Modeling Experts and Novices in Citizen Science Data for Species Distribution Modeling

Jun Yu, Weng-Keen Wong and Rebecca A. Hutchinson
{yuju,wong,rah}@eecs.oregonstate.edu

School of EECS
1148 Kelley Engineering Center
Oregon State University
Corvallis, OR 97331

October 11, 2010

Abstract

Citizen scientists, who are volunteers from the community that participate as field assistants in scientific studies [3], enable research to be performed at much larger spatial and temporal scales than trained scientists can cover. Species distribution modeling [6], which involves understanding species-habitat relationships, is a research area that can benefit greatly from citizen science. The eBird project [18] is one of the largest citizen science programs in existence. By allowing birders to upload observations of bird species to an online database, eBird can provide useful data for species distribution modeling. However, since birders vary in their levels of expertise, the quality of data submitted to eBird is often questioned. In this paper, we develop a probabilistic model called the *Occupancy-Detection-Expertise* (ODE) model that incorporates the expertise of birders submitting data to eBird. We show that modeling the expertise of birders can improve the accuracy of predicting observations of a bird species at a site. In addition, we can use the ODE model for two other tasks: predicting birder expertise given their history of eBird checklists and identifying bird species that are difficult for novices to detect.

Keywords: Applications, Bayesian Networks, Graphical Models, Species Distribution Modeling, Contrast Mining, Citizen Science

1 INTRODUCTION

The term *Citizen Science* refers to scientific research in which volunteers from the community participate in scientific studies as field assistants [3]. 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) [6] 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 [4]. However, for difficult-to-detect organisms, Fitzpatrick et al. [7] found differences between observations made by volunteers and by experienced scientists led to biases in their results.

The eBird project [18], launched in 2002 by the Cornell Lab of Ornithology and National Audubon Society, is one of the largest citizen science programs in existence. The eBird project maintains an online database that allows bird watchers (known as *birders*) to submit checklists that record the bird species they have seen or heard. eBird's goal is to maximize the utility and accessibility of the vast numbers of bird observations made each year by recreational and professional birders. As an example of the volume of data submitted, in January 2010, participants reported more than 1.5 million bird observations across North America.

SDM can, in theory, benefit greatly from data collected by eBird. The goal of SDM is to predict the presence/absence or abundance of a species at a geographic site. SDM is an important task in ecology for several reasons. SDM helps ecologists understand species-habitat relationships, which in turn helps ecologists predict biodiversity. Prediction of biodiversity is an important factor in conservation planning and reserve design. In addition, by understanding species-habitat relationships, ecologists can predict species invasions and identify areas at risk.

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 [11]. 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. Although the focus of this paper is on species distribution modeling, the occupancy / detection problem is representative of a more general problem in domains such as object recognition and surveillance 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 (eg. an additive Gaussian noise term), but in some situations the detection process is affected by conditions during detection and requires a more complex model.

Mackenzie et al. [17] 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 3.1. Recent work [12] has applied the OD model to citizen science checklist data similar to those from eBird. 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.

2 RELATED WORK

The goal of SDM is to understand the relationship between species and their habitat, which is characterized by a set of environmental features. Two very common groups of methods include envelope models, which determine a profile of suitable habitat in environmental feature space [2] and genetic algorithms such as GARP [21], which select rules that make the best predictions. Since SDM can be viewed as a supervised machine learning problem, many methods in statistics and machine learning have also been used, including GLMs/GAMs [1], Hierarchical Bayesian models [14], boosted regression trees [9], and Maximum Entropy models [19]. The Maximum Entropy models developed by Phillips et al. are primarily intended to be applied to presence-only data. Presence-only methods are not applicable to the eBird data set because eBird observers report both presence and absence information.

SDMs that do not explicitly model the detection process can incorrectly declare a site to be unsuitable habitat for a species when in fact, the species was simply not detected at that site. In order to address this issue, MacKenzie et al. [17] proposed a model to estimate site occupancy rates when the detec-

Table 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.
$B(Y_{it})$	The birder associated with checklist Y_{it}.
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.
$Logistic$	Logistic function $P(t) = \frac{1}{1+exp^{-t}}$.

tion probabilities are less than one and when multiple visits are made to each site. MacKenzie's model did not allow for false detections, which result from incorrectly declaring a species to be present at a site when in fact the site is unoccupied by that species. Royle and Link [20] further extended MacKenzie's site occupancy model by using a finite mixture model to account for false detections. Royle and Link concluded that false detections can induce extreme bias in estimates of site occupancy when they are not accounted for. Hutchinson and Dietterich [10] explore an EM approach to estimating parameters in site occupancy models; we adopt this EM approach in our paper.

3 METHODOLOGY

In this section, we first describe the OD model [17, 16]. 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.

3.1 The Occupancy-Detection Model

Figure 1 illustrates the OD model for a single species as a graphical model [13], 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 1, the true site occupancy at site i (Z_i) is latent. The dotted boxes in Figure 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 and the inner plate represents the number of visits T_i to the ith site. A summary of the random variables used in the OD model are given in Table 1.

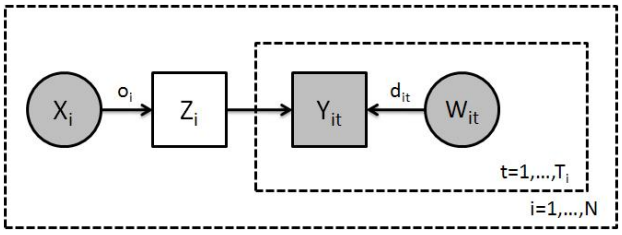

Figure 1: Graphical model representation of the Occupancy-Detection model for a single bird species.

The occupancy component of OD model models the occupancy status of the site (ie. 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 1.

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

The detection component captures the conditional probability of the observer detecting the species (ie. random variable Y_{it}), during a visit at site i and at time t conditioned on the site being occupied ie. $Z_i = 1$ 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 probability, the true detection probability d_{it} at site i and time t is modeled as a logistic function (parameterized by $\boldsymbol{\beta}$) of the detection features as shown in Equation 2.

$$\begin{aligned} d_{it} &= Logistic(\boldsymbol{W}_{it}; \boldsymbol{\beta}) \\ P(Y_{it}|Z_i, \boldsymbol{W}_{it}, \boldsymbol{\beta}) &= (Z_i d_{it})^{Y_{it}}(1 - Z_i d_{it})^{1-Y_{it}} \end{aligned} \tag{2}$$

Under the OD model, sites are visited multiple times and observations are made during each visit. The site detection history includes the observed presence

Table 2: Notation Used 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.
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.

or absence of the species on each visit at this site. The OD model makes two key assumptions. First, the population closure assumption [17] 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 false detections occur when observers incorrectly declare a species to be present at a site when the site is in fact unoccupied by that species. 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. Note that false absences, which occur when observers erroneously report the absence of a species when the site is in fact occupied by that species, are allowed by the OD model. False absences could be due to species that are hard to detect (eg. 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.

3.2 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. A graphical model representation of the ODE model for a single bird species is shown in Figure 2. A summary of the random variables and parameters used in the ODE model is given in Table 2.

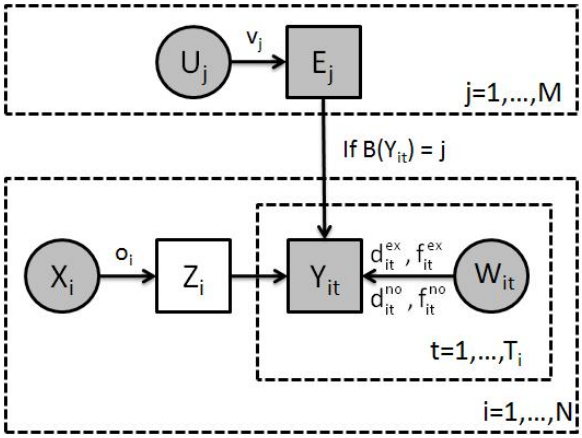

Figure 2: Graphical model representation of Occupancy-Detection-Expertise Model for a single bird species.

In the expertise component, E_j is a binary random variable capturing the expertise (ie. 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. We use the logistic function with parameters $\boldsymbol{\gamma}$ to model the expertise component as follows:

$$\begin{aligned} v_j &= Logistic(\boldsymbol{U}_j; \boldsymbol{\gamma}) \\ P(E_j|\boldsymbol{U}_j, \boldsymbol{\gamma}) &= v_j^{E_j}(1-v_j)^{1-E_j} \end{aligned} \tag{3}$$

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 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 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 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:

$$
\begin{aligned}
d_{it}^{ex} =& Logistic(\boldsymbol{W}_{it}; \boldsymbol{\beta}_1^{ex}) \\
\tilde{P_1^{ex}}(Y_{it}) =& (d_{it}^{ex})^{Y_{it}}(1 - d_{it}^{ex})^{1-Y_{it}} \qquad (4) \\
f_{it}^{ex} =& Logistic(\boldsymbol{W}_{it}; \boldsymbol{\beta}_0^{ex}) \\
\tilde{P_0^{ex}}(Y_{it}) =& (f_{it}^{ex})^{Y_{it}}(1 - f_{it}^{ex})^{1-Y_{it}} \qquad (5) \\
d_{it}^{no} =& Logistic(\boldsymbol{W}_{it}; \boldsymbol{\beta}_1^{no}) \\
\tilde{P_1^{no}}(Y_{it}) =& (d_{it}^{no})^{Y_{it}}(1 - d_{it}^{no})^{1-Y_{it}} \qquad (6) \\
f_{it}^{no} =& Logistic(\boldsymbol{W}_{it}; \boldsymbol{\beta}_0^{no}) \\
\tilde{P_0^{no}}(Y_{it}) =& (f_{it}^{no})^{Y_{it}}(1 - f_{it}^{no})^{1-Y_{it}} \qquad (7)
\end{aligned}
$$

$$
\begin{aligned}
P(Y_{it}|Z_i, \boldsymbol{W}_{it}, E_{B(Y_{it})}, \boldsymbol{\beta}) =& E_{B(Y_{it})}[Z_i \tilde{P_1^{ex}}(Y_{it}) + (1 - Z_i)\tilde{P_0^{ex}}(Y_{it})] + \\
& (1 - E_{B(Y_{it})})[Z_i \tilde{P_1^{no}}(Y_{it}) + (1 - Z_i)\tilde{P_0^{no}}(Y_{it})] \qquad (8)
\end{aligned}
$$

3.3 Parameter Estimation and Regularization

The ODE model requires a labeled set of expert and novice birders to estimate the model parameters using Expectation Maximization [5]. The EM algorithm maximizes the expected joint log-likelihood shown in Equation 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 9 below.

$$
\begin{aligned}
Q =& \mathbb{E}_{P(\boldsymbol{Z}|\boldsymbol{Y},\boldsymbol{E})}[\log P(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{E} | \boldsymbol{X}, \boldsymbol{U}, \boldsymbol{W})] \\
=& \mathbb{E}_{P(\boldsymbol{Z}|\boldsymbol{Y},\boldsymbol{E})}\Bigg[\sum_{j=1}^{M} \log P(E_j | \boldsymbol{U}_j, \boldsymbol{\gamma}) + \\
& \sum_{i=1}^{N} \Bigg[\log P(Z_i | \boldsymbol{X}_i, \boldsymbol{\alpha}) \sum_{t=1}^{T_i} \log P(Y_{it} | Z_i, \boldsymbol{W}_{it}, E_{B(Y_{it})}, \boldsymbol{\beta})\Bigg]\Bigg] \\
=& \mathbb{E}_{P(\boldsymbol{Z}|\boldsymbol{Y},\boldsymbol{E})}\Bigg\{\sum_{j=1}^{M} [E_j \log(v_j) + (1 - E_j)\log(1 - v_j)] + \\
& \sum_{i=1}^{N} \Bigg[Z_i \log(o_i) + (1 - Z_i)\log(1 - o_i) + \\
& \sum_{t=1}^{T_i} \log \Big[E_{B(Y_{it})}[Z_i P_1^{\tilde{ex}}(Y_{it}) + (1 - Z_i) P_0^{\tilde{ex}}(Y_{it})] + \\
& (1 - E_{B(Y_{it})})[Z_i P_1^{\tilde{no}}(Y_{it}) + (1 - Z_i) P_0^{\tilde{no}}(Y_{it})]\Big]\Bigg]\Bigg\} \\
=& \sum_{j=1}^{M} \Big[E_j \log(v_j) + (1 - E_j)\log(1 - v_j)\Big] + \\
& \sum_{i=1}^{N} \Bigg\{\tilde{P}(Z_i = 1)\Bigg[\log(o_i) + \sum_{t=1}^{T_i} \log\Big(E_{B(Y_{it})} P_1^{\tilde{ex}}(Y_{it}) \\
& + (1 - E_{B(Y_{it})}) P_1^{\tilde{no}}(Y_{it})\Big)\Bigg] + \\
& \tilde{P}(Z_i = 0)\Bigg[\log(1 - o_i) + \sum_{t=1}^{T_i} \log\Big(E_{B(Y_{it})} P_0^{\tilde{ex}}(Y_{it}) \\
& + (1 - E_{B(Y_{it})}) P_0^{\tilde{no}}(Y_{it})\Big)\Bigg]\Bigg\}
\end{aligned}
\tag{9}
$$

There are three regularization parameters $(\lambda_o, \lambda_d, \lambda_e)$ corresponding to the occupancy, detection and expertise components of the ODE model. We regu-

larize these three components using the penalty term in Equation 10.

$$r(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \lambda_o \frac{1}{2} \sum_{i=2}^{|\alpha|} \alpha_i^2 + \lambda_d \frac{1}{2} \sum_{i=2}^{|\beta|} \beta_i^2 + \lambda_e \frac{1}{2} \sum_{i=2}^{|\gamma|} \gamma_i^2 \tag{10}$$

In the M-step, EM determines the values of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ that maximize Equation 9. Using the gradients in Equations 11 - 13, we apply L-BFGS [15] to perform the optimization. Equation 12 is representative of the other parameters $\boldsymbol{\beta}_0^{ex}$, $\boldsymbol{\beta}_1^{no}$, and $\boldsymbol{\beta}_0^{no}$.

$$\begin{aligned}
\frac{\partial Q}{\partial \boldsymbol{\alpha}} &= \sum_{i=1}^{N} \frac{\partial Q}{\partial o_i} \frac{\partial o_i}{\partial \boldsymbol{\alpha}} - \lambda_o \boldsymbol{\alpha}^{(t)} \\
&= \sum_{i=1}^{N} (\tilde{P}(Z_i = 1) - o_i) \boldsymbol{X}_i - \lambda_o \boldsymbol{\alpha}^{(t)} \qquad (11) \\
\frac{\partial Q}{\partial \boldsymbol{\beta}_1^{ex}} &= \sum_{i=1}^{N} \sum_{t=1}^{T_i} \frac{\partial Q}{\partial \boldsymbol{\beta}_{it}^{ex}} \frac{\partial \boldsymbol{\beta}_{it}^{ex}}{\partial \boldsymbol{\beta}_1^{ex}} - \lambda_d \boldsymbol{\beta}_1^{ex(t)} \\
&= \sum_{i=1}^{N} \sum_{t=1}^{T_i} \tilde{P}(Z_i = 1) \frac{E_{B(Y_{it})} \tilde{P_1^{ex}}(Y_{it})(Y_{it} - d_{it}^{ex}) \boldsymbol{W}_{it}}{E_{B(Y_{it})} \tilde{P_1^{ex}}(Y_{it}) + (1 - E_{B(Y_{it})}) \tilde{P_1^{no}}(Y_{it})} - \lambda_d \boldsymbol{\beta}_1^{ex(t)} \qquad (12) \\
\frac{\partial Q}{\partial \boldsymbol{\gamma}} &= \sum_{j=1}^{M} \frac{\partial Q}{\partial v_j} \frac{\partial v_j}{\partial \boldsymbol{\gamma}} - \lambda_e \boldsymbol{\gamma}^{(t)} \\
&= \sum_{j=1}^{M} (E_j - v_j) \boldsymbol{U}_j - \lambda_e \boldsymbol{\gamma}^{(t)} \qquad (13)
\end{aligned}$$

An *identifiability* problem [20] 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.

3.4 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.

3.4.1 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 14.

$$
\begin{aligned}
&P(Z_i = 1 | \boldsymbol{X}_i, \boldsymbol{Y}_{i\cdot}, \boldsymbol{W}_{i\cdot}, \boldsymbol{U}) \\
&= \frac{P(\boldsymbol{Y}_{i\cdot}, Z_i = 1 | \boldsymbol{X}_i, \boldsymbol{W}_{i\cdot}, \boldsymbol{U})}{\sum_{z_i \in \{0,1\}} P(\boldsymbol{Y}_{i\cdot}, Z_i = z_i | \boldsymbol{X}_i, \boldsymbol{W}_{i\cdot}, \boldsymbol{U})} \\
&= \frac{\sum_{\boldsymbol{e}^i} P(\boldsymbol{Y}_{i\cdot}, Z_i = 1, \boldsymbol{E}^i = \boldsymbol{e}^i | \boldsymbol{X}_i, \boldsymbol{W}_{i\cdot}, \boldsymbol{U})}{\sum_{z_i \in \{0,1\}} \sum_{\boldsymbol{e}^i} P(\boldsymbol{Y}_{i\cdot}, Z_i = z_i, \boldsymbol{E}^i = \boldsymbol{e}^i | \boldsymbol{X}_i, \boldsymbol{W}_{i\cdot}, \boldsymbol{U})}
\end{aligned} \tag{14}
$$

where

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

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 (ie. 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 4.2.

3.4.2 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 15.

$$P(Y_{it} = 1|\boldsymbol{X}_i, \boldsymbol{W}_{it}, \boldsymbol{U}_{B(Y_{it})}) = \sum_{z_i \in \{0,1\}} \sum_{e \in \{0,1\}} P(Y_{it} = 1, Z_i = z_i, E_{B(Y_{it})} = e|\boldsymbol{X}_i, \boldsymbol{W}_{it}, \boldsymbol{U}_{B(Y_{it})}) \quad (15)$$

where

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

3.4.3 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 Y_{it}^j and $\boldsymbol{Y}_{i\cdot}^j$ extending our previous notation), let $\boldsymbol{W}_{it}^j$ be the detection features for 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 16.

$$\begin{aligned} P(E_j = 1|\boldsymbol{X}, \boldsymbol{Y}^j, \boldsymbol{W}, \boldsymbol{U}_j) &= \frac{P(E_j = 1, \boldsymbol{Y}^j|\boldsymbol{X}, \boldsymbol{W}, \boldsymbol{U}_j)}{\sum_{e_j \in \{0,1\}} P(E_j = e_j, \boldsymbol{Y}^j|\boldsymbol{X}, \boldsymbol{W}, \boldsymbol{U}_j)} \\ &= \frac{\sum_{\boldsymbol{z}^j} P(E_j = 1, \boldsymbol{Y}^j, \boldsymbol{Z}^j = \boldsymbol{z}^j|\boldsymbol{X}, \boldsymbol{W}, \boldsymbol{U}_j)}{\sum_{e_j \in \{0,1\}} \sum_{\boldsymbol{z}^j} P(E_j = e_j, \boldsymbol{Y}^j, \boldsymbol{Z}^j = \boldsymbol{z}^j|\boldsymbol{X}, \boldsymbol{W}, \boldsymbol{U}_j)} \end{aligned} \quad (16)$$

where

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

4 EVALUATION

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

Table 3: Occupancy, Detection and Expertise Features in the 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.

include results from a contrast mining task that illustrates the utility of the ODE model.

4.1 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 identified. Table 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 [18].

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

Table 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

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 knowledge of birder 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 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 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.

4.2 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 (LR) 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 5. Boldface indicates the best results. The $\star$ and † symbols indicate that the ODE model is a statistically significant improvement (paired t-test, $\alpha = 0.05$) over the LR and OD models respectively.

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. LR Model: 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 LR model. LR is a special case of a GLM, which is a common class of methods used for SDM by ecologists [1]. 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 both LR models using an L2-norm regularization term. As before, 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

Table 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***†

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 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 an 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.

$$\begin{aligned} P(Y_{it} = 1|\boldsymbol{X}_i, \boldsymbol{W}_{it}, \boldsymbol{\alpha}, \boldsymbol{\beta}) &= \sum_{z_i \in \{0,1\}} P(Z_i = z_i|\boldsymbol{X}_i, \boldsymbol{\alpha}) P(Y_{it} = 1|Z_i = z_i, \boldsymbol{W}_{it}, \boldsymbol{\beta}) \\ &= P(Z_i = 1|\boldsymbol{X}_i, \boldsymbol{\alpha}) P(Y_{it} = 1|Z_i = 1, \boldsymbol{W}_{it}, \boldsymbol{\beta}) \end{aligned} \tag{17}$$

3. ODE Model: The ODE model is trained using EM and the prediction of the detection random variable Y_{it} is based on Equation 15. The birder expertise is observed during training but unobserved during testing.

4.3 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. LR Model: 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. We use an 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 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 6 we tabulate the mean AUC for each species, with boldface entries indicating the best model and $\star$ indicating that the ODE model is a statistically significant improvement (paired t-test, $\alpha = 0.05$) over LR.

4.4 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 4.1 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 7.

Table 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***

Table 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

5 DISCUSSION

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 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 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.

When predicting expertise, the ODE model outperforms LR on all species except for *White-breasted Nuthatch* as shown in Table 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 then. 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.

In the contrast mining task, the results in Table 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 Δ_{TD} values are much larger. These results show that the ODE model is a promising approach for contrast mining, which can identify differences in how experts and novices report bird species.

6 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.

For future work, we would like we would like to replace the logistic regression parts of the ODE model with more flexible function approximators such as boosted trees [8] which allow non-linear interactions between the features. In addition, we would like to develop a semi-supervised approach to training the ODE model since a large amount of unlabeled data is available.

Acknowledgements

The authors would like to thank Marshall Iliff, Brian Sullivan, Chris Wood and Steve Kelling for their help with this paper, particularly with manually labelling the expertise of birders in the training data. This work was supported by NSF grant CCF 0832804.

References

[1] M. P. Austin. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecol. Modell.*, 157:101–118, 2002.

[2] G. Carpenter, A. N. Gillison, and J. Winter. Domain: a flexible modelling procedure for mapping potential distributions of plants and animals. *Biodiversity and Conservation*, 2:667–680, 1993.

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

[4] 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.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, 39(1):1–38, 1977.

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

[7] 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.

[8] J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001.

[9] J. H. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. *Ann. Stat.*, 28:337–407, 2000.

[10] R. A. Hutchinson and T. G. Dietterich. Parameter estimation in a hierarchical model for species occupancy. In *Neural Information Processing Systems (NIPS) Workshops: The Generative and Discriminative Learning Interface*, 2009.

[11] M. Kéry, B. Gardner, and C. Monnerat. Predicting species distributions from checklist data using site-occupancy models. *Journal of Biogeography*, 37:1851–1862, 2010.

[12] M. Kéry, J. A. Royle, H. Schmid, M. Schaub, B. Volet, G. Häfliger, and N. Zbinden. Site-occupancy distribution modeling to correct population-trend estimates derived from opportunistic observations. *Conservation Biology*, 24(5):1388–1397, 2009.

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

[14] A. M. Latimer, S. Wu, A. E. Gelfand, and J. John A. Silander. Building statistical models to analyze species distributions. *Ecological Applications*, 16(1):33–50, 2006.

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

[16] 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.

[17] 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.

[18] 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.

[19] S. J. Phillips, M. Dudik, and R. E. Schapire. A maximum entropy approach to species distribution modeling. In *Proceedings of the 21st International Conference on Machine Learning*, pages 83–91, 2004.

[20] 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.

[21] D. Stockwell and D. Peters. The garp modelling system: problems and solutions to automated spatial prediction. *Int. J. Geogr. Inform. Sci.*, 13:143–158, 1999.