entry in a checklist for that specific bird species), and 3) the expertise features associated with each birder. The observation history of each birder is used to construct two expertise features – the total number of checklists submitted and the total number of bird species

Table 3.3: Occupancy, Detection and Expertise features in eBird data set.

Occupancy Features	Comments
Population	Population per square mile.
Housing density	Number of housing units per square mile.
Housing percent vacant	Percentage of housing units.
Elevation	Elevation in meters from National Elevation Dataset.
Habitat_X	Percent of surrounding landscape that is habitat class X. There are 15 habitat classes.

Detection Features	Comments
Time of day	Time when observation started, ranging over [0; 24).
Observation duration	Duration of observation for the checklist, in hours.
Route distance	Distance traveled during observation period, in kilometers.

Expertise Features	Comments
Number of Checklists	Number of checklists submitted by a birder.
Number of species	Number of species identified by a birder.

identified. Table 3.3 shows 19 occupancy features, 3 detection features and 2 expertise features we use in the experiment. For more details on the occupancy and detection features, we refer the readers to the eBird Manual [60].

In our experiments we use eBird data from New York state during the breeding season (May to June) in years 2006-2008. We choose the breeding season because many bird species are more easily detected during breeding and because the population closure assumption is reasonably valid during this time period. Furthermore, we group the checklists within a radius of 0.16 km of each other into one site and each checklist corresponds to one visit at that grouped site. The radius is set to be small so that the site occupancy is constant across all the checklists associated with that grouped site. Checklists associated with the same grouped site but from different years are considered to be from different sites. We train on a training set with the expertise of birders hand-labeled by ornithologists working with the eBird project at the Cornell Lab of Ornithology. The birder expertise was determined through a variety of methods including personal contact, reputation in the birding community, number of checklists rejected during the data verification process, and manual inspection of checklists submitted to eBird. This training set consists of 32 expert birders (with 2352 checklists in total) and

Table 3.4: Bird species in each group.

Category	Bird Species
Group A	Blue Jay White-breasted Nuthatch Northern Cardinal Great Blue Heron
Group B	Brown Thrasher Blue-headed Vireo Northern Rough-winged Swallow Wood Thrush
Group C	Hairy Woodpecker Downy Woodpecker Purple Finch House Finch

88 novice birders (with 2107 checklists in total).

There are roughly 400 bird species that have been reported over the NY state area. Each bird species can be considered a different prediction problem. We evaluate our results over 3 groups with 4 bird species each as shown in Table 3.4. Group A consists of common bird species that are easily identified by novices and experts alike. Group B consists of bird species that are difficult for novices to detect. Experts typically detect Brown Thrashers, Blue-headed Vireos and Wood Thrushes by sound rather than sight. The Northern Rough-winged Swallow is extremely hard to identify because it flies very quickly and has subtle distinguishing traits that novices are usually unfamiliar with. Finally, Group C consists of two pairs of birds – Hairy and Downy Woodpeckers and Purple and House Finches. Novices typically confuse members of a pair for each other.

3.3.2 Task 1: Prediction of observations on a checklist

Since the *occupancy* status of the site Z_i is not available, we can use the *observation* of a bird species as a substitute. We evaluate the accuracy of the ODE model when predicting detections versus two other baseline models: a Logistic Regression model that does not separate the occupancy and the detection process and the classic OD model found in the ecology literature.

Evaluating predictions on spatial data is a challenging problem due to two key issues. First, a non-uniform spatial distribution of the data introduces a bias in which small regions with high sampling intensity have a very strong influence on the performance of the model. Secondly, spatial autocorrelation allows test data points that are close to training data points to be easily predicted by the model. To alleviate the effects of both of these problems, we superimpose a 9-by-16 checkerboard (each grid cell is roughly a 50 km x 33 km rectangle) over the data. The checkerboard grids the NY state region into black and white cells. Data points falling into the black cells are grouped into one fold and those falling into the white cells are grouped into another fold. The black and white sets are used in a 2-fold cross validation. We also randomize the checkerboarding by randomly positioning the bottom left corner to create different datasets for the two folds. We run 20 such randomization iterations and within each iteration, we perform a 2-fold cross validation. We compute the average AUC across all 20 runs and show the results in Table 3.5. Boldface results indicate the winner, $\star$ and $\dagger$ indicate ODE model is statistically significant better than LR and OD model with paired t-test.

We use a validation set to tune the regularization terms of three different models. Data in one fold is divided into a training set and a validation set by using a 2-by-2 checkerboard on each cell. More specifically, each cell is further divided into a 2-by-2 subgrid, in which the top left and bottom right subgrid cells are used for training and the top right and bottom left subgrid cells are used for validation. We evaluate all combinations of values $\{0, 0.001, 0.01, 0.1, 1\}$ for the regularization terms on the validation set, using the set of values that produce the best AUC on the validation set. For values of the regularization term less than 0.01, the results do not change by much.

1. Logistic Regression Model (LR): A typical machine learning approach to this problem is to combine the occupancy and detection features into a single set of features rather than separating occupancy and detection into two separate processes and modeling occupancy as a latent variable. Since we are interested in the benefit of distinctly modeling occupancy and detection by having occupancy as a latent variable, we use a baseline of a Logistic Regression model. Logistic Regression is a special case of a GLM, which is a common class of methods used for SDM by ecologists [2].

To set up this baseline algorithm, we use two LR models. The first LR model predicts the expertise of a birder using the expertise features of that birder. The probability of the birder being an expert is then treated as a feature associated with each checklist from that birder. The second LR predicts the detection Y_{it} using the occupancy features, detection features and the expertise probability computed from the first LR. We regularize those two LR models with L2-norm. Again we evaluate all combinations of values $\{0, 0.001, 0.01, 0.1, 1\}$ on the validation set and pick up the set of values that generate the best AUC.

2. OD Model: In order to incorporate birder expertise in the OD model, we also employ a LR to predict the birder expertise from the expertise features. We treat the probability of the birder being an expert as another detection feature associated with each checklist from that birder. Then, we use EM to train the OD model. To predict a detection, we first compute the expertise probability using coefficients from the first LR and then predict the detection as in Equation 3.17 using the occupancy features, detection features and the predicted expertise as an additional detection feature. There are three regularization parameters corresponding to the first LR model, occupancy component and detection component of the OD model. Similarly we use L2-norm for all three regularization terms. The best set of values in all combinations of values $\{0, 0.001, 0.01, 0.1, 1\}$ are chosen based on the validation set.

$$\begin{aligned} P(Y_{it} = 1 | \boldsymbol{X}_i, \boldsymbol{W}_{it}; \boldsymbol{\alpha}, \boldsymbol{\beta}) &= \sum_{z_i \in \{0,1\}} P(Z_i = z_i | \boldsymbol{X}_i; \boldsymbol{\alpha}) P(Y_{it} = 1 | Z_i = z_i, \boldsymbol{W}_{it}; \boldsymbol{\beta}) \\ &= P(Z_i = 1 | \boldsymbol{X}_i; \boldsymbol{\alpha}) P(Y_{it} = 1 | Z_i = 1, \boldsymbol{W}_{it}; \boldsymbol{\beta}) \end{aligned} \tag{3.17}$$

3. ODE Model: The ODE model is trained using EM and the prediction of the detection random variable Y_{it} is based on Equation 3.15. The birder expertise is observed during training but unobserved during testing.

Since true site occupancies are typically not available for real-world species distribution data sets, predicting species observations at a site is a reasonable substitute for evaluating the performance of a SDM. Table 3.5 indicates that the top performing model over all 12 species is the ODE model. The ODE model offers a statistically significant improvement over LR in all 12 species and over the OD model in 10 species. The two main advantages that the OD model has over LR are that it models occupancy separately from detection and it allows checklists from the same site i to share evidence through

Table 3.5: Average AUC for predicting detections on test set checklists for bird species.

Group A Bird Species	LR	OD	ODE
Blue Jay	0.6726	0.6881	**0.7104***†
White-breasted Nuthatch	0.6283	0.6262	**0.6600***†
Northern Cardinal	0.6831	0.7073	**0.7085***
Great Blue Heron	0.6641	0.6691	**0.6959***†

Group B Bird Species	LR	OD	ODE
Brown Thrasher	0.6576	0.6920	**0.6954***
Blue-headed Vireo	0.7976	0.8055	**0.8325***†
Northern Rough-winged Swallow	0.6575	0.6609	**0.6872***†
Wood Thrush	0.6579	0.6643	**0.6903***†

Group C Bird Species	LR	OD	ODE
Hairy Woodpecker	0.6342	0.6283	**0.6759***†
Downy Woodpecker	0.5960	0.5622	**0.6183***†
Purple Finch	0.7249	0.7458	**0.7659***†
House Finch	0.5725	0.5809	**0.6036***†

the latent variable Z_i. However, in 3 species, the OD model performs worse than the LR model. This decrease in AUC is largely due to the fact that the OD model does not allow for false detections. In contrast to the OD model, the ODE model allows for false detections by both novices and experts and it can incorporate the expertise of the observer into its predictions. Since the ODE model consistently outperforms the OD model, the improvement in accuracy is mainly due to these two advantages.

3.3.3 Task 2: Prediction of birder's expertise

Automated prediction of a birder's expertise can alleviate the onerous task of manually classifying a new birder as an expert or novice. In this experiment, we compare the ODE model with a Logistic Regression to predict the birder's expertise.

1. Logistic Regression Model (LR): To train a LR to predict a birder's expertise, every checklist is treated as a single data instance. The set of features for each data instance include occupancy features, detection features, and expertise features. To predict the expertise of a new birder, we first retrieve the checklists submitted by the birder, predict the birder's expertise on each checklist using LR, and then the predictions of expertise on each checklist are averaged to give the final probability. Again we use L2-norm to regularize the LR model and choose the best value in $\{0, 0.001, 0.01, 0.1, 1\}$ based on the validation set.

2. ODE Model: The ODE model is trained using EM and the prediction of birder's expertise is based on Equation 3.16.

We evaluate on the same twelve bird species using a 2-fold cross validation across birders. We randomly divide the expert birders and novice birders into half so that we have an equal number of expert birders as well as novice birders in the two folds. Assigning birders to each fold will assign checklists associated with each birder to the corresponding fold. We use a validation set to tune the regularization terms of both the LR model and the ODE model. Of all birders in the "training" fold, half of the expert birders and the novice birders in that fold are randomly chosen as the actual training set and the other half serve as the validation set. We tune the regularization terms of both the LR model and the ODE model using the range of values $\{0, 0.001, 0.01, 0.1, 1\}$ over each of the regularization parameters. Finally, we run 2-fold cross validation on the two folds and compute the AUC. For each bird species, we perform the 2-fold cross validation using 20 different random splits for the folds. In Table 3.6 we tabulate the mean AUC for each species, with boldface entries indicating the winner and $\star$ indicating statistically significant improvement using paired t-test.

When predicting expertise, the ODE model outperforms LR on all species except for *White-breasted Nuthatch* as shown in Table 3.6. The ODE model's results are statistically significant for Group B birds species, which are hard to detect, but not significant for Group A birds, which are common and much more obvious to detect. For Group C, the ODE model results are statistically significant for Hairy Woodpeckers, Purple Finches and House Finches. These results are consistent with behavior by novice birders. Both Purple Finch and Hairy Woodpeckers are rarer and experienced birders can identify them. In contrast, novices often confuse House Finches for Purple Finches and Downy Woodpeckers for Hairy Woodpeckers. Overall, the AUCs for most species are within the 0.70-0.80 range, which is an encouraging result for using the ODE model to predict the expertise of a birder.

Table 3.6: Average AUC for predicting birder expertise on a test set of birders for bird species.

Group A Bird Species	LR	ODE
Blue Jay	0.7265	**0.7417***
White-breasted Nuthatch	**0.7249**	0.7212
Northern Cardinal	0.7352	**0.7442**
Great Blue Heron	0.7472	**0.7661**
Group B Bird Species	LR	ODE
Brown Thrasher	0.7523	**0.7761***
Blue-headed Vireo	0.7869	**0.7981**
Northern Rough-winged Swallow	0.7792	**0.8052***
Wood Thrush	0.7675	**0.7937***
Group C Bird Species	LR	ODE
Hairy Woodpecker	0.7056	**0.7334***
Downy Woodpecker	0.7223	**0.7307**
Purple Finch	0.7481	**0.7739***
House Finch	0.7279	**0.7403***

3.3.4 Task 3: Contrast mining

In this contrast mining task, we identify bird species that are over/under reported by novices compared to experts. We compare the average Δ_{TD} values for Groups A and B, where Δ_{TD} is the difference of the true detection probabilities between expert and novice birders. We expect experts and novices to have similar true detection probabilities on species from Group A, which correspond to common, easily identified bird species. For Group B, which consists of species that are hard to detect, we expect widely different true detection probabilities. In order to carry out this case study, we first train the ODE model over all the data described in Subsection 3.3.1 for a particular species. Then for each checklist, we compute the difference between the expert's true detection probability and the novice's true detection probability. We average this value over all the checklists. The results are shown in Table 3.7.

In the contrast mining task, the results in Table 3.7 indicate that experts and novices appear to have very similar true detection probabilities for the common bird species in Group A. However, for the hard-to-detect bird species in Group B, the difference between the true detection probabilities of experts and novices are much larger. These results

Table 3.7: Average Δ_{TD} for Group A and B.

Group A Bird Species	Average Δ_{TD}
Blue Jay	0.0118
White-breasted Nuthatch	0.0077
Northern Cardinal	-0.0218
Great Blue Heron	0.0110

Group B Bird Species	Average Δ_{TD}
Brown Thrasher	0.1659
Blue-headed Vireo	0.1158
Northern Rough-winged Swallow	0.1618
Wood Thrush	0.0954

show that the ODE model is a promising approach for contrast mining, which can identify differences in how experts and novices report bird species.

3.4 Conclusion

We have presented the ODE model that has distinct components that capture occupancy, detection and observer expertise. We have shown that it produces more accurate predictions of species detections and birder's expertise than other models. More importantly, we can use this model to find differences between expert and novice observations of birds. This knowledge can be used to inform citizen scientists who are novice birders and thereby improve the reliability of their observations.

Chapter 4: Automated Data Verification in a Large-scale Citizen Science Project: a Case Study

4.1 Introduction

Recruiting volunteers in broad-scale citizen science projects to gather biodiversity information can generate enormous quantities of data across broad spatial and temporal domains [80]. However, maximizing the information gathered from citizen-science data depends on finding the proper balance between data quantity and quality [40]. 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 [59]. 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 [84].

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.

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 (e.g. unlike in [52]). 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. 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.

Our strategy for addressing data quality in eBird project 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 as our exemplar broad-scale citizen science project [80]. 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 [86] 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 [86] 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 [86], 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). 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.

4.2 Methodology

In this section, we first describe the expert-defined filter in eBird and illustrate the challenges due to the tremendous growth of observations submitted to eBird. Then, we describe the two-step automated data filter, which helps address the challenges of the expert-defined filter.

4.2.1 Expert-defined filter

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.

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.

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 2.1. Finally, due to the expansion of eBird, data filters need to be created for new regions without existing regional experts.

4.2.2 Automated data filter

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

4.2.2.1 Emergent data filter

The eBird database currently holds more than 100 million bird observations. These historical records can be used to filter unusual observations that require review, but allow entry of expected species within the expected times when species should occur. In particular, 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 4.1 shows an example of Chipping Sparrow's weekly

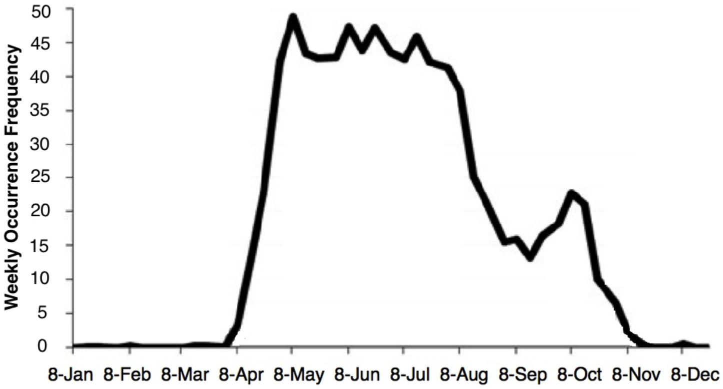


Figure 4.1: The weekly occurrence frequency for Chipping Sparrow in Tompkins Co., NY.

occurrence frequency across a year window in Tompkins Co., NY. The Chipping Sparrow (*Spizella passerina*) is a common breeding bird in upstate New York, but departs the region in the fall and rarely occurs in winter.

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 (e.g. 5%). The emergent data filter flags the observations of species falling below the threshold and processes the observations with the data quality model.

4.2.2.2 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 [26].

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 Chapter 3 because it was successful at accurately predicting an observer's expertise. Once we learn the ODE model from eBird data, we can perform inference to predict a birders' expertise based on their submissions. More details on predicting a birder's expertise is given in Section 3.2.4.3.

4.3 Evaluation

4.3.1 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.

4.3.2 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%.

4.3.3 Estimating an observer's expertise level

For our analysis, we used observations from the original eBird Reference Data [60] 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.

4.4 Results and Discussion

4.4.1 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 4.3 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. For example, Chipping Sparrows (see Figure 4.3 B) 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.

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.

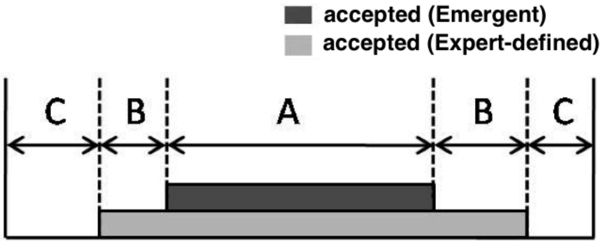


Figure 4.2: An illustration of the time periods covered by the expert-defined filter (light grey bar) and the emergent filter (dark grey bar). Observations falling with the bars are automatically accepted. Observations falling outside of the bars are flagged for review.

Figure 4.2 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 4.2 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

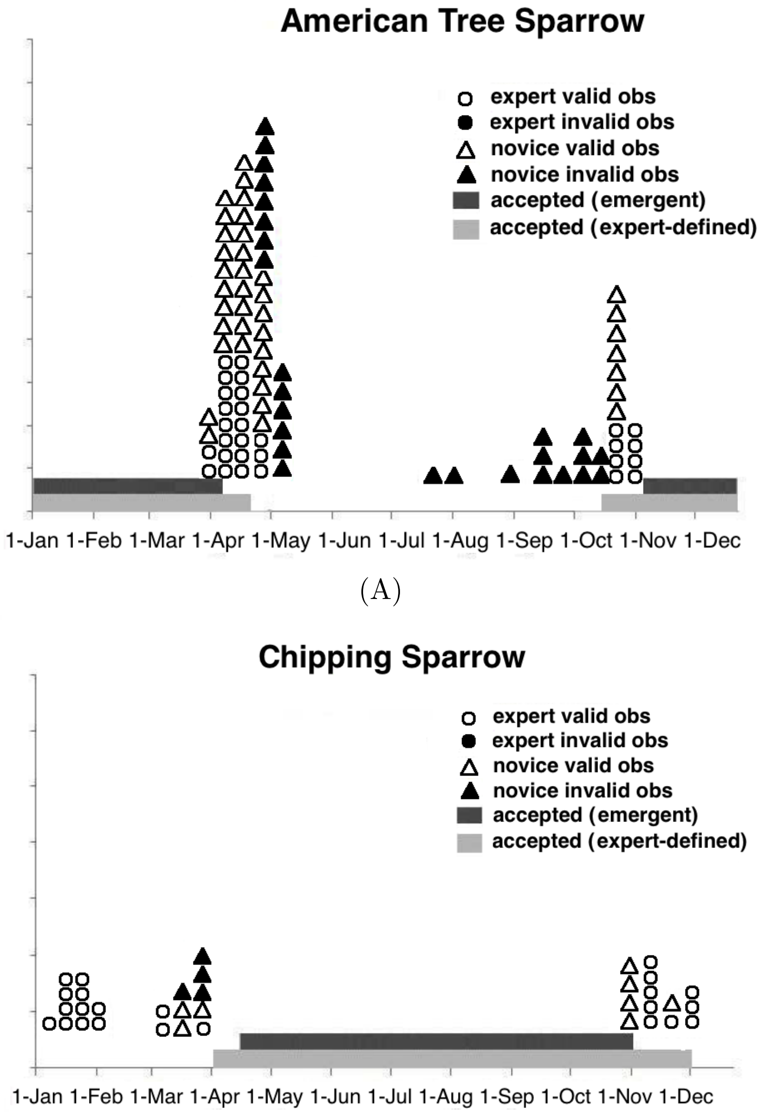
Finally, records falling in region C were flagged by both filters. Unlike for regions A and B, these records were in fact reviewed by experts and then designated to be either valid and added to eBird or designated invalid and discarded. We used the validity of these records as a measure of the accuracy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

4.4.2 The automated data filter

In Figures 4.3, 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 4.1 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 4.2 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.



(A)

(B)

Figure 4.3: Flagged observations for (A) American Tree Sparrow and (B) Chipping 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 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.

Table 4.1: Example of 9 bird species with different occurrence patterns in Tompkins Co., NY. The second and the 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.

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 4.4 (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

Table 4.2: 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

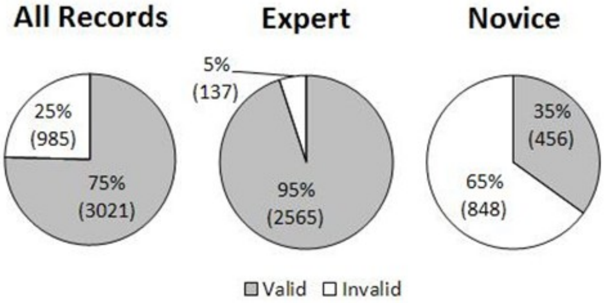


Figure 4.4: The fraction of valid and invalid records among all records in region C (left), then broken down by expert records (center) and by novice records (right). The number of records in each pie slice is shown in parentheses.

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.

4.5 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.

Chapter 5: Clustering Species Accumulation Curves to Identify Groups of Citizen Scientists with Similar Skill Levels

5.1 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 [80, 47], 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 [40] and conservation [63].

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.

To better leverage the citizen science data, it is important to cluster citizen scientists by their skill levels. Identifying groups of citizen scientists with similar skill levels can help understand behaviors between different groups of citizen scientists [73], develop automated data quality filters [85] and build more accurate species distribution models [86]. However, grouping citizen scientists by their skill levels can be challenging in some citizen science projects.

In a citizen science project, participants may play the role of either a *processor* or a *sensor*. When citizen scientists act as *processors*, the same processing tasks are usually repeated and multiple volunteers can be assigned to work on the same task by design. For example, Zooniverse [74] uses volunteers as processors by having them classify or extract information from images. The validation of the process can be through consensus of responses, or directly from an expert [75]. Thus the skill level of a citizen scientist can be measured based on the validity of one's finished tasks, allowing us to group participants based on their skill levels.

However, grouping citizen scientists is challenging when they act as *sensors* in a citizen science project. When they act as *sensors*, ground truth is rarely available and events can neither be repeated nor independently observed. Since their skill levels can not be measured by validating their finished tasks, it is challenging to group citizen scientists in citizen projects like eBird. In eBird, participants actively collect bird observational data over a broad spatial and temporal extent. Most sites are surveyed by a few participants, and there is no ground truth of species occupancies at a site to validate one's submissions.

In this work, we propose to identify a citizen scientist's skill level using species accumulation curves (SACs) [34]. In ecology, the SAC is a graph plotting the cumulative number of unique species observed as a function of the cumulative effort expended (e.g. time). SACs are typically used in the ecological literature to quantify species richness and to identify significant areas for conservation [9]. However, we repurpose the use of SACs as an effective measure of an observer's skill level to detect species. 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 previous study in the eBird project showed that SACs could distinguish eBird observers with different levels of participation and capture the evolution of eBird participants' skills over time.

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 and propose a learning algorithm based on Expectation-Maximization. These clusters can be used to classify birders into different skill levels, which can then be used to develop automated data quality filters [85] and to track how the skills of individual birders evolve over time. In our empirical study, we apply our clustering algorithm to eBird data and show that the skill levels corresponding to the resulting clusters are meaningful. Although we focus on eBird data in this study, the mixture model can be easily applied to other citizen science projects to measure a participant's involvement.

Table 5.1: Notation in the mixture of SACs model

Symbol	Description
M	Number of observers.
N_i	Number of checklists submitted by birder i.
K	Number of groups.
Z_i	Group membership (unobserved) of birder i.
$\boldsymbol{X}_{ij}$	effort birder i expends on checklist j.
Y_{it}	Number of unique species reported on checklist j of birder i.
$\boldsymbol{\beta}_k$	Parameters of group k.

5.2 Methodology

In this section, we first introduce the mixture of SACs model in the graphical model representation. Then we present a learning algorithm for the mixture model using Expectation-Maximization and show how to determine the number of components in the mixture model. Finally, we illustrate how to cluster a new birder based on his or her previous submissions.

5.2.1 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 observers in the same group are at similar skill levels. As eBird is our application domain, we use *observer* and *birder* interchangeably. Figure 5.1 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. A summary of the random variables used in the mixture of SACs model are given in Table 5.1.

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

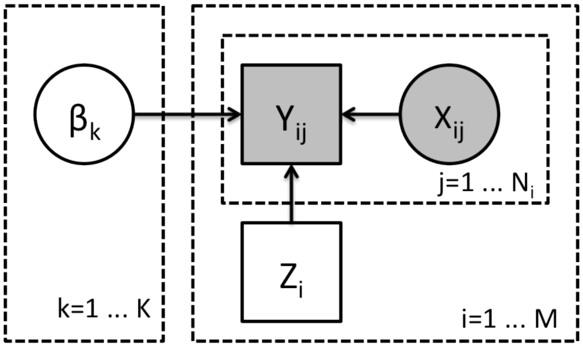


Figure 5.1: The mixture of Species Accumulation Curves model.

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 $\boldsymbol{\beta}_{Z_i} \cdot \boldsymbol{X}_{ij}$ where $\boldsymbol{\beta}_{Z_i}$ are the parameters of group Z_i. 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 5.1.

$$\begin{aligned}
\log P(\boldsymbol{Y}|\boldsymbol{X};\boldsymbol{\pi},\boldsymbol{\beta},\sigma^2) &= \sum_{i=1}^{M} \log P(\boldsymbol{Y}_{i\cdot}|\boldsymbol{X}_{i\cdot};\boldsymbol{\pi},\boldsymbol{\beta},\sigma^2) \\
&= \sum_{i=1}^{M} \log \left(\sum_{k=1}^{K} P(\boldsymbol{Y}_{i\cdot}, Z_i = k|\boldsymbol{X}_{i\cdot};\boldsymbol{\pi},\boldsymbol{\beta},\sigma^2) \right) \\
&= \sum_{i=1}^{M} \log \left(\sum_{k=1}^{K} P(Z_i = k;\boldsymbol{\pi}) \prod_{j=1}^{N_i} P(Y_{ij}|\boldsymbol{X}_{ij}, Z_i = k;\boldsymbol{\beta},\sigma^2) \right)
\end{aligned} \tag{5.1}$$

5.2.2 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 [14]. The EM algorithm iterates between performing the E-step and M-step until the difference of parameters between two consecutive iterations is below some threshold ϵ. 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 5.2. In addition, let r_{ik} denote the expected group membership of birder i belonging to group k.

$$\begin{aligned}
\mathcal{Q} &= E_{\boldsymbol{Z}|\boldsymbol{Y},\boldsymbol{X}}[\log(P(\boldsymbol{Y}, \boldsymbol{Z}|\boldsymbol{X}; \boldsymbol{\pi}, \boldsymbol{\beta}, \sigma^2))] \\
&= \sum_{i=1}^{M} E_{\boldsymbol{Z}|\boldsymbol{Y},\boldsymbol{X}}\left[\log \prod_{k=1}^{K} \Big(P(Z_i = k; \boldsymbol{\pi}) P(\boldsymbol{Y}_{i\cdot}|\boldsymbol{X}_{i\cdot}, Z_i = k; \boldsymbol{\beta}, \sigma^2)\Big)^{\mathbb{I}(Z_i=k)}\right] \\
&= \sum_{i=1}^{M}\sum_{k=1}^{K} E_{\boldsymbol{Z}|\boldsymbol{Y},\boldsymbol{X}}[\mathbb{I}(Z_i = k)] \log\left(P(Z_i = k; \boldsymbol{\pi}) \prod_{j=1}^{N_i} P(Y_{ij}|\boldsymbol{X}_{ij}, Z_i = k; \boldsymbol{\beta}, \sigma^2)\right) \\
&= \sum_{i=1}^{M}\sum_{k=1}^{K} r_{ik} \left[\log(P(Z_i = k; \boldsymbol{\pi})) + \sum_{j=1}^{N_i} \log(P(Y_{ij}|\boldsymbol{X}_{ij}, Z_i = k; \boldsymbol{\beta}, \sigma^2))\right] \\
&= \sum_{i=1}^{M}\sum_{k=1}^{K} r_{ik} \left[\log \pi_k + \sum_{j=1}^{N_i} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{(Y_{ij} - \boldsymbol{\beta}_k \cdot \boldsymbol{X}_{ij})^2}{2\sigma^2}\Big)\right)\right] \\
&= \sum_{i=1}^{M}\sum_{k=1}^{K} r_{ik} \log \pi_k + \sum_{i=1}^{M}\sum_{k=1}^{K} r_{ik} \sum_{j=1}^{N_i} \left(-\frac{(Y_{ij} - \boldsymbol{\beta}_k \cdot \boldsymbol{X}_{ij})^2}{2\sigma^2} - \log(\sqrt{2\pi\sigma^2})\right)
\end{aligned} \tag{5.2}$$

In the E-step, we keep the parameters $\{\boldsymbol{\pi}, \boldsymbol{\beta}, \sigma^2\}$ fixed and update the expected group membership r_{ik} for every birder i and group k. This expected membership can be computed as the posterior probability shown in Equation 5.3.

$$\begin{aligned}
r_{ik} &= P(Z_i = k|\boldsymbol{X}_{i\cdot}, \boldsymbol{Y}_{i\cdot}; \boldsymbol{\pi}, \boldsymbol{\beta}, \sigma^2) \\
&= \frac{P(Z_i = k; \boldsymbol{\pi}) \prod_{j=1}^{N_i} P(Y_{ij}|\boldsymbol{X}_{ij}, Z_i = k; \boldsymbol{\beta}, \sigma^2)}{\sum_{k'=1}^{K} P(Z_i = k'; \boldsymbol{\pi}) \prod_{j=1}^{N_i} P(Y_{ij}|\boldsymbol{X}_{ij}, Z_i = k'; \boldsymbol{\beta}, \sigma^2)}
\end{aligned} \tag{5.3}$$

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 λ to ensure that the constraint $\sum_{k=1}^{K} \pi_k = 1$ is satisfied.

$$\sum_{i=1}^{M} \frac{r_{ik}}{\pi_k} - \lambda = \sum_{i=1}^{M} r_{ik} - \lambda\pi_k = 0$$

Summing over all $k \in \{1, \cdots, K\}$, we find that $\lambda = \sum_i \sum_k r_{ik} = M$. Thus we plug λ into the equation above and get the updating equation for π_k in Equation 5.4.

$$\pi_k = \frac{1}{M} \sum_{i=1}^{M} r_{ik} \tag{5.4}$$

To estimate $\boldsymbol{\beta}_k$, we compute the gradient of $\boldsymbol{\beta}_k$ w.r.t. the expected complete log-likelihood $\mathcal{Q}$ in Equation 5.5. Notice that the gradient in Equation 5.5 has the same form as 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.

$$\frac{\partial \mathcal{Q}}{\partial \boldsymbol{\beta}_k} = \frac{1}{\sigma^2} \sum_{i=1}^{M} r_{ik} \sum_{j=1}^{N_i} (Y_{ij} - \boldsymbol{\beta}_k \boldsymbol{X}_{ij}) \boldsymbol{X}_{ij} \tag{5.5}$$

Finally, we compute the gradient of σ^2 w.r.t. the expected complete log-likelihood $\mathcal{Q}$ and the updating equation for the parameter σ^2 has the closed-form solution in Equation 5.6.

$$\sigma^2 = \frac{\sum_{i=1}^{M} \sum_{k=1}^{K} r_{ik} \sum_{j=1}^{N_i} (Y_{ij} - \boldsymbol{\beta}_k \boldsymbol{X}_{ij})^2}{\sum_{i=1}^{M} N_i} \tag{5.6}$$

Since the EM algorithm may converge to a local maximum of the expected complete log-likelihood function, depending on initialization of the parameters, we use random restart by assigning each birder to a group randomly. The expected membership $r_{i\cdot}$ specifies a soft clustering of birder i. To get the partition of birders in the training data, we assign each birder to the group with largest expected membership.

5.2.3 Determining the number of groups

To determine the number of groups in the data, we start the mixture model with only one group ($K = 1$) and gradually increase the value of K until it does not improve the average log-likelihood on a holdout set. The average log-likelihood is defined in Equation 5.7. Unlike the log-likelihood function in Equation 5.1, we compute the data likelihood of a birder by averaging the observation probability $P(Y_{ij}|\boldsymbol{X}_{ij}, Z_i = k; \boldsymbol{\beta}, \sigma^2)$ over all the observations from that birder.

$$\sum_{i=1}^{M} \log \left(\sum_{k=1}^{K} P(Z_i = k; \boldsymbol{\pi}) \frac{1}{N_i} \sum_{j=1}^{N_i} P(Y_{ij}|\boldsymbol{X}_{ij}, Z_i = k; \boldsymbol{\beta}, \sigma^2) \right) \tag{5.7}$$

5.2.4 Inference

After we learn the mixture of SACs model, we can infer a new birder's group membership based on their previous observations. Given the mixture model learned from training data, a new birder's membership can be computed as in Equation 5.8. Let $\boldsymbol{Y}'$ and $\boldsymbol{X}'$ denote the species count and effort duration in the new birder's observations and Z' denote the latent group membership of the new birder.

$$\underset{k}{\operatorname{argmax}}\, P(Z' = k|\boldsymbol{Y}', \boldsymbol{X}'; \boldsymbol{\pi}, \boldsymbol{\beta}, \sigma^2) = \underset{k}{\operatorname{argmax}}\, P(Z' = k; \boldsymbol{\pi}) \prod_{j=1}^{N_i} P(Y'_{ij}|\boldsymbol{X}'_{ij}, Z' = k; \boldsymbol{\beta}, \sigma^2) \tag{5.8}$$

5.3 Evaluation and Discussion

In our study, we evaluate the mixture of SACs model using eBird Reference Data [60] in four species-rich states that have high levels of year-round eBird participation (New York, Florida, Texas, and California). First, we show the effectiveness of SACs in characterizing the differences of eBird users with different levels of participation and capturing the evolution of birders' skills over time. Then, we test whether the mixture model can cluster eBird participants into groups based on their skill levels. However, evaluating the clustering is challenging due to the lack of the ground truth on their skills. Given the large number of birders, we can not validate the clusters by manually verifying each birder's submissions. Instead, we propose to validate the clusters based on an individual birder's ability to identify hard-to-detect species and use anecdotal information from the eBird project leaders. We also run the same analyses on eBird hotspots where the number of observers is relatively small, allowing us to manually verify their skills and validate the clustering results.

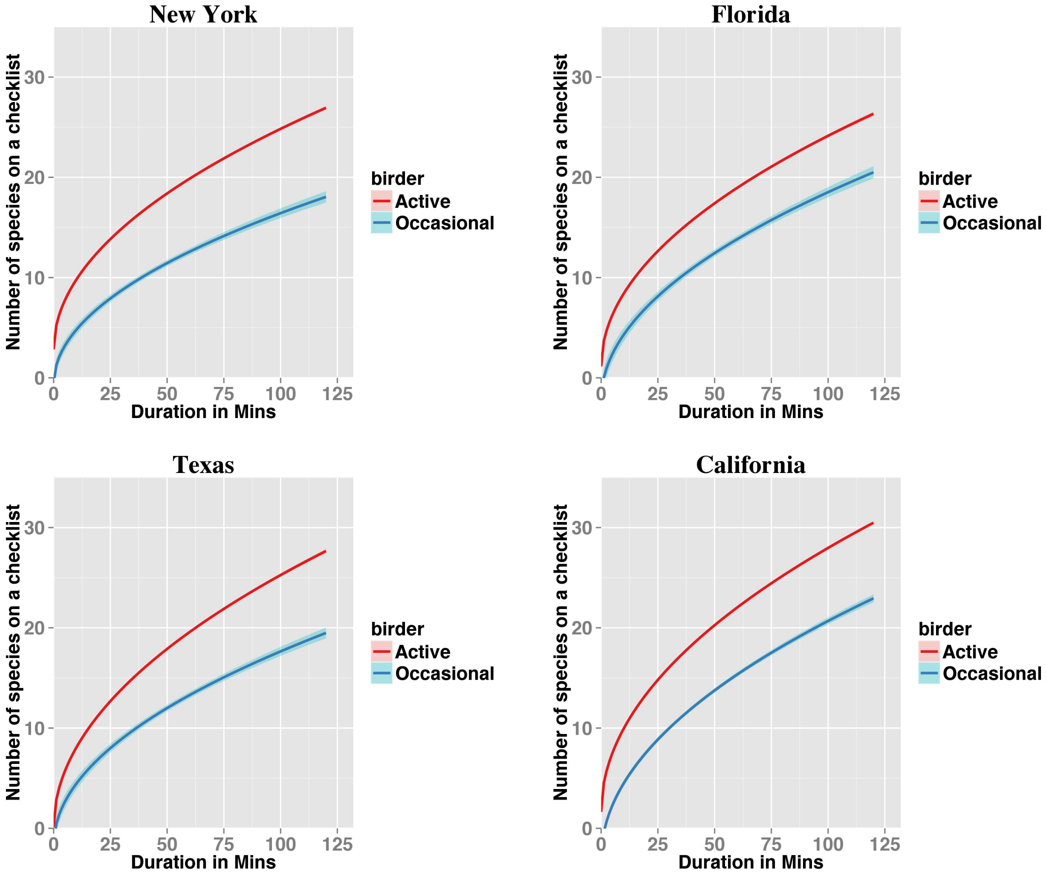


Figure 5.2: The species accumulation curves of the active and occasional eBird users in four states. The shaded area of a curve shows the 95% confidence interval.

5.3.1 Species Accumulation Curves

To show the SACs can be used to characterize birders' skill levels, we split all eBird participants into the active users and the occasional users, and show the differences of their SACs. In eBird, the top 10% of eBird participants contribute about 90% of eBird observations. These top 10% birders have submitted at least 60 checklists to eBird up till 2013. Thus active users are defined to be those eBird participants who submitted 60 or more checklists and the other eBird participants are occasional users. Intuitively, active users are more involved in eBird and should have better skill levels than the occasional users. To account for the spatial sampling bias, we use data only from sites where both active and occasional users have submitted at one checklist. We also limit our analysis to include checklists with duration less than 2 hours. Then we fit a SAC to the data from each group and present the curves in Figure 5.2. Across all four states, the active users have a much steeper SAC than the occasional users, indicating the active users are able to identify more species per unit time than the occasional users.

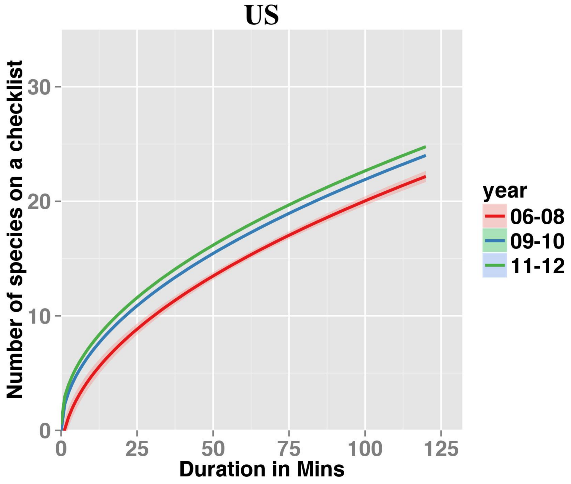


Figure 5.3: The species accumulation curves showing the evolution of eBird participants' skills over time. The shaded area of a curve shows the 95% confidence interval.

We also show that SACs can capture the evolution of birders' skill levels over time. In this analysis, we focus on a subset of eBird participants who started participating in eBird as occasional users and turned into active users as they became more engaged in eBird. First we split eBird data from the year 2006 to 2012 into three stages (06-08, 09-10 and 11-12) and identify active users within each stage as those top birders who contributed 90% of the data in that stage. Then the subset of eBird participants of interest are those who started as occasional users in stage 1 and turned into active users in stage 3, and their number of submissions kept increasing through these three stages. In this analysis, we identify this set of birders using the eBird data of the entire US and then fit a SAC of their submissions for each stage shown in Figure 5.3. As eBird participants became more active and involved in birding, their skill levels also evolved over time (their SACs became steeper). The improvement of their skills from stage 2 to stage 3 is less obvious compared to the improvement from stage 1 to stage 2 because identifying hard-to-detect species requires much more learning and experience.

5.3.2 Grouping eBird participants

We evaluate the mixture model in four different states using the eBird data in 2012. First, we remove the birders who submitted fewer than 20 checklists 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 number of distinct groups in the data, we split the data into training and validation sets. We train the mixture model on the training set with different values of $K \in \{1, \cdots, 6\}$, and then we calculate the average log-likelihood on the validation set. The best value of K is chosen when increasing K does not improve the average log-likelihood. In Figure 5.4, we show the average log-likelihood on the holdout data in four states. The graphs clearly show that there are 3 distinct groups in all four states.

Given the value of K chosen above, we re-estimate the mixture model using all the data in 2012 and show the SACs of different groups in four states in Figure 5.5. We sort the SACs by their slope coefficient β_1 in decreasing order so that the first group corresponds to the most skilled observers and the last group corresponds to the least skilled observers. The red curve corresponding to the top group has a consistently higher SAC than the other two groups across all four states. Birders in this top group are able to detect around 40 unique species during a 2-hour birding trip, while birders in group 2 and group 3 can only detect around 30 and 20 species. The 95% confidence intervals of the curves indicate that they are significantly different from each other across

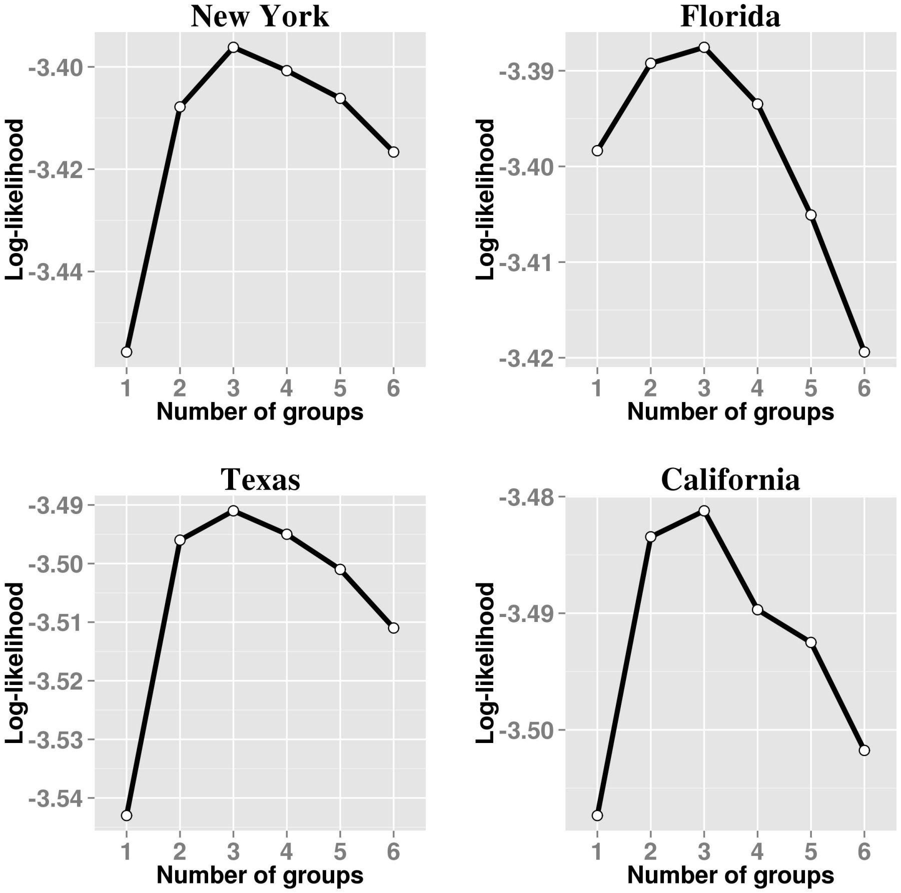


Figure 5.4: The average log-likelihood on the holdout data for different values of K in four states. The highest number indicates the best number of distinct groups found in that state.

all four states. Though the number of distinct groups are the same in all four states, the proportions of groups are very different. In New York and California, there are 7% and 12% participants falling into the top group as they are able to detect more species per unit of time. In Florida and Texas, the size of the top group is bigger, with 19% and 18% observers respectively. One explanation is that in New York, a small group of observers from the Cornell Lab of Ornithology are extremely skilled at identifying bird species; their skill levels distinguish them from the rest of the eBird participants in New York (the validation of these top group birders verified our hypothesis in Section 5.3.3).

State	Number of Birders (percent)			Averaged checklists per birder		
	G1	G2	G3	G1	G2	G3
New York	30 (7%)	155 (37%)	236 (56%)	407	215	152
Florida	63 (19%)	144 (45%)	117 (36%)	200	125	124
Texas	79 (18%)	196 (44%)	169 (38%)	132	157	99
California	91 (12%)	298 (40%)	352 (48%)	236	195	111

Table 5.2: The number of birders and the average number of checklists submitted per birder of each group in four states.

In Table 5.2, we report the number and proportion of observers and the average number of checklists submitted per birder in each group. The observers in the more skilled groups submit more checklists than observers in the less skilled groups. This matches our intuition that observers who are more active and involved in eBird project tend to be more skilled in detecting bird species. To demonstrate the differences in birding skills of birders across groups, we randomly choose two birders from each group in New York and show their SACs in Figure 5.6. Birders in the top group are able to accumulate species much faster especially in the first 30-45 minutes and can detect more than 30 species (sometimes more than 50 species) in the first 60 minutes, while birders in group 2 and group 3 are less skilled, reporting around 20 and 15 species respectively. Birders in group 3 can hardly detect more than 30 species in the first 120 minutes due to their limited experience and skills in bird watching.

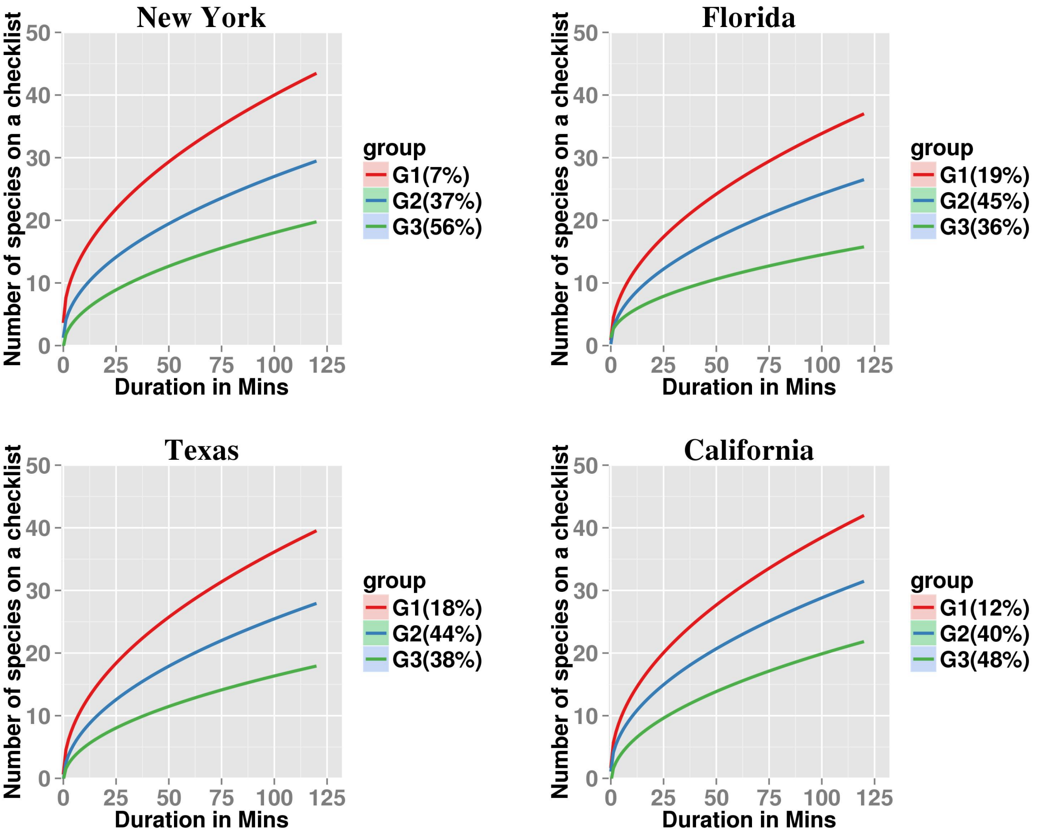
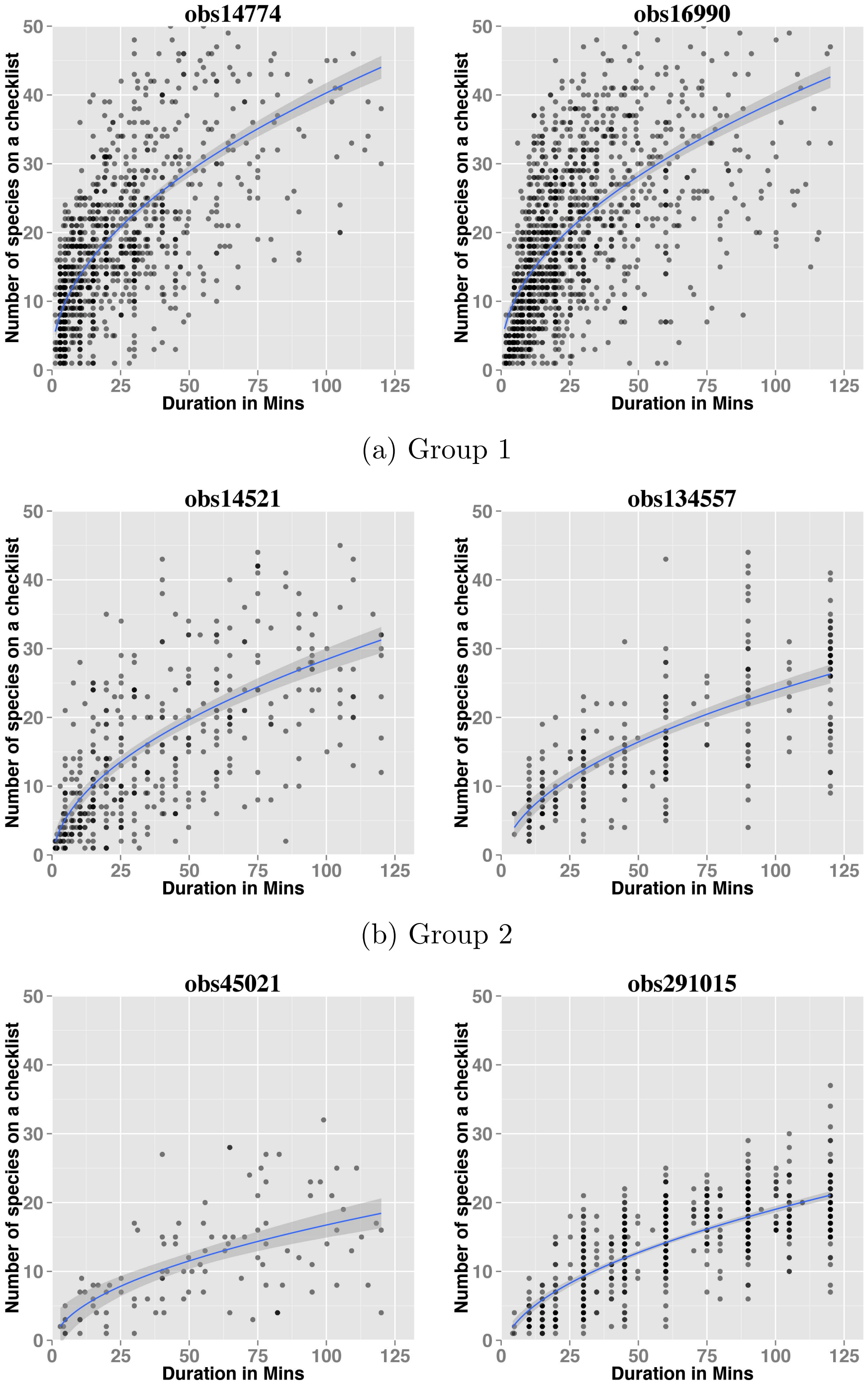


Figure 5.5: The species accumulation curves learned from the mixture of SACs model in four states. The number in the parenthesis indicates the proportion of birders in each group. The shaded area of a curve shows the 95% confidence interval.

5.3.3 Detection of hard-to-detect bird species

A good partition of birders leads to distinct differences in the skill levels of different groups. Since we do not have ground truth on birders' skills, we characterize their skill levels in terms of their ability to detect hard-to-detect bird species. Hard-to-detect species often require more experience and skills to be identified, e.g. some species can be detected by sound rather than by sight and some species can be detected only if observers know their habitats. In our experiment, we use 8 hard-to-detect species in each state suggested by experts at the Cornell Lab of Ornithology and calculate the



(a) Group 1

(b) Group 2

(c) Group 3

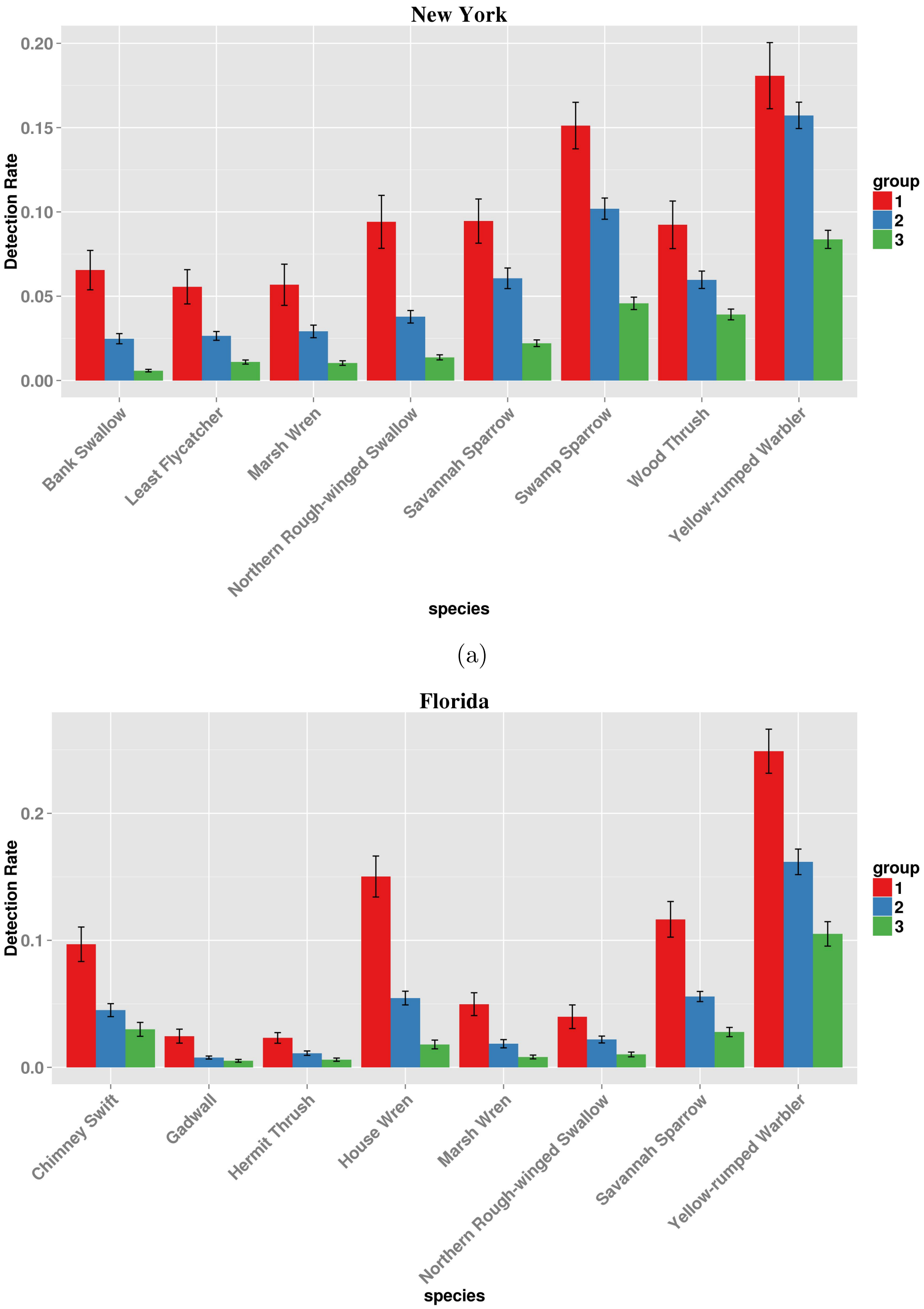
Figure 5.6: The species accumulation curves and the scatter plots of eBird participants from three groups in New York. The shaded area of a curve shows the 95% confidence interval. Each point represents a checklist submitted by the birder. The color of a point in the scatterplot specifies the number of checklists overlapped in the location. A darker color indicates more checklists overlapped at that point.

average detection rate of observers within each group. An observer's detection rate of a species is defined to be the percent of one's checklists that report the detection of that species. In Figure 5.7 and 5.8, we show the average detection rate of the hard-to-detect species in each group. The top group has the highest detection rate across all species in all four states, 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 statistically significant differences 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 New York while the other 5 observers are known to be reputable birders submitting high quality checklists to eBird. Thus, the mixture model is able to identify a group of top eBird contributors that are highly skilled birders and distinguish their behavior from the other groups of eBird participants.

5.3.4 eBird Hotspots

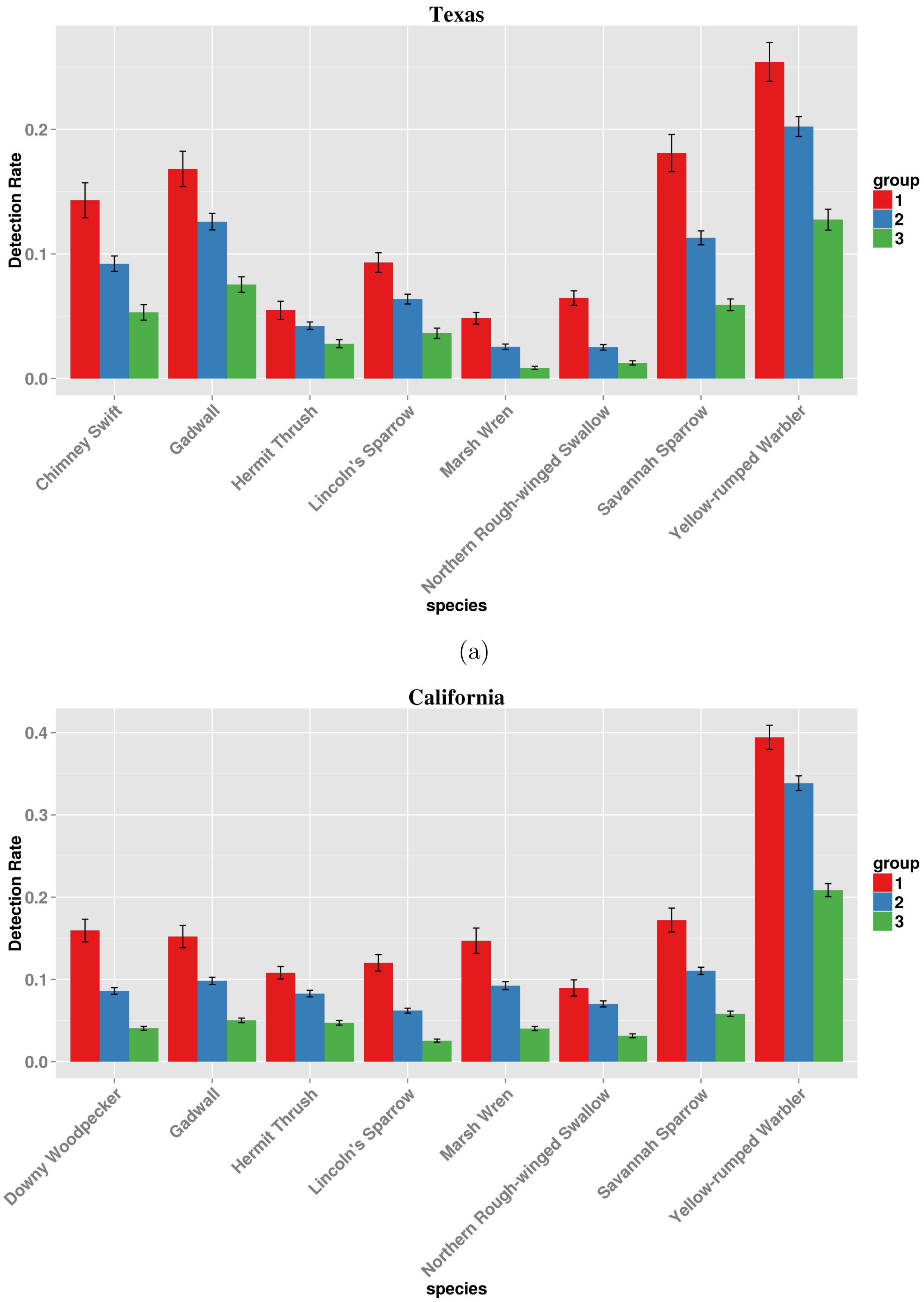
Since validating the clusters of birders at each location in a state is not viable, we run the same analyses on two eBird hotspots (*Stewart Park* and *Hammond Hill* in New York), where the number of observers allows us to manually verify the partition of birders. The eBird hotspots are public birding locations that are often heavily visited all year around. After training the mixture model using data submitted in those two hotspots, the model discovers 2 groups in Stewart Park and only 1 group in Hammond Hill. The SACs of these two eBird hotspots are shown in Figure 5.9. In Stewart Park, there are 25 birders submitting at least 10 checklists in 2012 and about half of the birders (13 birders) are classified into group 1. After manually verifying their identities and previous submissions, all 13 birders in group 1 have been verified to be expert birders and 10 out of the other 12 birders have been verified to be novice birders. There are two skilled birders being classified into group 2 because most of their submissions are short-time observations, making the curve fitting of their observations less accurate. In Hammond Hill, there are only 10 birders submitting at least 10 checklists in 2012 and all of them



(a)

(b)

Figure 5.7: (a): The average detection rate of three groups on 8 hard-to-detect species in New York. (b): the average detection rate of three groups on 8 hard-to-detect species in Florida. The error bars represent the standard error of detection rate within a group.



(a)

(b)

Figure 5.8: (a): the average detection rate of three groups on 8 hard-to-detect species in Texas. (b): the average detection rate of three groups on 8 hard-to-detect species in California. The error bars represent the standard error of detection rate within a group.

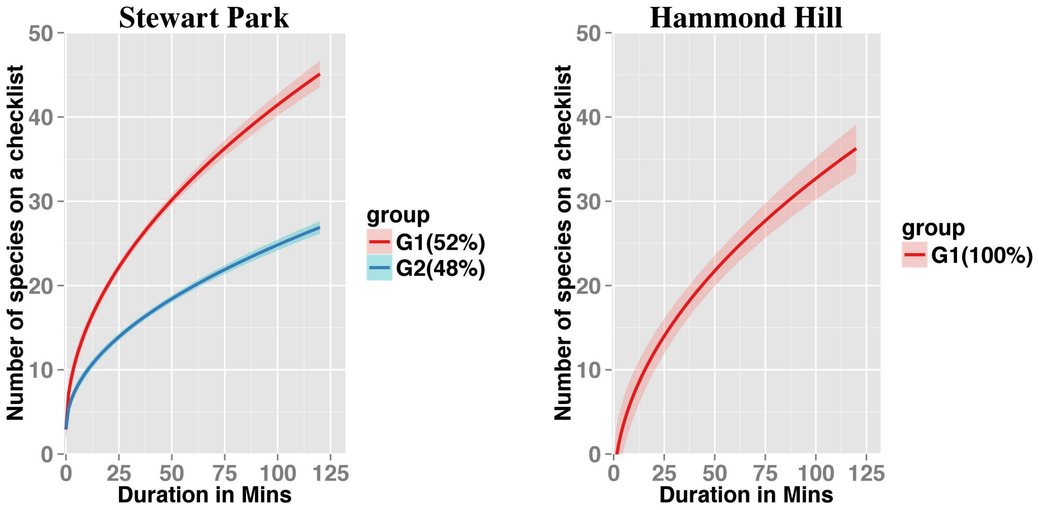


Figure 5.9: The species accumulation curves in eBird hotspots *Stewart Park* and *Hammond Hill*. The number in the parenthesis indicates the proportion of birders in that group. The shaded area of a curve shows the 95% confidence interval.

are verified to be expert birders. Thus, the mixture model is able to find the correct number of groups and cluster birders with similar skill levels into the same group.

5.4 Conclusion

Identifying groups of citizen scientists with similar skill levels is crucial in large-scale citizen science projects. Clustering citizen scientists by their skill levels allows better use of citizen science data, e.g. understanding different birding behaviors, developing automated data filters, and building species distribution models. In this work, we proposed to characterize an observer's skill based on species accumulation curves and developed a mixture of SACs model that was successful at identifying distinct groups of citizen scientists with similar skill levels in the eBird project. In addition, the clusters discovered from New York data do in fact correspond to groups that vary in their ability to observe hard-to-detect bird species.

Chapter 6: Modeling Misidentification of Bird Species by Citizen Scientists

6.1 Introduction

Species distribution models (SDMs) estimate the pattern of species occurrence on a landscape based on environmental features associated with each site. SDMs play an important role in predicting biodiversity and designing wildlife reserves [51, 62]. Learning accurate SDMs over a broad spatial and temporal scale requires large amounts of observational data to be collected. This scale of data collection is viable through *citizen science*, in which volunteers from the general public are encouraged to contribute data to scientific studies [11].

Although citizen scientists can contribute large quantities of data, data quality is a common source of concern with large-scale citizen science projects like eBird. In eBird, individuals vary greatly in their ability to identify organisms by species. Inexperienced observers either overlook or misidentify certain species and thus add noise to the data. For example, novice birders often confuse house finches with purple finches, which are similar in appearance. However, expert birders can distinguish between the two species, largely by where they are observed as house finches are often found in urban settings while purple finches are often found in forests. One way to reduce noise is to identify the invalid observations in the data verification process [85] and remove them from the eBird database. A more proactive way to improve data quality is to enhance the species identification skills of inexperienced observers and to help them correctly identify species that are commonly mistaken for each other. To accomplish this goal, we need to first discover groups of misidentified species from eBird data.

To discover groups of misidentified species, we extend the well known latent variable model in ecology, the *Occupancy-Detection* (OD) model [55], to the multiple species case. The OD model separates occupancy from detection and was developed under the assumption that data were collected by expert field biologists. As such, As such, the OD model assumes that there are no false positives in the data, since experts will not typically misidentify the species. The OD model does account for false negatives, which are common in species data since many species are secretive and hard to detect on surveys. Since citizen science data is collected less rigorously, the assumption of no false positives is questionable. For example in eBird, false positives arise when novice observers mistake one species for another. Previous work has incorporated the possibility of false positives into the OD model [71], and more recent work has adapted this to the citizen science context by distinguishing between experts and novices in the detection process [87].

The OD model and its variants are typically constructed for individual species, although some work has begun to address the co-occurrence patterns of pairs of species [56, 82]. In this work, we introduce the *Multi-Species Occupancy-Detection* (MSOD) model, which models the occurrence pattern of multiple species simultaneously and treats false positives as arising from the presence of another species with which the reported species is confused. This contrasts with previous work which treated species independently instead of linking false positives to the presence of an alternate species. We model the detection process in the MSOD model using the noisy-or parameterization, inspired by the QMR-DT network for medical diagnosis [77, 38, 43]. To discover species confusions, we propose an algorithm to learn both the model structure (i.e. species confusions) and the parameters of the SDMs from observational data.

Modeling occupancy and detection patterns for multiple species jointly has two important advantages. Firstly, discovering the patterns of confusion between species is useful in improving inexperienced observers' skills and eventually leading to better quality data in the eBird Human/Computer Learning Network [46]. Secondly, explicitly modeling the confusions and detection errors between species can improve the estimates of their occupancy patterns. Since the latent occupancy model is the true species distribution model of interest in this case, improvements in our ability to remove the nuisance detection process from the data allow more accurate ecological conclusions to be drawn.

In our study, we show that explicitly modeling the confusions and detection errors between species not only helps discover groups of confusing species, but also improves the estimates of the occupancy patterns of those species using synthetic data and real world eBird data.

Table 6.1: Notation in the Multi-Species Occupancy-Detection model.

Symbol	Description
N	Number of sites.
T_i	Number of visits at site i.
$\boldsymbol{X}_i$	Occupancy features at site i.
Z_{is}	Occupancy status (unobserved) of species s at site i.
$\boldsymbol{Z}_{i\cdot}$	Occupancy status (unobserved) of all the species at site i.
Y_{its}	Observed presence/absence of species s at site i, visit t.
$\boldsymbol{Y}_{it\cdot}$	Observed presence/absence of all the species at site i, visit t.
o_{is}	Occupancy probability of species s at site i.
d_{itrs}	Detection probability of species s at site i, visit t due to the presence of species r.
$\boldsymbol{\alpha}_s$	Occupancy parameters of species s.
$\boldsymbol{\beta}_{rs}$	Detection parameters of species s when species r is present.
λ_{o_s}	Occupancy regularization term of species s.
$\lambda_{d_{rs}}$	Detection regularization term of species s when species r is present.
$\lambda_{s_{rs}}$	Structural regularization term of species s when species r is present.

6.2 Methodology

In this section, we first show the Multi-Species Occupancy-Detection model in graphical model representation and illustrate the parameterization of its detection process. Then we present a learning algorithm to estimate the model structure and parameters in the MSOD model. Finally we show how to make inference for the site occupancy (Z) and the observations on the checklist (Y).

6.2.1 The Multi-Species Occupancy-Detection model

The *Multi-Species Occupancy-Detection* (MSOD) model is a bipartite Bayesian network consisting of observed and latent binary variables for every species as shown using plate notation in Figure 6.1. The outer plate represents N sites. The variable $\boldsymbol{X}_i$ denotes a vector of features that influence the occupancy pattern for the species (e.g. land cover type) and Z_{is} denotes the true occupancy status of species s at site i. The occupancy

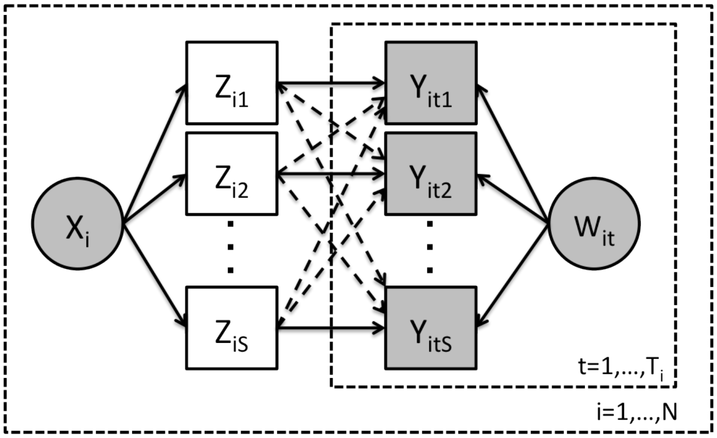


Figure 6.1: Graphical model representation of the Multi-Species Occupancy-Detection model.

patterns of different species depend on different habitats. Site i is surveyed T_i times. The variable $\boldsymbol{W}_{it}$ is a vector of features that affect the detectability of the species (e.g. time of day) and $Y_{its} \in \{0, 1\}$ indicates whether the species s was detected ($Y_{its} = 1$) on visit t at site i. A summary of the random variables used in the MSOD model is given in Table 6.1.

Structurally, the solid arrows in the plate diagram are fixed and known in advance; the dotted arrows are candidates to be added by the learning algorithm. In particular, we encode the fact that the presence of a species always causes the detection of that species by fixing the straight arrows in the model. However, other prior information can be easily encoded by adding or removing arrows in the model. When the structure of the MSOD model is fixed, the joint probability for the MSOD model is given in Equation 6.1.

$$\begin{aligned} P(\boldsymbol{Y}, \boldsymbol{Z} | \boldsymbol{X}, \boldsymbol{W}) &= \prod_{i=1}^{N} P(\boldsymbol{Y}_{i\cdot\cdot}, \boldsymbol{Z}_{i\cdot} | \boldsymbol{X}_i, \boldsymbol{W}_{i\cdot}) \\ &= \prod_{i=1}^{N} \prod_{s=1}^{S} P(Z_{is} | \boldsymbol{X}_i) \prod_{t=1}^{T_i} \prod_{s'=1}^{S} P(Y_{its'} | \boldsymbol{Z}_{i\cdot}, \boldsymbol{W}_{it}) \\ &= \prod_{i=1}^{N} \prod_{s=1}^{S} \left[P(Z_{is} | \boldsymbol{X}_i) \prod_{t=1}^{T_i} P(Y_{its} | \boldsymbol{Z}_{i\boldsymbol{\pi}(Y_{its})}, \boldsymbol{W}_{it}) \right] \end{aligned} \tag{6.1}$$

6.2.2 Parameterization

In the MSOD model, the species-specific occupancy model $P(Z_{is}|\boldsymbol{X}_i)$ of species s is parameterized as in the standard OD model. For each site i and each species s, we compute the probability o_{is} that site i is occupied by species s as $o_{is} = \sigma(\boldsymbol{X}_i \cdot \boldsymbol{\alpha}_s)$, where σ is the logistic function. Then the true occupancy Z_{is} of species s is generated by drawing from a Bernoulli distribution with parameter o_{is}. More specifically, the occupancy model can be written as follows.

$$\begin{aligned} o_{is} &= \sigma(\boldsymbol{X}_i \cdot \boldsymbol{\alpha}_s) \\ P(Z_{is}|\boldsymbol{X}_i; \boldsymbol{\alpha}_s) &= {o_{is}}^{Z_{is}}(1 - o_{is})^{1-Z_{is}} \end{aligned} \tag{6.2}$$

The detection probabilities ($P(Y_{its}|\boldsymbol{Z}_{i\cdot}, \boldsymbol{W}_{it})$ for each species s) depend on the occupancy status of species s (Z_{is}) and the occupancy status of other species s' that may be confused for species s. We model the detection process based on the noisy-or parameterization of the QMR-DT network for medical diagnosis [38, 77, 43]. The QMR-DT and MSOD models both consist of a set of latent causal variables (diseases and true species occupancies, respectively) and observed evidence variables (symptoms and observations, respectively). The key differences from the QMR-DT network are that the MSOD model has the same number of latent and observed variables and that the MSOD model needs to learn the partially unknown structure from data, whereas the QMR-DT network is constructed by experts from some number of diseases to a larger number of potential symptoms and the main task is to efficiently make inferences for the potential diseases given the symptoms.

More specifically, let d_{itrs} be the probability that at site i on visit t, species s is reported because species r is present. That is, $d_{itrs} = P(Y_{its} = 1|Z_{ir} = 1) = \sigma(\boldsymbol{W}_{it} \cdot \boldsymbol{\beta}_{rs})$. Therefore, true detection of a species s (straight arrows) denotes the detection due to the presence of itself, and false detection of a species (cross arrows) denotes the detection due to the presence of other species confused for s. Let $\boldsymbol{\gamma}$ be the adjacency matrix of $\{0, 1\}$ that represents the graph structure between the occupancy variable Z and the observation variable Y. $\boldsymbol{\gamma}_{rs} = 1$ if species r can be confused for species s (i.e. there exists an arrow from Z_{ir} to Y_{its}) and 0 otherwise. Additionally, we allow the leak probability d_{0s} of species s to be the probability of an observation when the occupancy of its parent nodes are all false. Due to the independence assumption in the noisy-or model, the probability of species s not being reported during visit t at site i ($P(Y_{its} = 0|\boldsymbol{Z}_{i\cdot}, \boldsymbol{W}_{it})$) can be fully factorized as in Equation 6.3. Thus, the probability of species s being reported is given in Equation 6.4.

$$P(Y_{its} = 0|\boldsymbol{Z}_{i\cdot}, \boldsymbol{W}_{it}) = (1 - d_{0s}) \prod_{r=1}^{S} (1 - d_{itrs})^{\gamma_{rs} Z_{ir}} \tag{6.3}$$

$$P(Y_{its} = 1|\boldsymbol{Z}_{i\cdot}, \boldsymbol{W}_{it}) = 1 - (1 - d_{0s}) \prod_{r=1}^{S} (1 - d_{itrs})^{\gamma_{rs} Z_{ir}} \tag{6.4}$$

6.2.3 Structure learning and parameter estimation

During training, we learn both the graph structure $\boldsymbol{\gamma}$ and the occupancy and detection parameters ($\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$). Given the unique bipartite graph structure of the MSOD model, we propose a structure learning approach using linear relaxation. We relax the constraint that the adjacency matrix $\boldsymbol{\gamma}_{rs} \in \{0, 1\}$ to $\boldsymbol{\gamma}_{rs} \in [0, 1]$, turning the integer program into a linear program. With this linear relaxation, we then estimate the MSOD model parameters using Expectation Maximization [14]. In the E-step, EM computes the expected occupancies $\boldsymbol{Z}_{i\cdot}$ for each site using Bayes rule. In the M-step, EM re-estimates the value of parameters $\{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}\}$ that maximize the expected log-likelihood in Equation 6.5.

$$\begin{aligned}
\mathcal{Q}(\Theta) &= \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{Y},\boldsymbol{X},\boldsymbol{W}} \Big[\log(P(\boldsymbol{Y}, \boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{W})) \Big] \\
&= \sum_{i}^{N} \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{Y},\boldsymbol{X},\boldsymbol{W}} \Big[\log \Big(P(Z_{i\cdot}|\boldsymbol{X}_i; \boldsymbol{\alpha}) \prod_{t}^{T_i} P(Y_{it\cdot}|Z_{i\cdot}, \boldsymbol{W}_{it}; \boldsymbol{\beta}) \Big) \Big] \\
&= \sum_{i}^{N} \sum_{\boldsymbol{Z}_{i\cdot}} \tilde{P}(\boldsymbol{Z}_{i\cdot}) \Big[\sum_{s}^{S} \log P(Z_{is}|\boldsymbol{X}_i; \boldsymbol{\alpha}_s) + \sum_{t}^{T_i} \log P(Y_{its}|Z_{i\cdot}, \boldsymbol{W}_{it}; \boldsymbol{\beta}_s) \Big] \\
&= \sum_{i}^{N} \sum_{\boldsymbol{Z}_{i\cdot}} \tilde{P}(\boldsymbol{Z}_{i\cdot}) \Big[\sum_{s}^{S} Z_{is} \log o_{is} + (1 - Z_{is}) \log(1 - o_{is}) + \\
&\qquad \sum_{t}^{T_i} Y_{its} \log \Big(1 - (1 - d_{0s}) \prod_{r}^{S} (1 - d_{itrs})^{\gamma_{rs} Z_{ir}} \Big) + \\
&\qquad (1 - Y_{its}) \log \Big((1 - d_{0s}) \prod_{r}^{S} (1 - d_{itrs})^{\gamma_{rs} Z_{ir}} \Big) \Big]
\end{aligned} \tag{6.5}$$

The expected occupancy at site i, $\tilde{P}(\boldsymbol{Z}_{i\cdot}) = P(\boldsymbol{Z}_{i\cdot}|\boldsymbol{X}_i, \boldsymbol{Y}_{i\cdot\cdot}, \boldsymbol{W}_{i\cdot})$, updated in the E-step can be computed as the posterior probability in Equation 6.6.

$$P(\boldsymbol{Z}_{i\cdot} = \boldsymbol{z}_{i\cdot}|\boldsymbol{X}_i, \boldsymbol{Y}_{i\cdot\cdot}, \boldsymbol{W}_{i\cdot}) = \frac{P(\boldsymbol{Y}_{i\cdot\cdot}, \boldsymbol{Z}_{i\cdot} = \boldsymbol{z}_{i\cdot}|\boldsymbol{X}_i, \boldsymbol{W}_{i\cdot})}{\sum_{\boldsymbol{z'}_{i\cdot}\in\{0,1\}^S} P(\boldsymbol{Y}_{i\cdot\cdot}, \boldsymbol{Z}_{i\cdot} = \boldsymbol{z'}_{i\cdot}|\boldsymbol{X}_i, \boldsymbol{W}_{i\cdot})} \tag{6.6}$$

In the M-step, EM determines the values of $\{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}\}$ that maximize the expected log-likelihood in Equation 6.5. Since there is no closed-form solution, we apply L-BFGS-B [8] to perform the optimization using the gradients in Equation 6.7, 6.8 and 6.9.

$$\begin{aligned}
\frac{\partial \mathcal{Q}}{\partial \boldsymbol{\alpha}_s} &= \sum_{i}^{N} \frac{\partial \mathcal{Q}_i}{\partial o_{is}} \frac{\partial o_{is}}{\partial \boldsymbol{\alpha}_s} \\
&= \sum_{i}^{N} \sum_{Z_{is}} \tilde{P}(Z_{is}) \Big(\frac{Z_{is}}{o_{is}} - \frac{1 - Z_{is}}{1 - o_{is}} \Big) o_{is}(1 - o_{is}) \boldsymbol{X}_i \\
&= \sum_{i}^{N} (\tilde{P}(Z_{is} = 1) - o_{is}) \boldsymbol{X}_i && (6.7) \\
\frac{\partial \mathcal{Q}}{\partial \boldsymbol{\beta}_{rs}} &= \sum_{i=1}^{M} \frac{\partial \mathcal{Q}_i}{\partial d_{itrs}} \frac{\partial d_{itrs}}{\partial \boldsymbol{\beta}_{rs}} \\
&= \sum_{i=1}^{M} \sum_{\boldsymbol{Z}_{i\cdot}} \tilde{P}(\boldsymbol{Z}_{i\cdot}) \sum_{t}^{T_i} \Bigg(\frac{Y_{its}}{1 - (1 - d_{0s}) \prod_k^S (1 - d_{itks})^{\gamma_{ks} Z_{ik}}} - 1 \Bigg) Z_{ir} \gamma_{rs} d_{itrs} \boldsymbol{W}_{it} \\
&= \sum_{i=1}^{M} \sum_{\boldsymbol{Z}_{i\cdot}} \tilde{P}(\boldsymbol{Z}_{i\cdot}) \sum_{t}^{T_i} \Bigg(\frac{Y_{its}}{P(Y_{its} = 1|\boldsymbol{Z}_{i\cdot}, \boldsymbol{W}_{it})} - 1 \Bigg) Z_{ir} \gamma_{rs} d_{itrs} \boldsymbol{W}_{it} && (6.8) \\
\frac{\partial \mathcal{Q}}{\partial \boldsymbol{\gamma}_{rs}} &= \sum_{i=1}^{M} \sum_{\boldsymbol{Z}_{i\cdot}} \tilde{P}(\boldsymbol{Z}_{i\cdot}) \sum_{t}^{T_i} \Bigg(1 - \frac{Y_{its}}{1 - (1 - d_{0s}) \prod_k^S (1 - d_{itks})^{\gamma_{ks} Z_{ik}}} \Bigg) log(1 - d_{itrs}) Z_{ir} \\
&= \sum_{i=1}^{M} \sum_{\boldsymbol{Z}_{i\cdot}} \tilde{P}(\boldsymbol{Z}_{i\cdot}) \sum_{t}^{T_i} \Bigg(1 - \frac{Y_{its}}{P(Y_{its} = 1|\boldsymbol{Z}_{i\cdot}, \boldsymbol{W}_{it})} \Bigg) log(1 - d_{itrs}) Z_{ir} && (6.9)
\end{aligned}$$

To avoid overfitting, we use L_2-regularization penalty on the occupancy and detection coefficients ($\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$). For the model structure parameter $\boldsymbol{\gamma}$, we apply L_1-regularization penalty to enforce sparsity in the model structure because there are often few species confusions among species. Since L_1-norm is not differentiable at value 0, we use a smooth function "$espL_1$" to approximate of the L_1-norm so that the non-differentiable optimization is transformed into a differentiable one [41]. More specifically, we have $|\gamma_{rs}| \approx \sqrt{\gamma_{rs}^2 + \epsilon}$ for a sufficiently small positive ϵ. The entire regularization terms are shown in Equation 6.10.

$$r(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{s=1}^{S} \left(\lambda_{o_s} \frac{1}{2} \sum_{i=2}^{|\boldsymbol{\alpha}_s|} \alpha_{si}^2 + \sum_{r=1}^{S} \left(\lambda_{d_{rs}} \frac{1}{2} \sum_{i=2}^{|\boldsymbol{\beta}_{rs}|} \beta_{rsi}^2 + \lambda_{s_{rs}} \sqrt{\gamma_{rs}^2 + \epsilon} \right) \right) \tag{6.10}$$

In the learned adjacency matrix, γ_{rs} specifies the probability of species r being confused for species s. We sort the entries in the learned adjacency matrix $\boldsymbol{\gamma}$ and then pick a threshold on $\boldsymbol{\gamma}$ by using a validation set. In particular, we greedily add cross edges (i.e. pairs of misidentified species) according to their probability of misidentification until the log-likelihood on the validation set does not improve. Once we determine the structure, we retrain the MSOD model with a fixed structure and estimate the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. When the structure is fixed, we can use the same learning approach described above except that $\boldsymbol{\gamma}_{rs}$ is now fixed to be either 0 or 1. In addition, we initialize the leak probability of each species in the MSOD model to the value of the leak probability learned by the ODLP model (the OD model with a learned leak probability). A flowchart in Figure 6.2 shows the process of learning the MSOD model. Exact computation of the expectations in 6.5 is computationally expensive with large value of S since it require summing over the configurations of S binary variables, resulting in 2^S terms. We will investigate speedups using variational approximations [43, 61, 68].

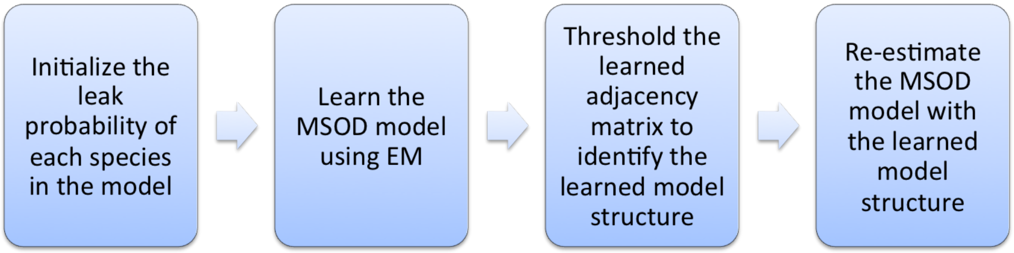

Initialize the leak probability of each species in the model
Learn the MSOD model using EM
Threshold the learned adjacency matrix to identify the learned model structure
Re-estimate the MSOD model with the learned model structure

Figure 6.2: The flowchart of learning the MSOD model.

An *identifiability* problem arises when estimating the MSOD model. This identifiability issue causes two symmetric but distinct sets of parameter values to be solutions to the EM procedure. For example, in a MSOD model of 2 species where both species can be confused for each other, we can convert one solution to an equivalent one by flipping the occupancies of both species and switching the true detection coefficients and the false detection coefficients within each species. While both of these solutions result in the same objective function in Equation 6.5, one solution yields a model that is more consistent with real world assumptions. To address this issue, we add a constraint to the objective function during training that biases EM towards the more desirable solution. This constraint encodes the fact that the detection probability of species s from the presence of itself is always higher than its detection probability from the presence of another species s' that is confused for species s (i.e. $\sum_{i,t} d_{itss} > \sum_{i,t} d_{its's}$).

6.2.4 Inference

The MSOD model can be used to predict the site occupancy of a specific species s (Z_{is}), or a set of species, and predict the observations of species s (Y_{its}) on a checklist. We describe the inference on these tasks in more detail below.

6.2.4.1 Prediction of site occupancy

We can use the MSOD model to compute the probability that the site is occupied by species s given the site environmental features and the observation history at that site. The occupancy probability of site i can be computed using Equation 6.11. In addition, let $\boldsymbol{Z}_{i\neg s}$ denote the occupancy variables of all species except for species s at site i.

$$\begin{aligned} P(Z_{is}=1|\boldsymbol{X}_i, \boldsymbol{Y}_{i\cdot\cdot}, \boldsymbol{W}_{i\cdot}) &= \frac{P(\boldsymbol{Y}_{i\cdot\cdot}, Z_{is}=1, \boldsymbol{Z}_{i\neg s}|\boldsymbol{X}_i, \boldsymbol{W}_{i\cdot})}{\sum_{z_{is}\in\{0,1\}} P(\boldsymbol{Y}_{i\cdot\cdot}, Z_{is}=z_{is}, \boldsymbol{Z}_{i\neg s}|\boldsymbol{X}_i, \boldsymbol{W}_{i\cdot})} \\ &= \frac{\sum_{\boldsymbol{z}_{i\neg s}} P(\boldsymbol{Y}_{i\cdot\cdot}, Z_{is}=1, \boldsymbol{Z}_{i\neg s}=\boldsymbol{z}_{i\neg s}|\boldsymbol{X}_i, \boldsymbol{W}_{i\cdot})}{\sum_{z_{is}\in\{0,1\}}\sum_{\boldsymbol{z}_{i\neg s}} P(\boldsymbol{Y}_{i\cdot\cdot}, Z_{is}=z_{is}, \boldsymbol{Z}_{i\neg s}=\boldsymbol{z}_{i\neg s}|\boldsymbol{X}_i, \boldsymbol{W}_{i\cdot})} \end{aligned} \tag{6.11}$$

where

$$P(\boldsymbol{Y}_{i\cdot\cdot}, Z_{is}=z_{is}, \boldsymbol{Z}_{i\neg s}=\boldsymbol{z}_{i\neg s}|\boldsymbol{X}_i, \boldsymbol{W}_{i\cdot}) = \prod_{r=1}^{S} P(Z_{ir}|\boldsymbol{X}_i;\boldsymbol{\alpha}_r)\prod_{t=1}^{T_i} P(Y_{its}|\boldsymbol{Z}_{i\cdot}, \boldsymbol{W}_{it};\boldsymbol{\beta}_{\cdot s})$$

Ecologists are sometimes interested in the co-occurrence of two or more species at a site in order to test some ecological hypotheses. We can use a similar formula as Equation 6.11 where we marginalize out the occupancy variable $\boldsymbol{Z}$ of species that we are not interested in and then calculate the posterior probability of co-occurrence for species of interest.

6.2.4.2 Predicting observations of species on a checklist

Since the true site occupancy is typically unavailable for evaluation in real-world field datasets, we often evaluate different SDMs based on the prediction of observation of a species at a site. Let $\boldsymbol{\pi}_s$ be the set of species that can be confused for species s in the MSOD model. To compute the probability of detecting species s at site i on visit t (Y_{its}), we marginalize out the occupancy variables of species in $\boldsymbol{\pi}_s$ as shown in Equation 6.12.

$$\begin{aligned} P(Y_{its} = 1|\boldsymbol{X}_i, \boldsymbol{W}_{it}) &= \sum_{\boldsymbol{z}_{i\boldsymbol{\pi}_s}} P(\boldsymbol{Y}_{its} = 1, \boldsymbol{Z}_{i\boldsymbol{\pi}_s} = \boldsymbol{z}_{i\boldsymbol{\pi}_s}|\boldsymbol{X}_i, \boldsymbol{W}_{it}) \\ &= \sum_{\boldsymbol{z}_{i\boldsymbol{\pi}_s}} P(Y_{its} = 1|\boldsymbol{Z}_{i\boldsymbol{\pi}_s} = \boldsymbol{z}_{i\boldsymbol{\pi}_s}, \boldsymbol{W}_{it}; \boldsymbol{\beta}_{\cdot s}) \prod_{k \in \boldsymbol{\pi}_s} P(Z_{ik} = z_{ik}|\boldsymbol{X}_i; \boldsymbol{\alpha}_k) \end{aligned} \tag{6.12}$$

6.3 Evaluation and Discussion

Evaluation of OD models and their variants is challenging because field data like eBird does not include the ground truth of site occupancy and we do not have access to the true model structure representing the “correct” species confusions. To evaluate the quality of the occupancy modeling component of the models, we use synthetic data and compare the learned model to the true model used to generate the data in predicting site occupancies and observations. Then on eBird data, we show the model structures learned for three case studies using sets of species known to be confused for each other and compare the performance of different models at predicting observations on a checklist.

6.3.1 Synthetic dataset

For the synthetic experiment, data is generated for 500 sites where the number of visits per site is randomly chosen from 1 to 3 with probability 50%/25%/25%. There are 4 occupancy covariates and 4 detection covariates drawn i.i.d from a standard normal distribution. A true structure over 5 species is generated by randomly adding 7 pairs of confusing species. Coefficients for the occupancy and detection models are also drawn i.i.d from standard normal distributions, and the leak probabilities for all species are set to be 0.01 as background noise. Furthermore, we constrain that the detection probability of a species s due to the presence of another species confused for s be smaller than the detection probability due to the presence of the species s itself. A training, validation and test dataset are generated following the generative MSOD model, and this entire process is repeated 30 times to generate 30 different datasets. This synthetic data is denoted by "Syn" in the result.

To test the robustness of the MSOD model, we inject different types of "noise" into the synthetic data and test the performance of the MSOD model against the "noise". First, we generate synthetic data with interactions between species occupancies, e.g. species competition and mutualism. In particular, we assume species 1 and 2 , and species 3 and 4 are pairs of competitors. The occupancy probability of species 2 at a site will be halved when species 1 occupies that site and same with species 3 and 4. Also, we assume species 3 and 5 have a mutualistic relationship and the occupancy probability of species 5 will increase by 20% (we truncate the occupancy probability at 1 when it goes beyond 1) at a site when species 3 occupies that site. We denote this synthetic data with occupancy interactions "Syn-I" in our discussion. Then, we generate synthetic data with non-linear occupancy covariates. More specifically, we generate the non-linear occupancy covariates ($X'_{i\cdot}$) from the original occupancy covariates ($X_{i\cdot}$) using the following transformations: $X'_{i1} = \sin(X_{i1} + 1)$, $X'_{i2} = \exp(X_{i2} - 1)$, $X'_{i3} = X_{i3} \cdot X_{i4}$, and $X'_{i4} = X_{i4}$. We denote this synthetic data "Syn-NL" in the discussion. In the last scenario, we make the synthetic data the most challenging by adding both species occupancy interactions and non-linear occupancy components ("Syn-I-NL") and test the performance of the MSOD model.

In our experiment, we compare the MSOD model against the standard OD model, a variant of the OD model called *ODLP*, which allows a learned leak probability in the OD model, and the *true* latent model in terms of predicting occupancy (Z) and observation (Y). To set the regularization terms of the occupancy (λ_o) and detection (λ_d) in the OD and ODLP models, we tune them over the set of values $\{0.01, 0.1, 1, 10\}$ based on the performance of the occupancy prediction. Instead of tuning the regularization terms of every species in the MSOD model separately, we set them to the best values found in the OD model of that species. In addition, we find that large regularization terms of false detections and model structure (λ_s) often lead to more accurate estimation of the structure in the MSOD model empirically, so we set both of them to be 10 in our experiment.

We report the area under the curve (AUC) and accuracy averaged over 30 datasets in Table 6.2 where both metrics are computed per species and averaged across species. On all four synthetic datasets, the standard OD model performs poorly because the *no false positives* assumption does not hold. The ODLP model improves slightly over the OD model because it allows false positives to be explained by the leak probability, but the leak probability itself can not accurately capture the noise from the detection process. The performance of the MSOD model is closest to the true model in predicting both occupancy and observation. Notice that the OD and MSOD model differ greatly in their prediction of occupancy even though their prediction of observations is fairly close. This indicates that the values of the latent occupancy variables are indeed very different from the values of the observation variables. As we allow species occupancy interactions and non-linear occupancy components in the data, the performance of the MSOD model decreases slightly and is still statistically better the OD and ODLP models. Furthermore, the MSOD model is more sensitive to the non-linear occupancy components in the data (about 3% decrease in terms of AUC in occupancy prediction) than the species occupancy interactions (1% decrease). In the most challenging case where both "noise" exist in the data, the performance of the MSOD model is still reasonably close to that of the true model.

To compare the learned model structure to the true model structure, we compute the *structural AUC*, which specifies the probability of ranking a true cross edge over an incorrect cross edge in the learned adjacency matrix. To calculate the structural AUC, we flatten the learned adjacency matrix and the true structure into two vectors and then calculate the AUC value from these two vectors. A structural AUC value of 1 indicates that the learning algorithm correctly ranks the true cross edge over the other cross edge

Table 6.2: The AUC and accuracy (and their standard errors) of occupancy and observation prediction averaged over 30 datasets in four synthetic experiments. The metrics are computed per species and averaged across species. Boldface results indicate the best performing model. $\star$ and $\dagger$ indicate the MSOD model is statistically better than the OD model and the ODLP model respectively.

(a) The synthetic dataset

Syn	**Occupancy ($\boldsymbol{Z}$)**		**Observation ($\boldsymbol{Y}$)**	
	AUC	Accuracy	AUC	Accuracy
TRUE	0.941 ± 0.004	0.881 ± 0.004	0.783 ± 0.004	0.756 ± 0.004
OD	0.849 ± 0.006	0.758 ± 0.006	0.751 ± 0.005	0.739 ± 0.004
ODLP	0.868 ± 0.006	0.780 ± 0.007	0.752 ± 0.005	0.741 ± 0.004
MSOD	**0.935 ± 0.005***†	**0.872 ± 0.006***†	**0.776 ± 0.004***†	**0.750 ± 0.004***†

(b) The synthetic dataset with species occupancy interactions

Syn-I	**Occupancy ($\boldsymbol{Z}$)**		**Observation ($\boldsymbol{Y}$)**	
	AUC	Accuracy	AUC	Accuracy
TRUE	0.943 ± 0.003	0.885 ± 0.004	0.776 ± 0.003	0.763 ± 0.005
OD	0.842 ± 0.005	0.731 ± 0.010	0.744 ± 0.004	0.746 ± 0.006
ODLP	0.865 ± 0.005	0.757 ± 0.010	0.746 ± 0.004	0.747 ± 0.006
MSOD	**0.925 ± 0.004***†	**0.862 ± 0.006***†	**0.763 ± 0.004***†	**0.755 ± 0.006***†

(c) The synthetic dataset with non-linear occupancy components

Syn-NL	**Occupancy ($\boldsymbol{Z}$)**		**Observation ($\boldsymbol{Y}$)**	
	AUC	Accuracy	AUC	Accuracy
TRUE	0.937 ± 0.003	0.878 ± 0.004	0.777 ± 0.005	0.762 ± 0.007
OD	0.837 ± 0.007	0.722 ± 0.010	0.739 ± 0.005	0.743 ± 0.007
ODLP	0.848 ± 0.007	0.734 ± 0.009	0.741 ± 0.005	0.744 ± 0.007
MSOD	**0.903 ± 0.006***†	**0.842 ± 0.007***†	**0.755 ± 0.004***†	**0.751 ± 0.007***†

(d) The synthetic dataset with species occupancy interactions and non-linear occupancy components

Syn-I-NL	**Occupancy ($\boldsymbol{Z}$)**		**Observation ($\boldsymbol{Y}$)**	
	AUC	Accuracy	AUC	Accuracy
TRUE	0.938 ± 0.003	0.878 ± 0.004	0.768 ± 0.003	0.759 ± 0.005
OD	0.832 ± 0.003	0.723 ± 0.011	0.731 ± 0.005	0.741 ± 0.006
ODLP	0.841 ± 0.006	0.735 ± 0.010	0.732 ± 0.004	0.742 ± 0.007
MSOD	**0.897 ± 0.006***†	**0.837 ± 0.008***†	**0.739 ± 0.005***†	**0.745 ± 0.007***†

in the model. In Table 6.3, we report the AUC for the learned model structure on four synthetic datasets. In the simplest case where there exists no "noise" in the data, the MSOD model archives the structural AUC value of 0.989. As we inject "noise" in the data, the structural AUC of the learned model structure only decreases slightly. In the most challenging case, the learning method can still achieve the structural AUC value of 0.970, indicating that the MSOD model almost always discovers the correct species confusions.

Syn	*Syn-I*	*Syn-NL*	*Syn-I-NL*
0.989 ± 0.012	0.980 ± 0.012	0.974 ± 0.010	0.970 ± 0.008

Table 6.3: The AUC (and its standard error) for the learned model structure of the MSOD model compared to the true model structure in four synthetic experiments. The AUC values are averaged over 30 datasets in each experiment.

6.3.2 eBird dataset

We also test the ability of the MSOD model to discover realistic species confusions on three case studies involving real-world eBird data, which was selected by consulting with experts at the Cornell Lab of Ornithology. We evaluated the MSOD model on subsets of eBird species that include some species known to be confused for each other and a distractor species with minimal similarity to the others. In the first case study, we consider the Sharp-shinned Hawk and Cooper's Hawk, and Turkey Vulture as the distractor species. In the second case study, we consider the Hairy Woodpecker and Downy Woodpecker, and Dark-eyed Junco as the distractor species. In the last case study, we consider the Purple Finch and House Finch, and Yellow-rumped Warbler as the distractor species.

In our experiment, we use eBird data from California in the year 2010 since eBird participation in California is high. We group the checklists within a radius of 0.16 km of each other into one site and each checklist corresponds to one visit at that grouped site. The radius is set to be small so that the site occupancy is constant across all the checklists associated with that grouped site. There are a total number of 3140 sites after grouping in California. For sites with more than 20 visits, we randomly sample 20 of them to include in the data. In our experiment, we use 19 occupancy features and 10 detection features shown in Table 6.4. For more details on the occupancy and detection covariates in the eBird data, we refer the readers to the eBird Manual [60].

To alleviate the effect of special autocorrelation in creating training and test data, we superimpose a checkerboard (each grid cell is roughly a 10 km by 10 km square) over the data in California. The checkerboard grids California into black and white cells. Data points falling into the white cells are grouped together as the test set. Then we further divide data in the black cells into 2-by-2 subgrids so that data falling into the top left and bottom right subgrids are grouped together as training set and data falling into the top right and bottom left are grouped together as validation set.

Table 6.4: Occupancy and detection features in eBird dataset used for evaluation.

Occupancy Features	Comments
Population	Population per square mile.
Housing density	Number of housing units per square mile.
Housing percent vacant	Percentage of housing units.
Elevation	Elevation in meters from National Elevation Dataset.
Habitat_X	Percent of surrounding landscape that is habitat class X. There are 15 habitat classes.

Detection Features	Comments
Time of day	Indicator variable of time (e.g. [0, 6), [6, 12), [12, 18), and [18, 24)).
Season	Indicator variable of season (e.g. Spring, Summer, Fall, and Winter).
Observation duration	Duration of observation for the checklist, in hours.
Route distance	Distance traveled during observation period, in kilometers.

6.3.2.1 Discovering species confusions

To learn the MSOD model on eBird data, we first estimate the leak probability of each spices by applying the ODLP model. Then we fix the leak probabilities of all species in the MSOD model and estimate the model structure and parameters described in Section 6.2.3. We show the learned model structures in Figure 6.3. The arrows specify the species confusions recovered by the MSOD model, e.g. Sharp-shinned Hawk and

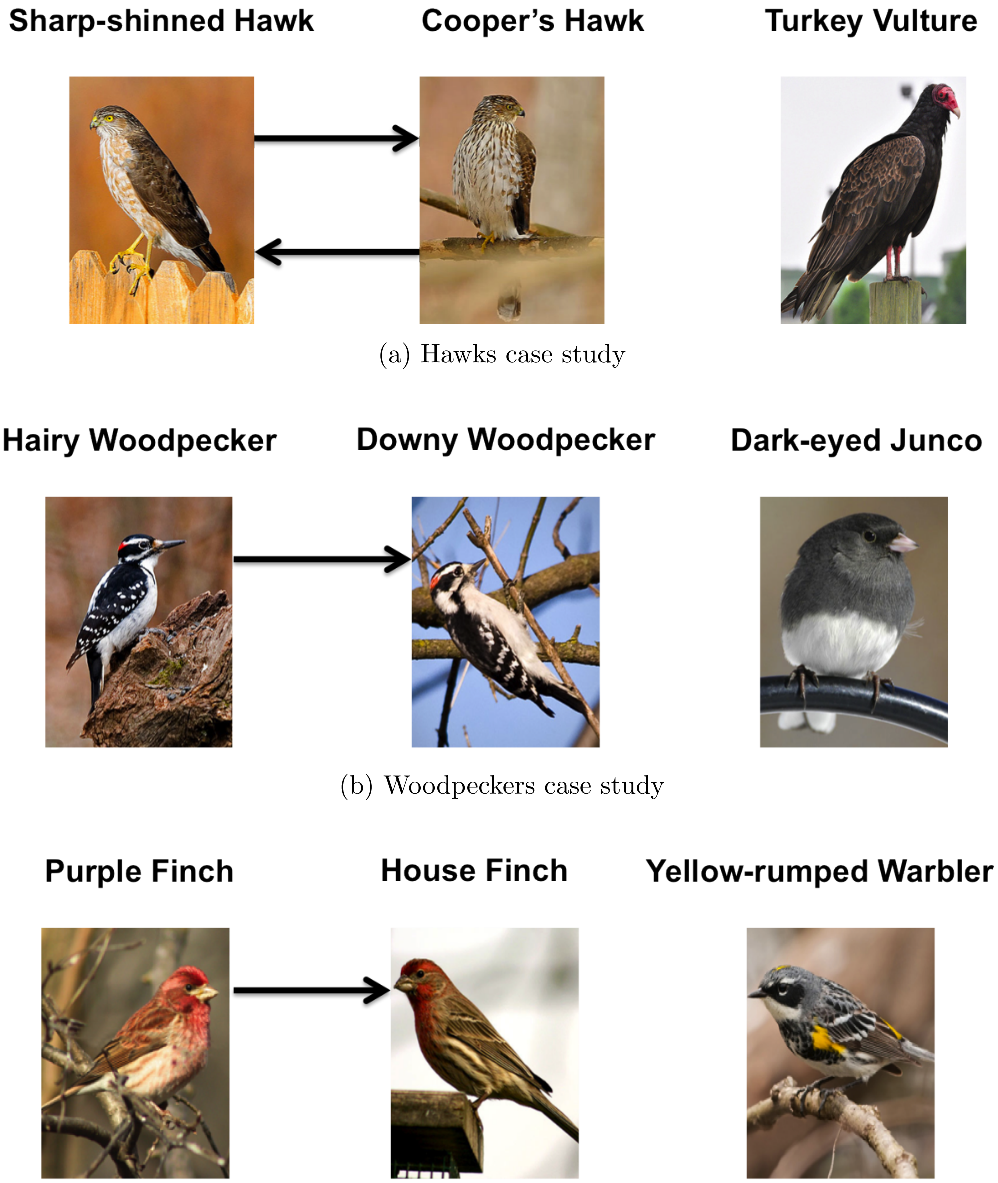
Cooper's Hawk are confused for each other, Hairy Woodpecker is likely to be confused for Downy Woodpecker, and Purple Finch is likely to be confused for House Finch. For all three cases, the structure recovered matches our expectations, and the confusion probability is higher on the arrow from the rarer species of the two to the more common one, indicating that inexperienced observers tend to misidentify the rarer species for the more common ones. Confusing rare species for the common ones often happens within entry-level observers, as they may not be aware of the rare species due to their lack of bird knowledge. Confusing the common species for the rare ones often happens within birders with certain birding skills as they are aware of the rare species, but lack the skills to distinguish them, thus resulting in an over-estimated distribution of the rare species.

6.3.2.2 Predicting the observations on a checklist

Since the species occupancies of a site is not available for evaluation, we use the prediction of observations on a checklist as a substitute. Given the learned structure in Section 6.3.2.1, we re-estimate the MSOD model using data in both training and validation set and predict the observations on checklists in the test set. To create different splits of the training and test sets, we randomize the checkerboarding by randomly positioning the bottom left corner to create 30 different datasets for evaluation. Then we compare the MSOD model with the standard OD model and the ODLP model as in the synthetic experiment. In Table 6.5, we report the AUC and accuracy of predicting observations for three case studies. In the MSOD model, all 6 species have statistically better AUC and accuracy compared to the OD model and 5 out of 6 species have statistically better AUC and accuracy compared to the ODLP model. The improvement of detection prediction of the MSOD model is minor as we expect given the results in the synthetic experiment.

6.4 Conclusion

We highlight two significant contributions of this work. Firstly, we introduce a novel multi-species occupancy-detection model. This is the first point of connection between the literature on ecological latent variable models and medical diagnosis with QMR-DT; we anticipate that further study of the similarities and differences between these models may yield more insight for one or both domains. Secondly, we show promising results



(a) Hawks case study

(b) Woodpeckers case study

(c) Finches case study

Figure 6.3: The arrows specify the species confusions recovered by the MSOD model. An arrow from species A to species B indicates that the presence of species A may result the detection of species B. We thank Chris Wood from the Cornell Lab of Ornithology for the images of each bird species.

Table 6.5: The AUC and accuracy of observation prediction for three eBird case studies. Boldface results indicate the winner, $\star$ and $\dagger$ indicate the MSOD model is statistically better than the OD and ODLP model respectively.

(a) The Hawks case study

	Sharp-shinned Hawk		**Cooper's Hawk**	
	AUC	Accuracy	AUC	Accuracy
OD	0.725 ± 0.005	0.967 ± 0.001	0.765 ± 0.003	0.912 ± 0.001
ODLP	0.737 ± 0.005	0.972 ± 0.001	0.770 ± 0.005	0.917 ± 0.002
MSOD	$\mathbf{0.757 \pm 0.003}^{\star\dagger}$	$\mathbf{0.976 \pm 0.001}^{\star\dagger}$	$\mathbf{0.780 \pm 0.002}^{\star\dagger}$	$\mathbf{0.923 \pm 0.001}^{\star}$

(b) The Woodpeckers case study

	Hairy Woodpecker		**Downy Woodpecker**	
	AUC	Accuracy	AUC	Accuracy
OD	0.833 ± 0.004	0.940 ± 0.001	0.761 ± 0.004	0.903 ± 0.001
ODLP	0.837 ± 0.004	0.944 ± 0.001	0.769 ± 0.004	0.909 ± 0.001
MSOD	$\mathbf{0.843 \pm 0.002}^{\star}$	$\mathbf{0.950 \pm 0.001}^{\star\dagger}$	$\mathbf{0.783 \pm 0.002}^{\star\dagger}$	$\mathbf{0.916 \pm 0.001}^{\star\dagger}$

(c) The Finches case study

	Purple Finch		**House Finch**	
	AUC	Accuracy	AUC	Accuracy
OD	0.807 ± 0.003	0.942 ± 0.001	0.758 ± 0.003	0.689 ± 0.002
ODLP	0.808 ± 0.003	0.943 ± 0.001	0.762 ± 0.003	0.696 ± 0.002
MSOD	$\mathbf{0.817 \pm 0.002}^{\star\dagger}$	$\mathbf{0.946 \pm 0.001}^{\star\dagger}$	$\mathbf{0.775 \pm 0.001}^{\star\dagger}$	$\mathbf{0.706 \pm 0.001}^{\star\dagger}$

of the MSOD model on both synthetic and eBird data. The ability to learn correct and reasonable networks holds great promise for quality control programs in citizen science data, and the ability to predict latent occupancy more accurately can help provide better species distribution models for conservation projects.

Chapter 7: Improving Predictions of Rare Species through Multi-species Distribution Modeling with an Ensemble of Classifier Chains

7.1 Introduction

There is emerging consensus that recent global change is rapidly altering species distributions [81, 66]; such changes are frequently quantified using species distribution models (SDMs) that combine observations of species occurrences with environmental factors to predict species distributions across time and locations where species data were not collected [27, 2]. Most work to date has focused on the development and refinement of single species SDMs. However, ecological interactions including competition, predation, and mutualism among species can affect species distributions [32]. Interest in capturing these ecological interactions in multi-species models is increasing [5, 48].

Past work on multi-species modeling has included model fitting approaches that analyze the parameters of fitted models to generate testable hypotheses about species interactions [65, 44]. A growing area of research in multi-species modeling involves developing predictive models [12, 27, 22, 10]. Intuitively, predictive multi-species models are expected to improve on the accuracy of single species models because they are able to leverage information about species interactions [35]. Multi-species models have been developed to capture similar responses by multiple species to environmental gradients [19], phylogenetic community structure [42] and spatial relationships [50]. Multi-species models may improve the predictive success for species that use similar habitats, represented by environmental covariates in the model, or because the species depend upon the same environmental factors. In particular, multi-species models may better predict the distribution of rare species [57, 25, 65] because of the direct or indirect interactions they have with common species that use the same geographic space.

Predictive multi-species distribution models typically differ in the order in which they assemble (seek groups of co-occurring species) and predict (seek relationships of a species to environmental factors): (1) assemble first, predict later, (2) predict first, assemble later, and (3) assemble and predict together [28]. Among these models, the *assemble and predict together* strategy [15, 64, 16] may improve the predictive success for rare species.

The field of machine learning has provided many effective approaches to single species distribution models. In particular, boosted regression trees (BRTs), which are able to model non-linear responses by discovering complex interactions among predictor variables [24] can achieve higher predictive success than other SDMs such as bioclimatic envelopes [1], genetic algorithms [78], and regression approaches such as Generalized Additive Models (GAMs) and Generalized Linear Models (GLMs) [23]. Machine learning also offers promise for providing effective multi-species models that employ the *assemble and predict together* strategy. The subfield of multi-label classification within machine learning contains many algorithms to predict multiple binary response variables (e.g. predicted presence/absence for multiple species) from covariates (e.g. environmental factors). In simple multi-label classification, response variables are considered to be independent, whereas more advanced techniques for multi-label classification leverage correlations between response variables [20]. Multi-label classification has been studied extensively in the machine learning community, where it has been applied to text labeling [31], image annotation [45] and gene function prediction [3]. We are unaware of any previous application of multi-label classifiers to species distribution modeling.

The goal of our study is to test whether incorporating information about species interactions in a multi-species model can improve predictive performance over a set of independent single-species models, with a focus on the predictive performance for rare species. Although several studies in the past have explored the differences between using a multi-species model versus a set of single-species models [36, 21, 4] only a handful have looked specifically at rare species prediction [10, 65]. Our work differs from these two past studies on rare species prediction in several ways. First, our scope is larger as we compare predictive performance over a larger number of datasets. Second, we use a multi-species model that captures more fine-grained species interactions than the hierarchical model [65]. Finally, we perform an analysis of our multi-species models to determine the model components (i.e. groups of species) that lead to improvements in the predictions of rare species. Unlike [10], we found that multi-species models tend to improve predictive performance over single-species models.

For our multi-species model, we chose a multi-label classifier known as the Ensemble of Classifier Chains (ECC) [70] for two reasons. First, ECC was the best performing multi-species model among the handful of models (specifically multivariate regression trees [15] and multivariate adaptive regression splines [21]) we evaluated in our preliminary experiments. Second, ECC can serve as an *outer wrapper* for single species models commonly used by ecologists (e.g. GLMs and BRTs) and combine these single species models into an ensemble, forming a multi-species model. Comparison of model performance among GLM and BRT models, each with and without ECC, provided controlled tests of the predictive success of multi-species vs. single-species models (Figure 7.1). We tested the predictive success and ecological relevance of these SDMs by using spatially and temporally explicit datasets sampled to capture variation in the underlying environment.

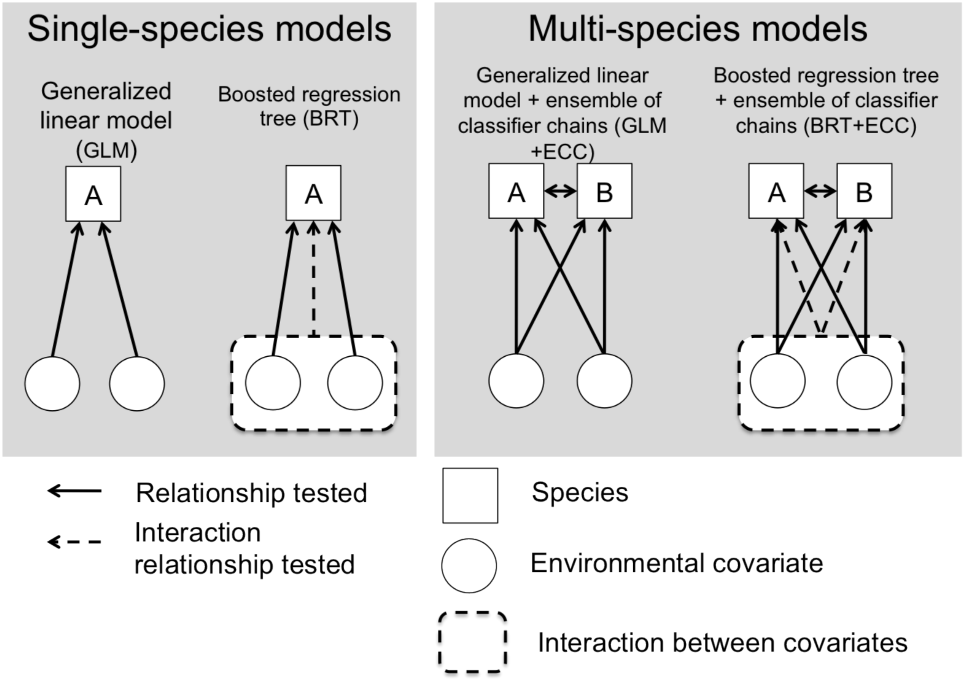

Single-species models
Generalized linear model (GLM)
Boosted regression tree (BRT)
A
A
Multi-species models
Generalized linear model + ensemble of classifier chains (GLM +ECC)
Boosted regression tree + ensemble of classifier chains (BRT+ECC)
A
B
A
B
Relationship tested
Interaction relationship tested
Species
Environmental covariate
Interaction between covariates

Figure 7.1: Schematic diagram of experiment to test how the inclusion of information on the presence of con-specifics (A, B) affects the performance of multi-species models (GLM+ECC, BRT+ECC) compared to single-species models (GLM, BRT). Single-species models were generalized linear models (GLMs) and boosted regression trees (BRTs); multi-species versions of these models were constructed using the ensemble of classifier chains (ECC).

We tested our models on long-term, landscape-scale records of mobile species (birds, insects) from two Long-Term Ecological Research (LTER) sites. These records included annual surveys of songbird distributions from 1999 to 2011 at the 3200-ha Hubbard Brook Experimental Forest (HBR) [6] and similar surveys from 2009 to 2011 at the 6200-ha HJ Andrews Experimental Forest (AND), as well as annual surveys of nocturnal moths from 1986 to 2008 at AND [37, 58]. We asked the following research questions:

1. How does prediction success compare for multi-species vs. single-species models?
2. Do multi-species models differentially improve prediction success for common versus rare species?
3. Which species have the greatest influence on predictive success for rare species in multi-species models?
4. Which ecological processes appear to explain model performance?

7.2 Methods

7.2.1 Study sites

The Andrews Forest occupies 64 km^2 in the conifer forest biome in western Oregon, USA. Mean annual precipitation is 2300 mm, but over 80% of precipitation falls between November and April. Mean daily temperature ranges from 2°C in January to 20°C in July. The landscape, which ranges from 400 to 1600 m elevation, is predominantly old-growth conifer forest dominated by Douglas-fir (Pseudotsuga menziesii) and western hemlock (Tsuga heterophylla); approximately 25% of the area is Douglas-fir plantations less than 50 years old.

The Hubbard Brook Experimental Forest occupies 31 km^2 in the temperate deciduous forest biome in northern New Hampshire, USA. The landscape, which ranges from 200 to more than 1000 m elevation, is predominantly second-growth forest following logging in the early 1900s, dominated by sugar maple (Acer saccharum), beech (Fagus grandifolia), and yellow birch (Betula alleghaniensis). Mean annual precipitation is 1400 mm, which is evenly distributed over the year. Mean daily temperature ranges from -9°C in January to 18°C in July.

7.2.2 Data

The *AND moth data* were obtained from the HJ Andrews Experimental Forest in Oregon, and consisted of a database of 32,352 individual records, representing 423 species of nocturnal moths collected by J. Miller over the period 1986 to 2006. Moths were sampled at 256 sites using a stratified random design including vegetation type, elevation, and proximity to streams. Each sampling event involved placing a light trap overnight for a single night; traps were collected the following day. Sites were sampled during the period of moth emergence (May to October), with about 20 sites sampled per night of sampling. Of the 256 sites, 157 were sampled at least twice, 32 were sampled more than 6 times, and 20 were sampled more than 40 times. Moths were identified, counted, and recorded according to date and location of collection. Sampling effort was uniformly distributed among sampling locations in study datasets; only species that were trapped at 5 or more sites were included in this analysis. A median of 42 species were trapped per site, and each of the 423 species occurred at a median of 24 sites. Given the rarity of most species and the long time period over which the data were collected, we were not able to estimate detection probability, so apparent absences in this analysis should not necessarily be considered true absences. Fifty-two environmental covariates of elevation, aspect, slope, vegetation type, mean monthly and annual temperature and precipitation were determined by overlaying sample coordinates on digital ortho-photos and GIS layers.

The *AND Bird data* and *HBR bird data* were obtained from ongoing sampling efforts at the HJ Andrews and Hubbard Brook Experimental Forests. At the AND, birds were sampled in 2009 and 2010 at 182 points using a systematic (300-m grid spacing) design stratified by elevation (460-1558 m), forest stand age (young plantation, old growth), and distance to road. At the Hubbard Brook Experimental Forest (HBR), birds were sampled at 371 points at 100-m or 200-m intervals along 15 north-south transects spaced 500 m apart, covering all elevations (240-936 m) and vegetation types.

During the peak of the avian breeding season (mid May through early July), each point was visited 6 (AND) or 3 (HBR) times; visits were separated by 1-2 weeks. The abundance of singing males of all bird species within 50 m was determined by using 10-min fixed-radius point counts [69]. Counts were conducted from 05:30 to 10:30 and did not occur during rain or strong wind (>15 kph). Points were surveyed in random order and four to five trained observers were rotated among points to reduce observer bias. Environmental covariates at AND were: field-sampled vegetation (species composition), vegetation structure derived from remotely sensed LiDAR data [33], average temperature, and elevation and distance to roads and streams determined from GIS. Environmental covariates at HBR were: elevation and vegetation characteristics from field measurements collected in 1999 and 2008. Bird presence/absence in each year was collected and extremely rare species that were detected at less than 0.5% of sites were omitted from HBR bird datasets.

7.2.3 Model development

We use the ECC algorithm as the predictive multi-species model in our analysis. In order to describe the ECC algorithm, we define the following notation. Let $X_1, \cdots, X_M$ be the set of M environmental covariates and $Y_1, \cdots, Y_S$ be the binary response variables where Y_s corresponds to the presence/absence of the sth species. The ECC algorithm relies on the concept of a classifier chain, which is a sequence of S binary classifiers. The classifier chain assumes an ordering of the response variables. For illustration, suppose the ordering follows the subscripts of the response variables $Y_1, \cdots, Y_S$. The first classifier in the chain predicts Y_1 by computing $P(Y_1|X_1, \cdots, X_M)$. The continuous-valued probability $\hat{Y}_1 = P(Y_1|X_1, \cdots, X_M)$ is used by the second classifier in the chain to predict $\hat{Y}_2 = P(Y_2|X_1, \cdots, X_M, \hat{Y}_1)$. Adding $\hat{Y}_1$ as an independent variable allows the model to account for possible correlations between species 1 and 2. During the training phase of our implementation, we used the predicted labels rather than the true labels as independent variables for the next classifier in the chain. We found this modification from the original algorithm consistently improved the prediction performance in our experiments. In general, the sth classifier predicts $\hat{Y}_s = P(Y_s|X_1, \cdots, X_M, \hat{Y}_1, \cdots, \hat{Y}_{s-1})$, where $\hat{Y}_s$ is the probability of the sth species occurring at that site given $(X_1, \cdots, X_M)$, which are the M environmental covariates at that site, and $(\hat{Y}_1, \cdots, \hat{Y}_{s-1})$, which are the predicted presences/absences of the preceding $s - 1$ species in the ordering at that site. We refer to the models for $P(Y_s|X_1, \cdots, X_M, \hat{Y}_1, \cdots, \hat{Y}_{s-1})$ as base classifiers. The base classifiers can be implemented using any of the common models used for single-species prediction (e.g. GLM and BRT).

Read et al. [70] construct an ensemble of classifier chains using L randomly-generated chain orderings. Each classifier chain in this ensemble is trained on a bootstrap replicate of the original training data that is created by sampling the original training data with replacement. Such bootstrapping is necessary because the accuracy of the results depends on the ordering of the response variables predicted by the classifier chains. To make a prediction with an ensemble of classifier chains, ECC stores the predicted probabilities for each species from each classifier chain. To compute an overall predicted probability for a particular species, ECC averages the probability for that species across chains. Importantly, this approach allows prediction of presence/absence for the taxa of interest $(Y_1, \cdots, Y_S)$ at sites that were not in the training set.

7.2.4 Temporal and spatial autocorrelation

To avoid bias from temporal autocorrelation created by birds or moths occurring at the same sites over multiple years, each bird dataset was modeled separately. To reduce bias due to spatial autocorrelation, each set of sampling points was assigned to training, validation, and testing subsets. Study sites were overlaid with a grid (approximately 1.8 by 2.4 km for AND and 0.82 by 1.2 km for HBR) and alternate grid cells were assigned to training, validation, or testing groups at a ratio of 25%/25%/50%. This checkerboard approach to validation accounts for spatial autocorrelation at scales smaller than the grid cell, which exceeds the scales at which spatial autocorrelation might be expected due to habitat patchiness or conspecific attraction [7]. The number of sites in the training, validation and testing subsets of each dataset is shown in Table 7.1.

Dataset	Training Sites	Validation Sites	Testing Sites
AND Moth	79	63	114
AND Bird (2009)	53	43	86
AND Bird (2010)	53	43	86
HBR Bird (1999)	103	94	174
HBR Bird (2008)	103	94	174

Table 7.1: The number of sites in the training, validation and testing subsets.

It is important to note that our validation approach also eliminates the potential for overfitting models. Overfitting occurs when a model fits not only the signal, but also the noise in the training data, resulting in optimistically biased estimates of model performance on the training data. More complex models (with more variables) will often perform better on the training set because they can fit the data more closely. To compare models of differing complexity fairly, we report performance for all models on a holdout test set. If overfitting occurred, we would expect more complex models to perform more poorly than simpler models on these holdout test data [76].

Parameters of the models were fit using the training data set. Some parameters of models required tuning, which we performed by using the validation data set. Specifically, for BRT, the depth of the trees and the number of trees were tuned, while for GLM, parameters controlling the regularization of the regression model were tuned (specifically we tuned the parameter α, which trades off L1 and L2 regularization, and the regularization parameter γ, which weights the regularization penalty term). The best performing model from the validation set, as determined using AUC, was then used on the holdout test dataset. Predicted vectors from model output were compared to the true observed vectors from the test data to evaluate model performance.

7.2.5 Experiments and evaluation

We conducted four experiments with the models. Experiment 1 addressed how prediction success compared for multi- vs. single-species models. In this experiment, we compared the performance of two single-species classifiers against their multi-label classifier extensions. The single-species classifiers were boosted regression trees (BRT) and elastic-net regularized general linear models (GLM). The multi-label classifier was the Ensemble of Classifier Chains (ECC) algorithm that produces the multi-label algorithms ECC-BRT and ECC-GLM.

To determine whether ECC multi-species models differentially improved prediction success for common versus rare species in Experiment 2, we compared the performance of the ECC multi-label classifier versus the single-species models on subsets of common (i.e. detections occurring at >40% of sites), moderately common (detections occurring at 10-40% of sites), and rare species (<10% prevalence of detections). Thus, for datasets with multiple years of data, different species were defined as *rare* in each year. Importantly, because we did not account for imperfect detections in these models, *rare* should be interpreted as reflecting few detections of a species. This could be a function of either true rarity, or low detection probability.

To determine which species need to be included in a multi-species model in order to most improve its predictive success on rare species in Experiment 3, we tested how the performance of the ECC models for rare species was affected by removing the common and/or moderately common species. In this experiment, ECC-BRT models were trained on environmental covariates and various combinations of species groups: (1) rare + moderately common + common species (ECC-RMC, equivalent to ECC-BRT in Experiment 1); (2) rare + common species (ECC-RC); (3) rare + moderately common species (ECC-RM); and (4) rare species (ECC-R).

To determine which species interactions appeared to explain model performance in Experiment 4, we tested how the performance of a multi-label classifier for multi-species modeling of rare species was affected by the removal of individual species that we hypothesized, based on the literature, interacted with various focal rare species via heterospecific attraction or competition. In this experiment, we used the ECC-BRT models for the 2 or 3 rare species that had the greatest improvement in performance of ECC-BRT compared to the BRT model; each ECC-BRT model was rerun $S - 1$ times (where S = total number of species) with one species dropped at a time. We then identified the species whose removal from the model caused the greatest decrease in model performance. The moth drop-one experiment was structured slightly differently, because there are more than 400 moth species in the dataset, which would be too computationally expensive to run $S - 1$ times. For rare moth multi-species models, we experimentally dropped 5-8 species one at a time that were (1) common, but do not necessarily co-occur with the target species, (2) moderate to rare and known to co-occur with the target species, and (3) moderate to rare and known not to co-occur with the target species.

In each experiment, the performance of the models was compared using the area under the receiver-operating curve (AUC), which is a measure of the relationship between the true positive rate and the false positive rate over the full range of discrimination thresholds [57]. The AUC measures the ability of a model to discriminate between sites where a species is present, vs. those where it is absent. An AUC of 0.5 indicates that a model has no discriminatory ability, while a score of 1 indicates that presences and absences are perfectly discriminated.

To evaluate performance of multi-species models, we computed a *species-based AUC*, defined as the average of the AUCs for each species in the multi-species model [21]. We compared performance of multi- to single-species models using the Δ species-based AUC (also denoted ΔAUC in the text), defined as the difference between the species-based AUC of multi-species vs. single-species models.

7.2.6 Hypothesis Testing

Testing for statistical significance between the single-species and multi-species models in Experiment 1 was challenging because the species AUCs were not independent of each other, which precluded the use of conventional hypothesis tests. We thus developed a paired-AUC difference bootstrap test to establish statistical significance for Experiment 1. This test compared the results of a multi-species model that predicted the presence/absence of all species simultaneously at a site (called method A) with a set of single species models (called method B). A description of these two methods follows. For method A, let $(u_1, \cdots, u_N)$ be N score vectors, with one score vector per site. The score vector $u_i = (u_i^1, \cdots, u_i^S)$ is a vector of S species predictions for site i, in which each component u_i^s is the probability of species s being present at site i. For method B, let $(v_1, \cdots, v_N)$ be N score vectors, with one score vector per site, i.e. the score vector $v_i = (v_i^1, \cdots, v_i^S)$ is a vector of S species predictions for site i. We paired the score vectors from methods A and B as $D = (u_1, v_1), \cdots, (u_N, v_N)$ and drew a bootstrap sample D_1 of these pairs by sampling N score pairs with replacement from D. From D_1, we computed the species-based AUC for method A and the species-based AUC for method B. We then computed Δ_1 by subtracting the species-based AUC for method B from the species-based AUC for method A. We repeated this process 999 more times, resulting in differences $\Delta_1, \cdots, \Delta_{1000}$. We sorted these differences in ascending order to obtain the 26th and 975th elements in the list. These elements formed a 95% bootstrap confidence interval on the difference. When the confidence interval did not include zero, we rejected the null hypothesis H_0 of no difference between the predictive performance (based on AUC) of the multi-species model and the set of single species models in favor of the alternate hypothesis H_1 that there was a difference. This paired-AUC bootstrap test is only applicable when comparing models that have the same number of species, hence it could not be used to compare models for rare vs. common species.

7.3 Results

Multi-species models created with ECC multi-label classifier extensions (ECC-BRT, ECC-GLM) had higher predictive success compared to single-species, environmental-covariate-only models (BRT and GLM) for all datasets, based on average species-based AUC. For bird species, average improvements in AUC were modest (approximately 0.02) for ECC-BRT compared to BRT (mean Δ species-based AUC = 0.019, range = [-0.028, 0.247]) and for ECC-GLM compared to GLM (mean Δ species-based AUC = 0.018, range = [-0.049, 0.156]). For moth species, average improvements in AUC were approximately 0.01 for ECC-BRT compared to BRT (mean Δ species-based AUC = 0.011, range = [-0.128, 0.192]), and about 0.02 for ECC-GLM compared to GLM (mean Δ species-based AUC = 0.017, range = [-0.199, 0.363]) in Figure 7.2. The majority of species (70% of bird species, 58% of moth species) were more accurately predicted by the multi-species models using boosted regression trees (ECC-BRT) than by the single-species models (BRT). Similarly, 60% of bird species and 56% of moth species were more accurately predicted by the multi-species models using generalized linear models (ECC-GLM) than by the single-species models (GLM). All multi-species models produced statistically significant improvements over their single-species counterparts, based on the paired-AUC difference bootstrap test in Table 7.2.

Although the increase in AUC for multi- vs. single-species models was small when averaged over all species in the dataset, the predictive advantages of the multi-species approach is more salient when we repeated the analysis over rare and common species and compared the average increase in AUC for these two groups in Figure 7.3. For rare species, multi-species models had significantly higher prediction success than their single-species counterparts (BRT and GLM models) for all five datasets, based on paired-AUC difference bootstrap tests in Table 7.2. In contrast, for common species, multi-species models had significantly higher predictive success than their single-species counterparts for only six of the ten model comparisons made using the five datasets in Table 7.2. The magnitude of the improvement in predictive success was small on average for common and rare species, but it differed greatly among species, datasets, and years. Models for only a few rare species experienced improvements in species-based AUC greater than 0.05 (e.g. two of nine rare species in the AND Bird 2010 dataset and two of twelve rare species in the HBR 1999 and 2008 datasets).

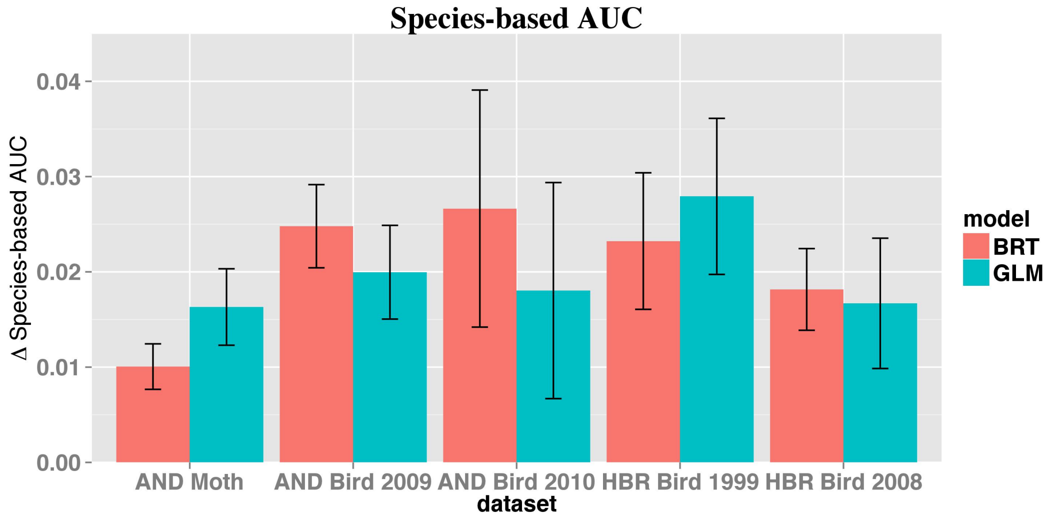
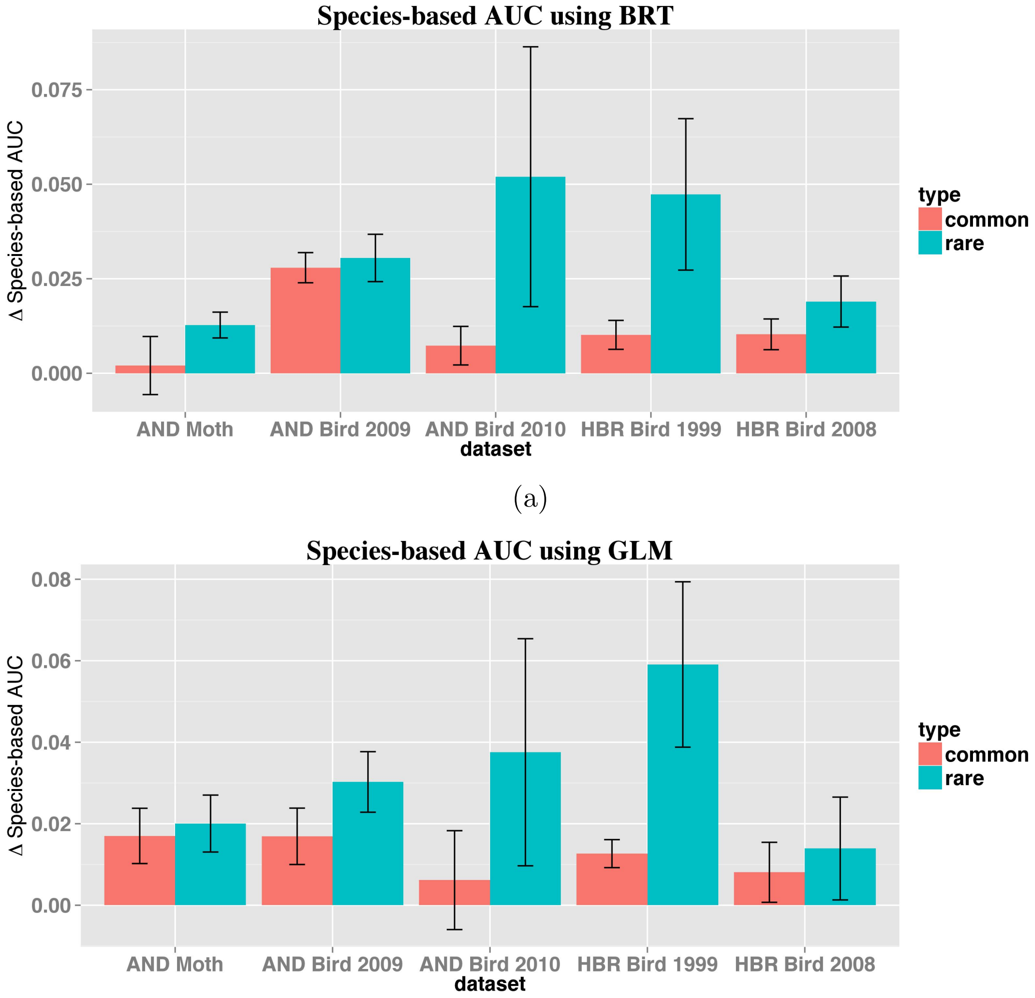


Figure 7.2: The difference in predictive success between multi-species and single-species models using BRT and GLM as base classifiers. A positive Δ species-based AUC indicates that a multi-species model has greater predictive success than single-species models. Standard errors for Δ species-based AUC are computed over S species.

Multi-species models for rare species had higher predictive success than multi-species models for common species in all five datasets when BRT was used as the base learner. Similarly, multi-species models using GLM as the base learner had higher predictive success for rare compared to common species in all five datasets, with substantially higher AUC values for rare species in HBR bird 1999 (mean AUC = 0.06 in Figure 7.3). Standard errors are close to overlapping zero in some cases, but the paired-AUC test could not be used to test for significant differences because of unequal numbers of species. The improvement in predictive success due to multi-species models differed between years in the same sites (e.g. common species at AND 2009 vs. AND 2010; rare species at HBR 1999 vs. HBR 2008), possibly because species defined as common or rare differed between years.

Multi-species models for rare species tended to have higher predictive success than the corresponding single-species models when more species were included in the model. In Figure 7.4, for all five datasets, the multi-species models that achieved the greatest improvement in predictive success were those that included the largest number of species,



(a)

(b)

Figure 7.3: Difference in predictive success of multi-species vs. single-species models for rare vs. common species, based on differences in species-based AUC for boosted regression trees (a) and general linear models (b) of five species datasets. The standard errors for Δ species-based AUC are computed over the species in each group.

Species	Dataset	Δ BRT Mean	Δ BRT 95% CI	Δ GLM Mean	Δ GLM 95% CI
All	AND Moth	0.010	(0.002, 0.018)	0.021	(0.004, 0.031)
All	AND Bird 2009	0.022	(0.012, 0.036)	0.023	(0.003, 0.033)
All	AND Bird 2010	0.024	(0.016, 0.036)	0.016	(0.000, 0.025)
All	HBR Bird 1999	0.017	(0.008, 0.026)	0.029	(0.021, 0.037)
All	HBR Bird 2008	0.014	(0.011, 0.018)	0.013	(0.006, 0.021)
Rare	AND Moth	0.009	(0.003, 0.013)	0.023	(0.000, 0.037)
Rare	AND Bird 2009	0.018	(0.013, 0.023)	0.038	(0.007, 0.071)
Rare	AND Bird 2010	0.030	(0.018, 0.036)	0.045	(0.004, 0.068)
Rare	HBR Bird 1999	0.027	(0.004, 0.051)	0.059	(0.022, 0.077)
Rare	HBR Bird 2008	0.014	(0.007, 0.020)	0.005	(0.001, 0.018)
Common	AND Moth	**0.003**	**(-0.007, 0.014)**	0.028	(0.015, 0.042)
Common	AND Bird 2009	0.021	(0.016, 0.026)	0.006	(0.002, 0.011)
Common	AND Bird 2010	**0.002**	**(-0.002, 0.008)**	**0.002**	**(-0.012, 0.019)**
Common	HBR Bird 1999	0.008	(0.005, 0.010)	0.009	(0.001, 0.017)
Common	HBR Bird 2008	0.014	(0.010, 0.018)	**0.002**	**(-0.005, 0.011)**

Table 7.2: The results of the Paired AUC Bootstrap Difference test on each dataset for all, rare and common species.

i.e. included rare, moderately common, and common species. We found considerable differences across species in the improvement produced by multi-species models, reflected in the large standard errors around means. Although the improvement in predictive success due to multi-species models were small for most species, for certain species, improvements were fairly large (Δ species-based AUC of > 0.10). The improvement in predictive success due to multi-species models also differed between years for bird datasets, because species defined as common or rare also differed between years.

We used *drop-one* experiments to explore the species interactions and the potential ecological processes that appeared to explain model performance. As expected, *drop-one* experiments revealed that multi-species model predictions for rare species were sensitive to the inclusion of species whose occurrences were correlated with the target species. However, the species that benefitted from multi-species modeling, and the associated species that improved predictions, differed between years. For example, in the AND bird datasets, *drop-one* experiments for the olive-sided flycatcher in 2009 and American

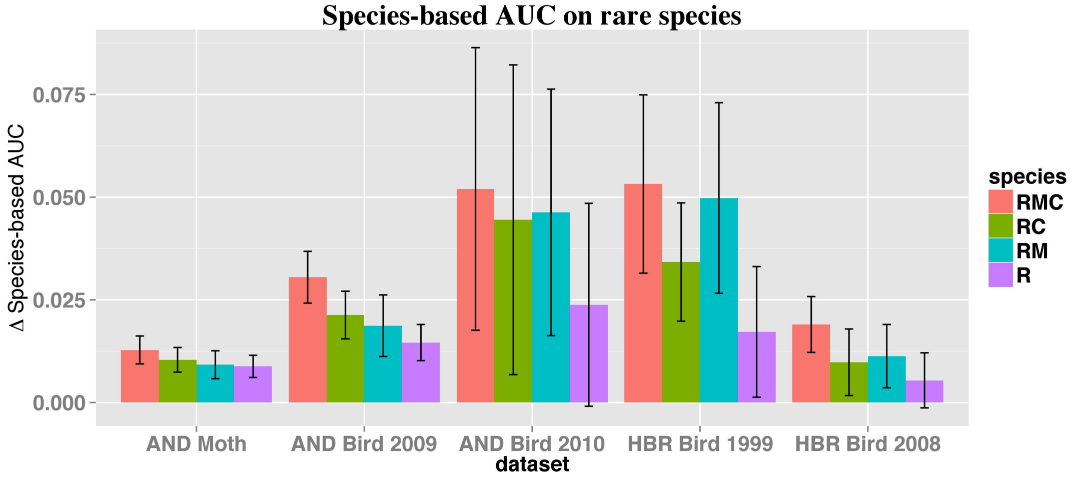


Figure 7.4: Difference in predictive success of multi-species vs. single-species models for four subsets of species (RMC, RC, RM, and R, where R = rare, M = moderately common, and C = common), based on differences in species-based AUC for boosted regression tree models (ECC-BRT vs. BRT) for five species datasets. For example, ECC-R refers to the difference in predictive success of multi-species models compared to single-species models, when only rare species are included. Error bars show the standard error of Δ species-based AUCs across all the rare species in the dataset.

robin in 2010 showed the greatest improvement from multi-species compared to single-species modeling in Figure 7.5. The model for olive-sided flycatcher in 2009 showed the greatest decline in species-based AUC after the removal of eight species (hermit warbler, dark-eyed junco, MacGillivray's warbler, Pacific-slope flycatcher, winter wren, Swainson's thrush, Wilson's warbler, and western tanager, with Δ AUC ranging from -0.045 to -0.060). These eight, relatively common, species often co-occur with the rare olive-side flycatcher in ecotones between mature forest and early successional habitats. In the HBR bird datasets, *drop-one* experiments for the wood thrush and rose-breasted grosbeak in 1999 and the black and white warbler and scarlet tanager in 2008 showed the largest improvement from multi-species compared to single-species modeling in Figure 7.6 and 7.7. The model for wood thrush in 1999 showed the greatest decline in species-based AUC after the removal of three species (winter wren, American redstart, and scarlet tanager, with Δ AUC ranging from -0.044 to -0.052). Model accuracy for the wood thrush was improved by the inclusion of the American redstart, a common species

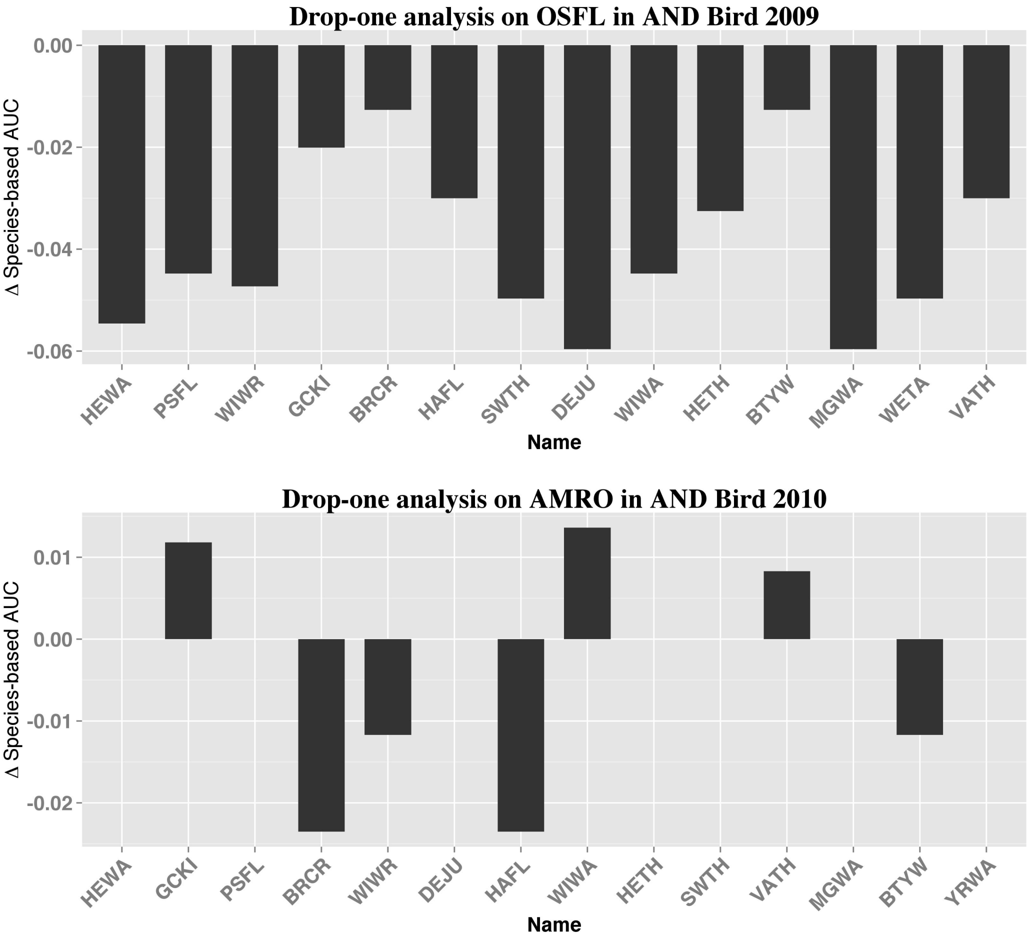


Figure 7.5: Drop-one analysis on AND Bird 2009 and 2010 data. (top) Change in species-based AUC for the rare species, olive-sided flycatcher in 2009, associated with dropping all species one by one. The changes for the 14 most influential species are shown. (bottom) Change in species-based AUC for the rare species, American robin in 2010, associated with dropping all species one by one. The changes for the 14 most influential species are shown.

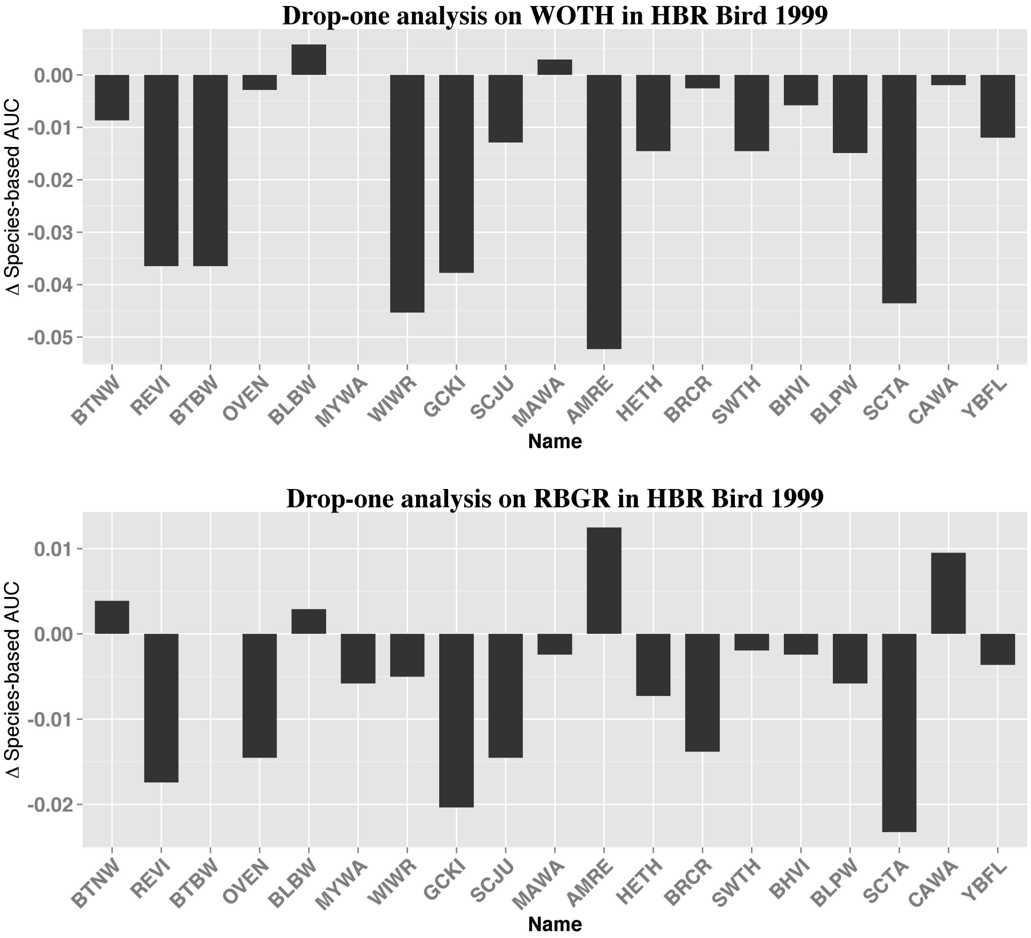
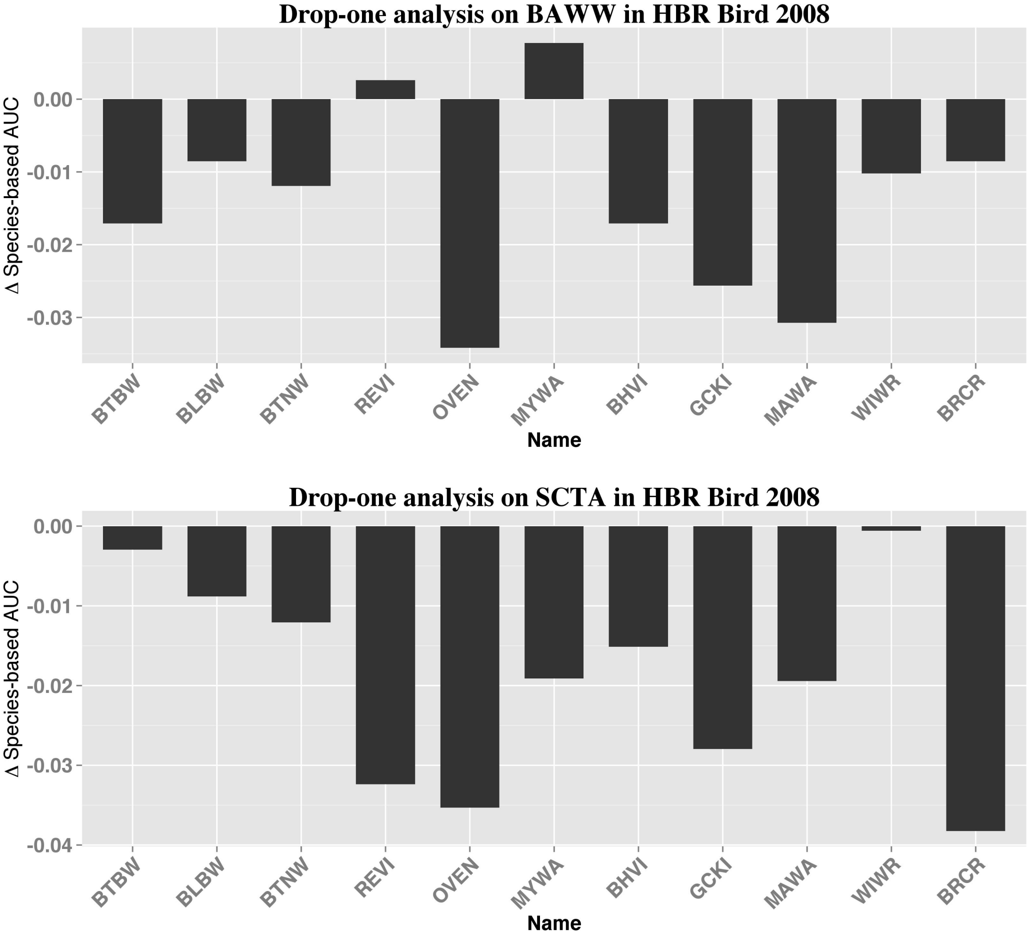


Figure 7.6: Drop-one analysis on HBR Bird 1999 data. (top) Change in species-based AUC for the rare species, wood thrush and (second from top) rose grosbeak, in 1999, associated with dropping all species one by one. The changes for the 19 most influential species are shown.


Drop-one analysis on BAWW in HBR Bird 2008

Drop-one analysis on SCTA in HBR Bird 2008

Figure 7.7: Drop-one analysis on HBR Bird 2008 data. Change in species-based AUC for the rare species, black-and-white warbler (top) and scarlet tanager (bottom) in 2008, associated with dropping all species one by one. The changes for the 11 most influential species are shown.

associated with second-growth deciduous forest, whose occurrence was significantly correlated with that of the wood thrush. Although the wood thrush is generally associated with mature forest, it can also be found along forest edges and in the dense understory of second-growth stands. In the AND nocturnal moth dataset, *drop-one* experiments Zanclognatha jacchusalis and Lithophane baileyi showed the greatest improvement from multi-species compared to single-species modeling in Figure 7.8. The model for Zanclognatha jacchusalis showed the greatest decline in species-based AUC when two rare species Lambdina fiscellaria and Eurois astricta were removed from the model. These three species occurred in the same habitat at different time periods, or in different habitats at the same time periods, or shared food preferences for vegetation that occurred in the same habitat. The model for Lithophane baileyi experienced the greatest decline in AUC when three rare species Nepytia phantasmaria, Lambdina fiscellaria and Lasionycta perplexa were removed from the model; these species occurred in similar habitats [39].

7.4 Discussion

7.4.1 Performance of multi-species vs. single-species models

In both boosted regression trees and generalized linear models, multi-species models improved predictive success over single species models. The improvement was greater for BRTs than for GLMs and for rare compared to commonly occurring species. These results suggest that accounting for interactions between environmental covariates (the benefits of BRT) and leveraging information from other species (the benefits of ECC) both improved model performance over simpler models without these features. Although multi-species models have more parameters and can better describe the training datasets irrespective of any ecological explanatory power, our predictive success is measured over a test dataset that is separate from the training dataset used to fit the models. In addition, both GLMs and BRTs guard against overly complex models. GLMs have a regularization term that penalizes the magnitude of the model coefficients and BRTs guard against overfitting by limiting the depth of the trees in the ensemble. The improvements in predictive success were relatively small when averaged across all species, but improvements were quite large for some rare species in some years. Our findings are consistent with those of similar studies that have compared multi-species models with single species models. For instance, [21] report that multi-species (MARS) models applied to datasets with approximately 30 species improved the median AUC by about 0.013 and that 50 to 85% of species were more accurately predicted by using a multi-species model. In our study, which included 23 (AND) and 44 (HBR) species, multi-species models improved success for similar proportions of species - 60% for GLM and 70% for BRT. Thus, the ECC algorithm, which is an approach from the field of machine learning, seems comparable with the performance of other multi-species models previously used for species distribution modeling.

Multi-species models appear to leverage information about widespread species to aid predictions of distributions and abundance of rare species that are difficult to model accurately because of the scarcity of observations [25]. In our analysis, model performance improved for rare species to a greater extent than for common species in a single-species model experiment that tested the effect of including interactions among environmental covariates (i.e. BRT vs. GLM) and in multi-species model experiments that tested the effect of including species interactions (i.e. ECC-BRT vs. BRT and ECC-GLM vs. GLM). Model performance for multi-species models of rare species improved consistently as a greater number of widespread species were included in the analysis. However, multi-species models did not improve performance for all rare species. In particular, multi-species methods improved SDM model performance for a greater proportion of rare bird species than for rare moth species.

We defined rare species as those with a prevalence of detections of <10% in the dataset for any given year; sampling effort was uniformly distributed among sampling locations in study datasets. When this definition was applied to multi-year datasets (birds at AND, birds at HBR), some species were rare in one year, but not in another. Although multi-species models improved prediction success for rare species much more than for common species, species whose prediction success was improved by multi-species modeling in one year were not necessarily improved (or rare) in another year. Also, for a very speciose dataset (moths at AND), multi-species models improved prediction success for some, but not all, rare species.

It is important to acknowledge that imperfect detection of rare species is a critical issue; since we do not know the true occupancy of our sites, we cannot use occupancies as ground truth to measure predictive performance. Instead, we treat detections (i.e. species presences and absences) as the ground truth for predictive success, as has been done in past work on SDMs [24, 50, 19]. The errors due to detection, however, do not affect our general conclusion that multi-species models improve the predictive performance of models for detecting rare species. Developing multi-species occupancy-detection models is an active area of research [88, 18].

7.4.2 Mechanisms for improved performance of multi-species distribution models

A principal value of multi-species distribution models is their potential to provide insights into species interactions and response to environmental change. Because species may directly influence the presence of other species via competition, facilitation or predation [83], we would expect multi-species models to perform better than single species models. Several indirect *interactions* may also have contributed to the greater predictive success of multi-species models in our study: (a) a species, which is difficult to predict from environmental covariates, may share measured environmental covariates with more common species, hence improving predictions, (b) species are predicted by environmental covariates, but also interact through some biotic mechanism (e.g. mutualism, competition) and this adds confidence to predictions from environmental covariates, and (c) species may co-occur as the result of important, missing (unmeasured) environmental covariates in Figure 7.9. These three sets of indirect mechanisms are more likely to produce greater improvement for rare than common species. For example, multi-species models for olive-sided flycatcher in 2009 and American robin in 2010 at AND, which were rare, were likely improved by the co-occurrence of other species sharing similar vegetation types old growth forest in the case of the flycatcher, and early successional forest in the case of the robin (Figure 7.9 a). At HBR, multi-species models for rose-breasted grosbeak in 1999 were improved by the addition of golden-crowned kinglets that had a negative association with rose-breasted grosbeaks. Kinglets are birds of the transition zone and spruce-fir forest, whereas rose-breasted grosbeaks occur only in northern hardwood forests, so their preference for very different vegetation types was likely the key cause of improved multi-species model performance (Figure 7.9 b). Given the very different nesting and foraging habits of these two species (Sallabanks et al. 1999, Swanson et al. 2012) , it is highly unlikely that improved predictive success reflects competitive interactions. Multi-species models for the very rare nocturnal moth, Zanclognatha jacchusalis, were improved by the inclusion of other very rare moth species with which they are known to share food preferences; this is most likely an example of a covariate not included in the models (Figure 7.9 c).

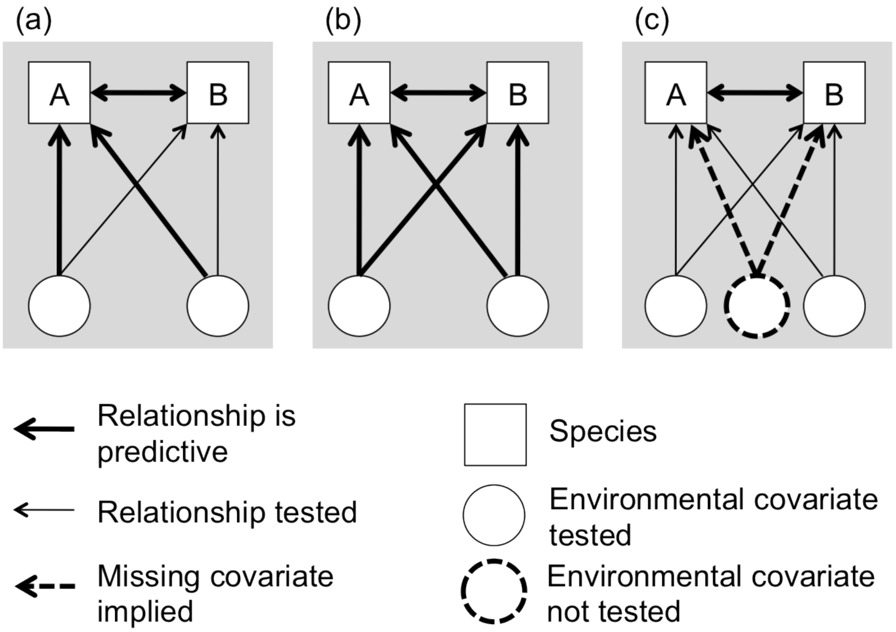

(a)
(b)
(c)
A
B
A
B
A
B
Relationship is predictive
Relationship tested
Missing covariate implied
Species
Environmental covariate tested
Environmental covariate not tested

Figure 7.9: Three different sets of conditions under which ecological processes result in improved species distribution models when boosted regression trees (BRTs) are combined with ensemble of classifier chains (ECC). (a) Species B cannot be effectively predicted from environmental covariates (e.g. because it is rare), but it appears to share A's environmental covariates, (b) Species A and B are predicted by environmental covariates, but also interact through some biotic mechanism (e.g. mutualism, competition) and this adds confidence to predictions from environmental covariates, and (c) The relationship between species A and B suggests that an unmeasured environmental covariate may influence both species.

To test whether co-occurrence data could be leveraged to increase predictive success, we used datasets collected at fine spatial resolutions, but broad spatial extents as recommended in [83]. In previous work, multi-species models did not improve predictions over single species models in datasets with low resolution (sample points spaced at 10 to 50 km) [4, 10]. However, multi-species models out-performed single-species models in datasets with higher resolution (sample points spaced <1 km) [21]. Interspecific interactions such as heterospecific attraction and competition tend to occur at the scale of individual animal territories or home ranges. Also, the advantage of multi-species models over single-species models may decline as the grain size of studies increases relative to the scale of habitat patchiness, because of the chance of not sampling some habitats increases with grain size.

Although multi-species models improved predictive success for the majority of both bird and moth species, the species that were defined as *rare* changed across years, and the improvements in predictive success for specific species achieved by multi-species models were not consistent from year to year. Thus, multi-species models, applied to a single year of data may not perform well when attempting to predict distributions over time, potentially limiting their capacity to detect species responses to environmental change. Many studies have predicted that biotic interactions will change over time as species respond individualistically to changes in environmental conditions [67, 79] and species immigration may produce unpredictable changes to community composition [6]. On the other hand, multi-species models constructed from longer or larger datasets may provide defensible baselines for detecting changes in relationships among species and habitats. Further work is needed to explore this issue.

7.4.3 Future research directions for modeling multi-species distributions

For computer scientists, some unique properties of species data may lead to new research directions for multi-label classification. First, almost none of the current work on multi-label algorithms focuses on accurately predicting the presence/absence of rare species, despite the importance of this topic to ecologists and conservation biologists. Second, interactions between species are complex, with community structure driven at least partially by facilitation, competition and predation. Many existing multi-label algorithms model only pairwise linear correlations between species and thus are inadequate for helping ecologists discover the more complex interactions that can occur in nature. Third, algorithms that can discover complex patterns of correlations often focus purely on making accurate predictions and do not easily suggest explanations as to why a particular prediction was made; however, such explanations are important for generating testable hypotheses about the interactions among species and among species and environmental variables. Therefore, further work on multi-label classifiers could focus on developing approaches that permit comparing the relative importance of two major forces that structure ecological communities – abiotic factors and biological interactions among species.

7.5 Conclusion

Our results indicate that multi-species models resulted in modest improvements to single species models. These improvements were greatest for rare species, which are traditionally difficult to predict. Modeling experiments with multi-species models also suggested testable hypotheses for ecological patterns and processes, including species interactions. These findings point to new research directions, including: developing new multi-label algorithms for predicting rare species more accurately, modeling more complex interactions beyond pairwise species correlations, and producing more easily interpretable multi-label models. These research directions may provide valuable contributions both to the fields of machine learning and ecology.

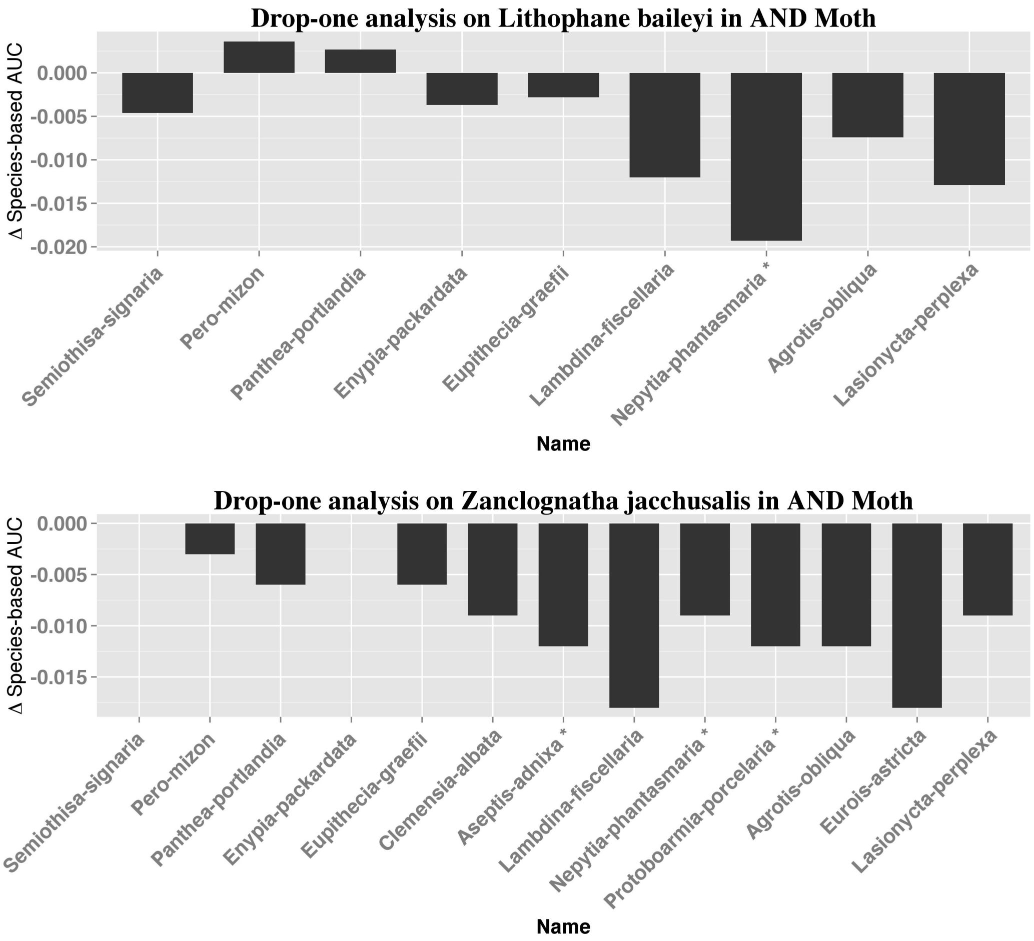


Figure 7.8: Drop-one analysis on AND Moth data for 1986 to 2008. (top) The change in species-based AUC for the rare species, Zanclognatha jacchusalis, associated with dropping thirteen selected species one by one. (bottom) change in species-based AUC for the rare species, Lithophane baileyi, 1986 to 2008, associated with dropping nine selected species one by one.

Chapter 8: Conclusions and Future Work

8.1 Contributions

Citizen Science encourages volunteers from the general public to participate in scientific studies, allowing data to be collected at much larger spatial and temporal scales. Although citizen scientists can contribute large volumes of data, data quality is often a concern due to variability in the skills of volunteers. Therefore it is crucial to improve the quality of the citizen science data so that more accurate information can be extracted. In my thesis, I investigate applying machine learning techniques to improve the quality of data submitted to citizen science projects.

The context of my work is eBird, which is one of the largest citizen science projects in existence. In the eBird project, citizen scientists act as a large global network of human sensors, recording observations of bird species and submitting these observations to a centralized database where they are used for ecological research such as species distribution modeling and reserve design. There are two aspects to improve the quality of eBird data. First, we must be able to identify invalid observations and remove them from the eBird database so that more accurate species distribution models can be created. Secondly, since eBird participants act as trainable human sensors, we can extract knowledge from data to teach the inexperienced birders and improve their birding skills so that they can contribute better quality data.

My thesis addresses problems in both aspects. To identify invalid observations, we develop an automated data verification process (Chapter 4) which leverages a birder's skill level. A birder's skill level can be graded using the Occupancy-Detection-Expertise model (Chapter 3) when the labeled birders are available or using the mixture of Species Accumulation Curves model (Chapter 5) when there are no labeled birders. Also we propose a Multi-Species Occupancy-Detection model to detect misidentification of bird species (Chapter 6) and this information can be used to help inexperienced birders distinguish groups of confusing species. In conclusion, my study shows that machine learning can be used to improve data quality in eBird by modeling an observer's skill level, developing an automated data verification model and discovering groups of misidentified species.

8.2 Future Work

In the Occupancy-Detection-Expertise work, there are two directions that we would like to explore. First, since non-linear interactions often exist in the real world data, we would like to replace the logistic regression parts of the ODE model with more flexible function approximators such as boosted trees which allow non-linear interactions between features. In addition, we would like to learn the ODE model in a semi-supervised fashion. Since most of the eBird participants are unlabeled, leveraging their observations during training may help create more accurate SDMs.

In the automated data verification work, we would like to improve the expertise prediction of the ODE model by including more expertise covariates. For example, an individual's regional expertise may be important, as a birder can be an expert observer in their home region, but less so outside of that region. In addition, we would like to test this automated data filter more broadly across the US.

In the mixture of Species Accumulation Curves work, we would like to determine the number of groups from data using nonparametric Bayesian approach. This allows the mixture model to adjust the number of groups automatically as new observational data are collected. Furthermore, we plan to extend this model to capture the evolution of an observer's skill level over time. For example, birders may switch from one group to a group of better skill level as their skills improve through practice over the years.

In the Multi-Species Occupancy-Detection work, the exact inference is computationally expensive with large number of species, so we would like to speed up the learning using variational approximations [43, 61, 68]. Currently, the MSOD model does not capture the correlations between the occupancy status of species, including competition, predation, and mutualism among species. We plan to extend the MSOD model to capture species interaction in the future.

In the multi-species distribution modeling work, we plan to develop learning algorithms which can take weights of species into account during learning [17] so that the predictions improve more on certain species of interest (e.g. the endangered species). Another future direction is to develop multi-species distribution models which can incorporate prior knowledge of species interactions from ecologists during learning.

8.3 Acknowledgements

This thesis is based upon work supported by the National Science Foundation under Grants No. 0941748 and 1209714. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors, and do not necessarily reflect the views of the National Science Foundation.

Bibliography

[1] M. B. Araujo, R. G. Pearson, W. Thuiller, and M. Erhard. Validation of species-climate impact models under climate change. *Global Change Biology*, 11:1504–1513, 2005.

[2] M. P. Austin. Spatial prediction of species distribution: an interface between ecological theory and statistical modeling. *Ecological Modelling*, 157:101–118, 2002.

[3] Z. Barutcuoglu, R. E. Schapire, and O. G. Troyanskaya. Hierarchical multi-label prediction of gene function. *Bioinformatics*, 22:830–836, 2006.

[4] A. Baselga and M. Araujo. Individualistic vs community modeling of species distributions under climate change. *Ecography*, 32:55–65, 2009.

[5] A. Baselga and M. B. Araujo. Do community-level models describe community variation effectively? *Biogeography*, 37:1842–1850, 2010.

[6] M. G. Betts, J. J. Nocera, and A. S. Hadley. Settlement in novel habitats induced by social information may disrupt community structure. *Condor*, 112:265–273, 2010.

[7] M.G. Betts, A.W. Diamond, G.J. Forbes, M.-A. Villard, and J. Gunn. The importance of spatial autocorrelation, extent and resolution in predicting forest bird occurrence. *Ecological Modelling*, 191:197–224, 2006.

[8] R. Byrd, P. Lu, J. Nocedal, and C. Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

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

[10] D. S. Chapman and B. V. Purse. Community versus single-species distribution models for british plants. *Biogeography*, 38:1524–1535, 2011.

[11] J. P. Cohn. Citizen science: Can volunteers do real research? *BioScience*, 58(3):192–197, 2008.

[12] G. De'ath. Boosted trees for ecological modeling and prediction. *Ecology*, 88:243–251, 2007.

[13] D. G. Delaney, C. D. Sperling, C. S. Adams, and B. Leung. Marine invasive species: validation of citizen science and implications for national monitoring networks. *Biological Invasions*, 10(1):117–128, 2008.

[14] 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.

[15] G. Death. Multivariate regression trees: A new technique for modeling species-environment relationships. *Ecology*, 83:1005–1117, 2002.

[16] T. Dirnbock, S. Dullinger, M. Gottfried, C. Ginzler, and G. Grabherr. Mapping alpine vegetation based on image analysis, topographic variables and canonical correspondence analysis. *Applied Vegetation Science*, 6:85–96, 2003.

[17] Janardhan Rao Doppa, Alan Fern, and Prasad Tadepalli. Output space search for structured prediction. In *ICML*, 2012.

[18] R. M. Dorazio, M. Kery, J. A. Royle, and M. Plattner. Models for inference in dynamic metacommunity systems. *Ecology*, 91:2466–2475, 2010.

[19] P. K. Dunstan, S. D. Foster, and R. Darnell. Model based grouping of species across environmental gradients. *Ecological Modelling*, 222:955–963, 2011.

[20] A. Elisseeff and J. Weston. A kernel method for multi-labelled classification. In *Advances in Neural Information Processing Systems*, pages 681–687, 2002.

[21] J. Elith and J. Leathwick. Predicting species distributions from museum and herbarium records using multi-response models fitted with multivariate adaptive regression splines. *Diversity and Distributions*, 13:265–275, 2007.

[22] J. Elith and J. R. Leathwick. Species distribution models: ecological explanation and prediction across space and time. *Annual Review of Ecology and Systematics*, 40:677–697, 2009.

[23] J. Elith, J. R. Leathwick, and T. Hastie. A working guide to boosted regression trees. *J Anim Ecol*, 77:802–813, 2008.

[24] Jane Elith, C. H. Graham, R. P. Anderson, M. Dudik, S. Ferrier, A. Guisan, R. J. Hijmans, F. Huettmann, J. R. Leathwick, A. Lehmann, J. Li, L. G. Lohmann, B. A. Loiselle, G. Manion, C. Moritz, M. Nakamura, Y. Nakazawa, J. M. Overton, A. T. Peterson, S. J. Phillips, K. Richardson, R. Scachetti-Pereira, R. E. Schapire, J. Soberon, S. Williams, M. S. Wisz, and N. E. Zimmermann. Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 29:129–151, 2006.

[25] R. Engler, A. Guisan, and L. Rechsteiner. An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. *Applied Ecology*, 41:263–274, 2004.

[26] K. A. Ericsson and N. Charness. Expert performance: Its structure and acquisition. *American Psychologist*, 49:525–747, 1994.

[27] S. Ferrier, M. Drielsma, G. Manion, and G. Watson. Extended statistical approaches to modelling spatial pattern in biodiversity in northeast new south wales. ii. community-level modelling. *Biodiversity and Conservation*, 11(12):2309–2338, 2002.

[28] S. Ferrier and A. Guisan. Spatial modelling of biodiversity at the community level. *Journal of Applied Ecology*, 43:393–404, 2006.

[29] Daniel Fink, Wesley M. Hochachka, Benjamin Zuckerberg, David W. Winkler, Ben Shaby, M. Arthur Munson, Giles Hooker, Mirek Riedewald, Daniel Sheldon, and Steve Kelling. Spatiotemporal exploratory models for broad-scale survey data. *Ecological Applications*, 20:21312147, 2010.

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

[31] N. Ghamrawi and A. McCallum. Collective multi-label classification. In *CIKM*, pages 195–200, 2005.

[32] W. Godsoe and L. J. Harmon. How do species interactions affect species distribution models? *Ecography*, 35:810–820, 2012.

[33] S.J. Goetz, D. Steinberg, M.G. Betts, R.T. Holmes, P.J. Doran, R. Dubayahand, and M. Hofton. Lidar remote sensing variables predict breeding habitat of a neotropical migrant bird. *Ecology*, 96:1569–1576, 2010.

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

[35] A. Guisan and C. Rahbek. Sesam - a new framework integrating macroecological and species distribution models for predicting spatio-temporal patterns of species assemblages. *Biogeography*, 38:1433–1444, 2011.

[36] A. Guisan, S. B. Weiss, and A. D. Weiss. Glm versus cca spatial modeling of plant species distribution. *Plant Ecology*, 143:107–122, 1999.

[37] P. C. Hammond and J. C. Miller. Comparison of the biodiversity of lepidoptera within three forested ecosystems. *Annals of the Entomological Society of America*, 91:323–328, 1998.

[38] David Heckerman. A tractable inference algorithm for diagnosing multiple diseases. In *Proceedings of the Fifth Conference on Uncertainty in Artificial Intelligence*, pages 163–172, 1989.

[39] S. A. Highland, J. C. Miller, and J. A. Jones. Determinants of moth diversity and community in a temperate mountain landscape: vegetation, topography, and seasonality. *Ecosphere*, 2013.

[40] 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 andEvolution*, 27(2):130–137, 2012.

[41] Su in Lee, Honglak Lee, Pieter Abbeel, and Andrew Y. Ng. Efficient l1 regularized logistic regression. In *AAAI*, 2006.

[42] A. R. Ives and M. R. Helmus. Generalized linear mixed models for phylogenetic analyses of community structure. *Ecological Monographs*, 81:511–525, 2011.

[43] Tommi S. Jaakkola and Michael I. Jordan. Variational probabilistic inference and the qmr-dt network. *Journal of Artificial Intelligence Research*, 10:291–322, 1999.

[44] M. M. Jackson, M. G. Turner, S. M. Person, and A. R. Ives. Seeing the forest and the trees: multilevel models reveal both species and community patterns. *Ecosphere*, 3:79, 2012.

[45] S. Ji, L. Tang, S. Yu, and J. Ye. A shared-subspace learning framework for multi-label classification. *ACM TKDD*, 4(2):8:1–8:29, 2010.

[46] S. Kelling, J. Gerbact, D. Fink, C. Lagoze, W-K. Wong, J. Yu, T. Damoulas, and C. Gomes. ebird: A human/computer learning network for biodiversity conservation and research. In *Proceedings of the Twenty-Fourth Annual Conference on Innovative Applications of Artificial Intelligence*, 2012.

[47] Steve Kelling, Carl Lagoze, Weng-Keen Wong, Jun Yu, Theodoros Damoulas, Jeff Gerbracht, Daniel Fink, and Carla P. Gomes. ebird: A human/computer learning network to improve conservation and research. *AI Magazine*, 34(1):10–20, 2013.

[48] W. D. Kissling. Towards novel approaches to modelling biotic interactions in multispecies assemblages at large spatial extents. *Biogeography*, 2012.

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

[50] A. M. Latimer, S. Banerjee, H. Sang, E. S. Mosher, and J. A. Silander. Hierarchical models facilitate spatial analysis of large data sets: a case study on invasive plant species in the northeastern united states. *Ecology Letters*, 12:144–154, 2009.

[51] John Leathwick, Atte Moilanen, Malcolm Francis, Jane Elith, Paul Taylor, Kathryn Julian, Trevor Hastie, and Clinton Duffy. Novel methods for the design and evaluation of marine protected areas in offshore waters. *Conservation Letters*, 1:91–102, 2008.

[52] 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 surveyn. *Monthly Notices of the Royal Astronomical Society*, 389:1179–1189, 2008.

[53] D. C. Liu and J. Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical Programming*, 45(3):503–528, 1989.

[54] 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.

[55] 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.

[56] Darryl I. MacKenzie, Larissa L. Bailey, and James D. Nichols. Investigating species co-occurrence patterns when species are detected imperfectly. *Journal of Animal Ecology*, 73:546–555, 2004.

[57] S. Manel, H. C. Williams, and S. J. Ormerod. Evaluating presence-absence models in ecology: the need to account for prevalence. *Applied Ecology*, 38:921–931, 2001.

[58] J. C. Miller, P.C. Hammond, and D.N.R. Ross. Distribution and functional roles of rare and uncommon moths across a coniferous forest landscape. *Annals of the Entomological Society of America*, 96:847–855, 2003.

[59] A. Munson, R. Caruana, D. Fink, W. Hochachka, M. Iliff, K. Rosenberg, D. Sheldon, B. 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, 2010.

[60] 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.

[61] Andrew Y. Ng and Michael I. Jordan. Approximate inference algorithms for two-layer bayesian networks. In *Advances in Neural Information Processing Systems*, volume 12, 1999.

[62] Emily Nicholson and Hugh P. Possingham. Making conservation decisions under uncertainty for the persistence of multiple species. *Ecological Applications*, 17(1):251–265, 2007.

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

[64] J. D. Olden. Species-specific approach to modeling biological communities and its potential for conservation. *Conservation Biology*, 17:854–863, 2003.

[65] O. Ovaskainen and J. Soininen. Making more out of sparse data: hierarchical modeling of species communities. *Ecology*, 92:289–295, 2011.

[66] C. Parmesan and G. Yohe. A globally coherent fingerprint of climate change impacts across natural systems. *Nature*, 421:37–42, 2003.

[67] R.G. Pearson and T.P. Dawson. Predicting the impacts of climate change on the distribution of species: are bioclimatic envelope models useful? *Global Ecology and Biogeography*, 12:361–371, 2003.

[68] John C. Platt, Emre Kiciman, and David A. Maltz. Fast variational inference for large-scale internet diagnosis. In *Advances in Neural Information Processing Systems*, volume 20, 2007.

[69] C. J. Ralph. Monitoring bird populations by point counts. *Agriculture, editor*, 1995.

[70] J. Read, B. Pfahringer, G. Holmes, and E. Frank. Classifier chains for multi-label classification. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 254–269, 2009.

[71] J. A. Royle and W. A. Link. Generalized site occupancy models allowing for false positive and false negative errors. *Ecology*, 87(4):835–841, 2006.

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

[73] J. R. Sauer, B. G. Peterjohn, and W. A. Link. Observer differences in the north american breeding bird survey. *The Auk*, 111(1):50–62, 1994.

[74] Neil Savage. Gaining wisdom from crowds. *Commun. ACM*, 55(3):13–15, 2012.

[75] S. A. Sheppard and Loren Terveen. Quality is a verb: the operationalization of data quality in a citizen science community. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, pages 29–38, 2011.

[76] S.M. Shirley, Z. Yang, R.A. Hutchinson, J.D. Alexander, K. McGarigal, and M.G. Betts. Species distribution modelling for the people: unclassified landsat tm imagery predicts bird occurrence at fine resolutions. *Diversity and Distributions*, 19:855–866, 2013.

[77] M.A. Shwe, B. Middleton, D.E. Heckerman, M. Henrion, E.J. Horvitz, H.P. Lehmann, and G.F. Cooper. Probabilisitc diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base. *Methods of Information in Medicine*, 30:241–255, 1991.

[78] A. K. Stockman, D. A. Beamer, and J. E. Bond. An evaluation of a garp model as an approach to predicting the spatial distribution of non-vagile invertebrate. *Diversity and Distributions*, 12:81–89, 2006.

[79] D. Stralberg, D. Jongsomjit, C. A. Howell, M. A. Snyder, J. D. Alexander, J. A. Wiens, and T. L. Root. Re-shuffling of species with climate disruption: a no-analog future for california birds? *Plos One*, 4, 2009.

[80] 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.

[81] P. M. Vitousek, H. A. Mooney, J. Lubchenco, and J. M. Melillo. Human domination of earths ecosystems. *Science*, 277:494–499, 1997.

[82] J. Hardin Waddle, Robert M. Dorazio, Susan C. Walls, Kenneth G. Rice, Jeff Beauchamp, Melinda J. Schuman, and Frank J. Mazzotti. A new parameterization for estimating co-occurrence of interacting species. *Ecological Applications*, 20(5):1467–1475, 2010.

[83] M. S. Wisz, J. Pottier, W. D. Kissling, L. Pellissier, J. Lenoir, C. F. Damgaard, C. F. Dormann, M. C. Forchhammer, J-A. Grytnes, A. Guisan, R. K. Heikkinen, T. T. Høye, I. Kuhn, M. Luoto, L. Maiorano, M-C. Nilsson, S. Normand, E. Ockinger, N. M. Schmidt, M. Termansen, A. Timmermann, D. A. Wardle, P. Aastrup, and

J-C. Svenning. The role of biotic interactions in shaping distributions and realized assemblages of species: implications for species distribution modelling. *Biological Reviews*, 2012.

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

[85] 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.

[86] J. Yu, W.-K. Wong, and R. Hutchinson. Modeling experts and novices in citizen science data for species distribution modeling. Technical report, Oregon State University, 2010. http://hdl.handle.net/1957/18806.

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

[88] E. F. Zipkin, J. A. Royle, D. K. Dawson, and S. Bates. Multi-species occurrence models to evaluate the effects of conservation and management actions. *Biological Conservation*, 143:479–484, 2010.