A Latent Variable Model for Discovering Bird Species Commonly Misidentified by Citizen Scientists

Jun Yu and **Rebecca A. Hutchinson** and **Weng-Keen Wong**
Department of EECS
Oregon State University
{yuju, rah, wong}@eecs.orst.edu

Abstract

Data quality is a common source of concern for large-scale citizen science projects like eBird. In the case of eBird, a major cause of poor quality data is the misidentification of bird species by inexperienced contributors. A proactive approach for improving data quality is to discover commonly misidentified bird species and to teach inexperienced birders the differences between these species. To accomplish this goal, we develop a latent variable graphical model that can identify groups of bird species that are often confused for each other by eBird participants. Our model is a multi-species extension of the classic occupancy-detection model in the ecology literature. This multi-species extension requires a structure learning step as well as a computationally expensive parameter learning stage which we make efficient through a variational approximation. We show that our model can not only discover groups of misidentified species, but by including these misidentifications in the model, it can also achieve more accurate predictions of both species occupancy and detection.

Introduction

Species distribution models (SDMs) estimate the pattern of species occurrence on a landscape by correlating observations of the species with environmental features. SDMs play an important role in modeling biodiversity and designing wildlife reserves (Leathwick et al. 2008). 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 the general public is encouraged to contribute data to scientific studies (Cohn 2008). For example, eBird (Sullivan et al. 2009; Kelling et al. 2013) is one of the largest citizen science projects in existence, relying on a global human sensor network of bird-watchers to report their observations of birds, identified by species, to a centralized database.

Although citizen scientists can contribute large quantities of data, data quality can be a concern (Hochachka et al. 2012). 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, inexperienced birders often confuse house finches with purple finches, which are similar in appearance, but occupy different habitats. One way to reduce noise is to identify and remove the invalid observations using a data verification model (Yu et al. 2012). A more proactive way is to discover which species are often confused for each other and to teach inexperienced observers to correctly identify species them.

To discover groups of misidentified species, we extend the classic single species *Occupancy-Detection* (OD) model (MacKenzie et al. 2002) from the ecology literature to handle multiple species simultaneously. The OD model is an SDM that separates the biological process of *occupancy*, which is a latent variable describing whether a species lives at a site, from the observational process of *detection*, which describes whether a species will be observed at a site it occupies. Separating occupancy from detection allows the OD model to account for false negatives, which are common in species data since many species are secretive and hard to detect on surveys. The OD model was also developed under the assumption that data were collected by expert field biologists and thus assumes that there are no false positives in the data. Citizen science data, however, is collected less rigorously, making this assumption questionable. Previous work has incorporated the possibility of false positives into the OD model (Royle and Link 2006). More recent work has modeled false positives in the citizen science context by distinguishing between experts and novices in the detection process (Yu, Wong, and Hutchinson 2010).

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 for a species as arising from misidentifications of other species. Modeling occupancy and detection patterns for multiple species jointly has two important advantages. Firstly, the patterns of species confusion that are discovered can be used to teach inexperienced observers and improve their skills. Secondly, explicitly modeling detection errors due to observer misidentification between species can improve the estimates of the occupancy patterns of these species. Since the latent occupancy is the true variable of interest, improvements in our ability to account for the detection process allow for more accurate ecological conclusions to be drawn. In our study, we show that explicitly modeling

observer confusion between species not only helps to discover groups of misidentified species, but also improves the estimates of the occupancy patterns of those species.

Occupancy-Detection model

In species distribution modeling, the primary goal is to estimate a habitat model for the species of interest, but the true occupancy status of the study sites is typically observed only indirectly. Figure 1 shows a plate diagram of the single-species *Occupancy-Detection* (OD) model, proposed by MacKenzie et al. (MacKenzie et al. 2002) to separate the detection process from occupancy. 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, while its occupancy status remains constant. 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.

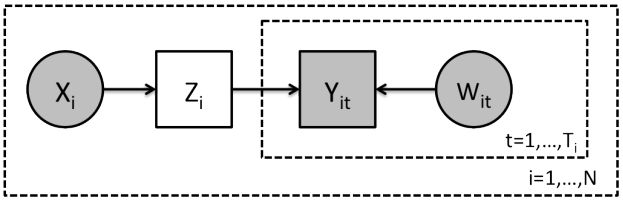


Figure 1: The Occupancy-Detection model.

The structure of the OD model corresponds to the following generative process. For each site i, we compute the probability o_i that site i is occupied as $o_i = \sigma(\boldsymbol{X}_i \cdot \boldsymbol{\alpha})$, where $\sigma(\cdot)$ is the logistic function. Then, the true occupancy Z_i is generated by drawing from a Bernoulli distribution with parameter o_i. Next, the site is visited T_i times. At each visit t, we compute the detection probability $d_{it} = \sigma(\boldsymbol{W}_{it} \cdot \boldsymbol{\beta})$. Finally, the observation Y_{it} is generated by drawing from a Bernoulli distribution with parameter $Z_i d_{it}$. If the site is not occupied ($Z_i = 0$), then $Y_{it} = 0$ with probability 1, but if $Z_i = 1$, then $Y_{it} = 1$ with probability d_{it}. This encodes the assumption that there are *no false positives* in the data.

Multi-Species Occupancy-Detection model

The *Multi-Species Occupancy-Detection* (MSOD) model consists of observed ($\boldsymbol{Y}$) and latent binary variables ($\boldsymbol{Z}$) for every species as shown using plate notation in Figure 2. Z_{is} denotes the occupancy status of species s at site i and Y_{its} denotes the observation of species s at site i on visit t. 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. The joint probability distribution for the MSOD model is given below where $\boldsymbol{Z}_{i\cdot}$ refers

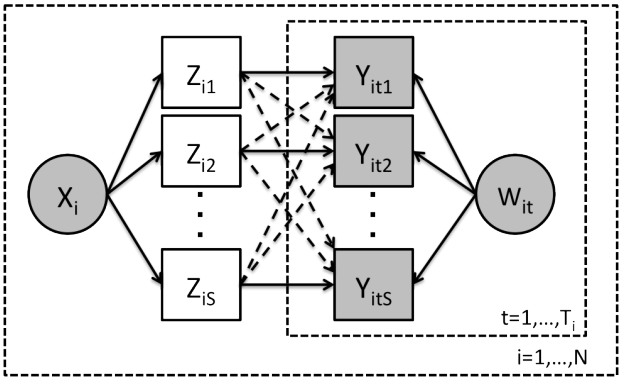


Figure 2: The Multi-Species Occupancy-Detection model.

to all the S latent occupancy variables at site i.

$$\ell = \prod_{i=1}^{N} P(\boldsymbol{Y}_{i\cdot\cdot}, \boldsymbol{Z}_{i\cdot} | \boldsymbol{X}_i, \boldsymbol{W}_{i\cdot})$$
$$= \prod_{i=1}^{N} \left[\prod_{r=1}^{S} P(Z_{ir} | \boldsymbol{X}_i) \right] \left[\prod_{t=1}^{T_i} \prod_{s=1}^{S} P(Y_{its} | \boldsymbol{Z}_{i\cdot}, \boldsymbol{W}_{it}) \right]$$

Parameterization

In the MSOD model, the species-specific occupancy models ($P(Z_{ir}|\boldsymbol{X}_i)$ for each r) are parameterized as in the OD model, where $Z_{ir} \sim Bernoulli(o_{ir})$ and $o_{ir} = \sigma(\boldsymbol{X}_i \cdot \boldsymbol{\alpha}_r)$. 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 that may be confused for species s. We model the detection probability based on a noisy-or parameterization (Heckerman 1989; Shwe et al. 1991). 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, $d_{itrs} = P(Y_{its} = 1|Z_{ir} = 1) = \sigma(\boldsymbol{W}_{it} \cdot \boldsymbol{\beta}_{rs})$. 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. In contrast, the probability of species s being reported during visit t at site i cannot be fully factorized, as shown below; in this case, 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.

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

Structure learning and parameter estimation

During training, we learn both the graph structure and the model parameters ($\{\boldsymbol{\alpha}, \boldsymbol{\beta}\}$). We start the learning by assuming that the bipartite graph between $\boldsymbol{Y}$ and $\boldsymbol{Z}$ is fully connected and then estimate the MSOD model parameters using Expectation Maximization (Dempster, Laird, and Rubin 1977). If we know that certain species are not easily confused for each other, we can incorporate this by initializing the model structure with no cross edge between those

species. In the E-step, EM computes the expected occupancies $\boldsymbol{Z}_{i\cdot}$ for every site i using Bayes rule. In the M-step, since there is no closed-form solution, we use L-BFGS (Liu and Nocedal 1989) to re-estimate the model parameters $\{\boldsymbol{\alpha}, \boldsymbol{\beta}\}$ that maximize the expected log-likelihood in Equation 1. We use random restarts to avoid getting trapped in local optima. We also add a constraint to the objective function during training that 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 misidentified for species s. Without this constraint, spurious cross-edges will be added to the model to account for the detection of species s.

$$\begin{aligned}\mathcal{Q} &= E_{\boldsymbol{Z}|\boldsymbol{Y},\boldsymbol{X},\boldsymbol{W}}[\log(P(\boldsymbol{Y},\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{W}))] \\ &= \sum_{i=1}^{N}\Big[\sum_{r=1}^{S} E_{Z_{ir}|\boldsymbol{Y}_{i\cdot\cdot},\boldsymbol{X}_i,\boldsymbol{W}_{i\cdot}}[\log(P(Z_{ir}|\boldsymbol{X}_i))] + \\ &\quad \sum_{t=1}^{T_i}\sum_{s=1}^{S} E_{\boldsymbol{Z}_{i\cdot}|\boldsymbol{Y}_{i\cdot\cdot},\boldsymbol{X}_i,\boldsymbol{W}_{i\cdot}}[\log(P(Y_{its}|\boldsymbol{Z}_{i\cdot},\boldsymbol{W}_{it}))]\Big]\end{aligned} \tag{1}$$

After learning the model parameters for the model with the fully connected structure, we refine the model structure using greedy search. More specifically, we sort all the cross edges (i.e. pairs of misidentified species) by their averaged detection probability ($\bar{d}_{\cdot\cdot rs}$) on the training data and then greedily add cross edges according to this metric until the log-likelihood on a holdout validation set does not improve. Once we determine the structure, we retrain the MSOD model with a fixed structure and estimate the model parameters. In addition, we initialize the leak probability of each species in the MSOD model to the value of the leak probability learned by modeling each species individually with a modified single-species OD model that contains a learned leak probability (called the ODLP model in our experiments). Furthermore, we use the L_2 penalty to regularize the model parameters (λ_o for occupancy parameters and λ_d for detection parameters).

Inference

The MSOD model can be used to predict the site occupancy of a specific species s (i.e. Z_{is}) or a set of species, and predict the observations of species s (i.e. Y_{its}) on a checklist. The occupancy probability of site i can be computed using the following equation where $\boldsymbol{Z}_{i\neg s}$ denote the occupancy variables of all species except for species s at site i.

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

Since the true site occupancy is typically unavailable for evaluation on real-world field datasets, we evaluate SDMs based on how well they predict the observation of a species at a site. The probability of detecting species s at site i on visit t can be computed as follows where $\boldsymbol{\pi}_s$ denotes the set of species that can be misidentified as species s.

$$\begin{aligned}&P(Y_{its}=1|\boldsymbol{X}_i,\boldsymbol{W}_{it}) \\ &= \sum_{\boldsymbol{z}_{i\boldsymbol{\pi}_s}} P(\boldsymbol{Y}_{its}=1, \boldsymbol{Z}_{i\boldsymbol{\pi}_s}=\boldsymbol{z}_{i\boldsymbol{\pi}_s}|\boldsymbol{X}_i,\boldsymbol{W}_{it})\end{aligned}$$

Variational Learning

The computation of expectations in Equation 1 is expensive with large values of S since it requires summing over the configurations of S binary variables, resulting in 2^S terms. To speed up the learning, we use the variational learning to reduce the computational cost. The key observation is that the inference is intractable because the detection probability $P(Y_{its}|\boldsymbol{Z}_{i\cdot},\boldsymbol{W}_{it})$ cannot be factorized when $Y_{its}=1$. Therefore, we use a fully factorized lower bound to approximate this detection probability based on Jensen's inequality (Jaakkola and Jordan 1999). We introduce the variational parameters q_{itrs} where $q_{it\cdot s}$ defines a multinomial distribution (i.e. $\sum_{r=1}^{S} q_{itrs}=1$) specifying the importance of each species for detecting species s on visit t at site i.

$$\begin{aligned}&\log P(Y_{its}=1|\boldsymbol{Z}_{i\cdot},\boldsymbol{W}_{it}) \\ &= \log\Big(1-\exp\big(-\theta_{0s}-\sum_{r=1}^{S} Z_{ir}\theta_{itrs}\big)\Big) \\ &= f(\theta_{0s}+\sum_{r=1}^{S} Z_{ir}\theta_{itrs}) \\ &\geq \sum_{r=1}^{S} q_{itrs} f\left(\theta_{0s}+\frac{Z_{ir}\theta_{itrs}}{q_{itrs}}\right)\end{aligned}$$

where $\theta_{0s}=-\log(1-d_{0s})$, $\theta_{itrs}=-\log(1-d_{itrs})$, $f(x)=\log(1-\exp(-x))$ is a concave function. Now the detection probability is in the fully factorized form.

During learning, instead of maximizing the expected log-likelihood in Equation 1, we maximize its lower bound approximation $\mathcal{Q}^\star$ by plugging in the lower bound of the detection probability $P(Y_{its}=1|\boldsymbol{Z}_{i\cdot},\boldsymbol{W}_{it})$ (Singliar and Hauskrecht 2006). The variational EM iterates between updating the variational parameters q_{itrs} and the expected occupancy Z_{ir} in the E-step and optimizing the model parameter $\{\boldsymbol{\alpha}, \boldsymbol{\beta}\}$ in the M-step until it converges.

$$\begin{aligned}\mathcal{Q} &\geq \mathcal{Q}^\star \\ &= \sum_{i=1}^{N}\sum_{r=1}^{S}\Big[E_{Z_{ir}|\boldsymbol{Y}_{i\cdot\cdot},\boldsymbol{X}_i,\boldsymbol{W}_{i\cdot}}[\log(P(Z_{ir}|\boldsymbol{X}_i))] + \\ &\quad \sum_{t=1}^{T_i}\sum_{s=1}^{S} E_{Z_{ir}|\boldsymbol{Y}_{i\cdot\cdot},\boldsymbol{X}_i,\boldsymbol{W}_{i\cdot}}[\log(P(Y_{its}|Z_{ir},\boldsymbol{W}_{it}))]\Big]\end{aligned}$$

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.

Synthetic dataset

For the synthetic data experiments, data is generated for 1000 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. The occupancy and detection processes in this data are linear functions of their respective covariates. A true structure over 5 species is generated by randomly adding 7 cross-edges (in addition to the five 'straight' or 'self' edges). 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 the detection probability of a species s due to the presence of another species confused for s to be smaller than the detection probability due to the presence of the species s itself. Training, validation, and test datasets 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 results.

To test the robustness of the MSOD model, we also generate data from models that differ from the assumptions of the MSOD model. First, we generate synthetic data with interactions between species occupancies, simulating 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; the same behavior occurs 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% at a site when species 3 occupies that site; we truncate the occupancy probability at 1 when it goes beyond 1. We denote this synthetic data with occupancy interactions "Syn-I" in the results. The second alternative data generation process is denoted "Syn-NL." In this setting, 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 $X'_{i4} = X_{i4}$. In the last data generation scenario, we make the synthetic data the most challenging by adding both species occupancy interactions and non-linear occupancy components (denoted "Syn-I-NL").

We compare the standard OD model against the exact inference (MSOD-E) and the variational inference (MSOD-V) MSOD models in terms of predicting occupancy (Z) and observation (Y). In addition, we include results for a variant of the OD model called *ODLP*, which includes a learned leak probability in the OD model, and the ground truth model that generated the data (called *TRUE*). We tune the regularization terms of the occupancy (λ_o) and detection (λ_d) processes in the OD and ODLP models over the set of values $\{0.01, 0.1, 1, 10\}$ based on the performance of the occupancy prediction on a holdout dataset. Instead of tuning the regularization terms of every species in the MSOD model separately, we run a less time-consuming pre-processing step in which we fit an OD model to each species individually and set the regularization terms in the MSOD model to the best value found by the OD model of that species.

Table 1: The AUC (and the standard errors) of occupancy and detection prediction averaged over 30 datasets in the 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 using the paired t-test.

	Syn dataset	
Model	**Occupancy (Z)**	**Observation (Y)**
TRUE	0.941 ± 0.004	0.783 ± 0.004
OD	0.849 ± 0.006	0.751 ± 0.005
ODLP	0.868 ± 0.006	0.752 ± 0.005
MSOD-E	$\mathbf{0.937 \pm 0.005}^{\star\dagger}$	$\mathbf{0.777 \pm 0.004}^{\star\dagger}$
MSOD-V	$0.908 \pm 0.007^{\star\dagger}$	$0.768 \pm 0.005^{\star\dagger}$
	Syn-I dataset	
Model	**Occupancy (Z)**	**Observation (Y)**
TRUE	0.943 ± 0.003	0.776 ± 0.003
OD	0.842 ± 0.005	0.744 ± 0.004
ODLP	0.865 ± 0.005	0.746 ± 0.004
MSOD-E	$\mathbf{0.928 \pm 0.004}^{\star\dagger}$	$\mathbf{0.766 \pm 0.004}^{\star\dagger}$
MSOD-V	$0.899 \pm 0.008^{\star\dagger}$	$0.757 \pm 0.005^{\star\dagger}$
	Syn-NL dataset	
Model	**Occupancy (Z)**	**Observation (Y)**
TRUE	0.937 ± 0.003	0.777 ± 0.005
OD	0.837 ± 0.007	0.739 ± 0.005
ODLP	0.848 ± 0.007	0.741 ± 0.005
MSOD-E	$\mathbf{0.907 \pm 0.006}^{\star\dagger}$	$\mathbf{0.755 \pm 0.004}^{\star\dagger}$
MSOD-V	$0.874 \pm 0.008^{\star\dagger}$	0.748 ± 0.007
	Syn-I-NL dataset	
Model	**Occupancy (Z)**	**Observation (Y)**
TRUE	0.938 ± 0.003	0.768 ± 0.003
OD	0.832 ± 0.003	0.731 ± 0.005
ODLP	0.841 ± 0.006	0.732 ± 0.004
MSOD-E	$\mathbf{0.897 \pm 0.006}^{\star\dagger}$	$\mathbf{0.739 \pm 0.005}^{\star\dagger}$
MSOD-V	$0.866 \pm 0.010^{\star\dagger}$	0.736 ± 0.004

In Table 1, we report the area under the ROC curve (AUC) averaged over 30 datasets; in each dataset, the AUC is 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 cannot accurately capture the noise from the detection process due to species misidentification. The performance of the MSOD model is closest to

the true model in predicting both occupancy and detection. As we allow species occupancy interactions and non-linear occupancy components in the data, the performance of the MSOD model decreases slightly, but it 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).

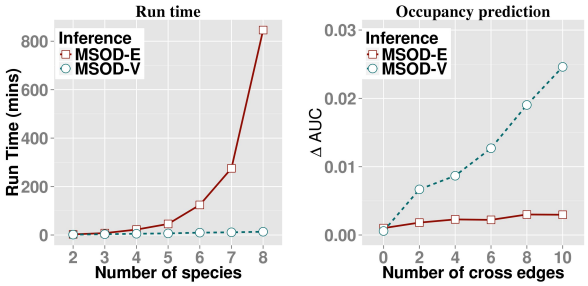
The OD and MSOD models differ greatly in their performance when predicting occupancy even though their performance when predicting detection is fairly close. This difference indicates that the values of the latent occupancy variables are indeed distinct from the values of the detection variables. Consequently, modeling occupancy as a separate latent process from detection is important, especially when other species can be mistakenly reported for the true species.

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 all the true cross edges over the other cross edges in the model. In Table 2, we report the structural AUC for the learned model structure on the four synthetic datasets. In the simplest case with the "Syn" dataset, the MSOD model achieves the structural AUC value of 0.989. As the synthetic data varies from the MSOD assumptions, 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 cross edges corresponding to species confusions in our synthetic datasets.

Syn	*Syn-I*	*Syn-NL*	*Syn-I-NL*
0.99 ± 0.01	0.98 ± 0.01	0.97 ± 0.01	0.97 ± 0.01

Table 2: The structural 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 structural AUC values are averaged over 30 datasets in each experiment.

Next, we compare the running time and prediction performance between the exact inference and the variational inference. First, we generate synthetic datasets for an increasing number of species S where $S \in [2, 8]$. For a dataset of S species, we randomly add S pairs of misidentified species and report the running time of one random restart of EM iteration in Figure 3 (a). As expected, the running time of the exact inference grows exponentially with the number of species, while the variational inference grows linearly. Then, we evaluate the prediction performance of the exact and variational inference with increasing complexity of the model structure. We generate synthetic datasets of 5 species and gradually increase the number of cross edges added from 0 to 10 with an increment of 2. We report the difference of occupancy prediction against the *true* latent model in Figure 3 (b). As the model structure gets more complex, the exact inference is very stable and robust, while the performance of the variational inference decreases slightly. The choice of exact or variational inference in the MSOD model will depend on application-specific trade-offs between running time and prediction performance.



(a) (b)

Figure 3: (a) The running time of the exact and variational inference with increasing number of species. (b) The difference of occupancy prediction against the true *latent* model (Δ AUC) with increasing complexity in the model structure. Both results are averaged over 10 datasets.

eBird dataset

We also test the ability of the MSOD model to discover species misidentifications on three case studies involving real-world eBird data, using species selected by experts at the Cornell Lab of Ornithology. We evaluated the MSOD model on subsets of eBird species that include pairs of species known to be confused for each other and a distractor species with minimal similarity to the others. The case studies include: Sharp-shinned Hawk and Cooper's Hawk (with Turkey Vulture as the distractor species), Hairy Woodpecker and Downy Woodpecker (with Dark-eyed Junco as the distractor species), and Purple Finch and House Finch (with Yellow-rumped Warbler as the distractor species).

In the experiment, we use data from California in the year 2010 since eBird participation in California is high. We group checklists (of which species were observed on a particular bird-watching event) 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 (e.g. population, housing density, housing vacancy, elevation and habitat class) and 10 detection features (e.g. time of day, season, observation duration and distance travelled). For more details on the eBird covariates, we refer the readers to the eBird Manual (Munson et al. 2009).

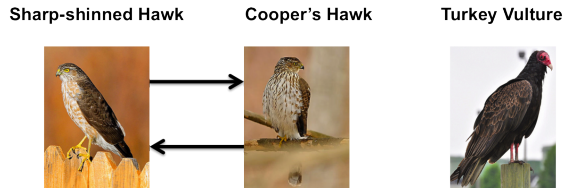
To alleviate the effect of spatial autocorrelation in creating training and test data, we superimpose a checkerboard over the data from California, with approximately 10 km x 10 km grid cells. If we "color" the checkerboard black and white, data points falling into the white cells are grouped together as the test set. Each black cell is further subdivided into a 2-by-2 subgrid so that data falling into the top left and bottom right subgrids form the training set and data falling into the top right and bottom left form the validation set.

Discovering species misidentifications To fit the MSOD model to eBird data, we first estimate the leak probability of each species 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 as described previously. We show the learned model structures in Figure 4. 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.

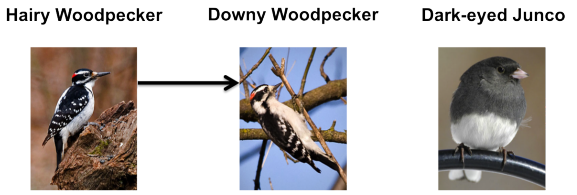
Predicting checklist observations Since ground truth on species occupancy is not available, we use the prediction of observations as a substitute. After learning the structure, we re-estimate the MSOD model using data in both the training and validation sets and predict the observations on checklists in the test set. We create 30 different train/test splits by randomly positioning the bottom left corner of the checkboard. Then we compare the MSOD model against the OD and ODLP models. In Table 3, we report the AUC and accuracy of predicting detections for the three case studies. The MSOD model results in statistically significant improvements in AUC on 6/6 species compared to the OD model and 5/6 species compared to the ODLP model.

Conclusion

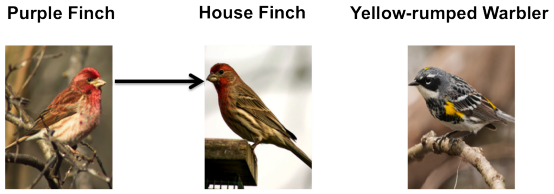
We introduce the Multi-Species Occupancy-Detection model to identify species misidentifications. Our experiments show that the model is not only capable of identifying groups of misidentified species but it also improves predictions of both species occupancy and detection. These results are promising for our goal of improving data quality for citizen science data by identifying difficult species for citizen scientists. Furthermore, the ability to accurately predict occupancy can improve species distribution models for conservation projects. For future work, we will investigate scaling up the model to even larger numbers of species and sites.



(a) Hawks case study



(b) Woodpeckers case study



(c) Finches case study

Figure 4: An arrow from species A to species B indicates that the MSOD model learns that the presence of species A affects the detection of species B.

Table 3: The AUC of observation prediction for three eBird case studies. Boldface results indicate the best model, $\star$ and $\dagger$ indicate the MSOD model is statistically better than the OD and ODLP model respectively using the paired t-test.

	Hawks case study	
Model	Sharp-shinned Hawk	Cooper's Hawk
OD	0.725 ± 0.005	0.765 ± 0.003
ODLP	0.737 ± 0.005	0.770 ± 0.005
MSOD	$\mathbf{0.757 \pm 0.003}^{\star\dagger}$	$\mathbf{0.780 \pm 0.002}^{\star\dagger}$

	Woodpeckers case study	
Model	Hairy Woodpecker	Downy Woodpecker
OD	0.833 ± 0.004	0.761 ± 0.004
ODLP	0.837 ± 0.004	0.769 ± 0.004
MSOD	$\mathbf{0.843 \pm 0.002}^{\star}$	$\mathbf{0.783 \pm 0.002}^{\star\dagger}$

	Finches case study	
Model	Purple Finch	House Finch
OD	0.807 ± 0.003	0.758 ± 0.003
ODLP	0.808 ± 0.003	0.762 ± 0.003
MSOD	$\mathbf{0.817 \pm 0.002}^{\star\dagger}$	$\mathbf{0.775 \pm 0.001}^{\star\dagger}$

References

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

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

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

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

Jaakkola, T. S., and Jordan, M. I. 1999. Variational probabilistic inference and the qmr-dt network. *Journal of Artificial Intelligence Research* 10:291–322.

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

Leathwick, J.; Moilanen, A.; Francis, M.; Elith, J.; Taylor, P.; Julian, K.; Hastie, T.; and Duffy, C. 2008. Novel methods for the design and evaluation of marine protected areas in offshore waters. *Conservation Letters* 1:91–102.

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

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

Munson, M. A.; Webb, K.; Sheldon, D.; Fink, D.; Hochachka, W. M.; Iliff, M.; Riedewald, M.; Sorokina, D.; Sullivan, B.; Wood, C.; and Kelling, S. 2009. The ebird reference dataset, version 1.0. Cornell Lab of Ornithology and National Audubon Society, Ithaca, NY.

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

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

Singliar, T., and Hauskrecht, M. 2006. Noisy-or component analysis and its application to link analysis. *Journal of Machine learning Research* 2189–2213.

Sullivan, B. L.; Wood, C. L.; Iliff, M. J.; Bonney, R.; Fink, D.; and Kelling, S. 2009. ebird: A citizen based bird observation network in the biological sciences. *Biological Conservation* 142(10):2282–2292.

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

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