Modeling Misidentification of Bird Species by Citizen Scientists

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

Rebecca A. Hutchinson
Department of EECS
Oregon State University
Corvallis, OR 97331
rah@eecs.orst.edu

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

Abstract

Data quality is a common source of concern with large-scale citizen science projects like eBird. In the case of eBird, poor quality data is often due to misidentification of bird species by inexperienced contributors. One approach for improving data quality is to identify commonly misidentified bird species and to teach inexperienced birders the differences between these species. In this paper, we develop a latent variable model, based on a multi-species extension of the classic occupancy-detection model in the ecology literature, that we can apply to eBird data to discover pairs of bird species that observers often confuse for each other.

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 [8, 11]. 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 [2]. For example, eBird [14, 7] 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, to a centralized database.

Although citizen scientists can contribute large quantities of data, data quality can be a concern [5, 15]. 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. One way to handle noise is to explicitly model the observer's expertise in SDMs [16]. A more proactive way to improve data quality is to improve the species identification skills of inexperienced observers and helping them correctly identify species that are commonly mistaken for each other.

To discover groups of misidentified species, we extend a well-known latent variable model in ecology, the *Occupancy-Detection* model, to the multiple species case. The multi-species OD model treats false positives for a species as arising from misidentifications of other species. We propose an approach to learn both the model structure (i.e. species confusions) and the parameters of the model from observational data. 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.

2 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. Mackenzie et al. [9] proposed a well-known site occupancy model in ecology, which we call the Occupancy-Detection (OD) model, that separates occupancy from detection. Figure 1 shows a plate diagram of the single-species OD model. 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.

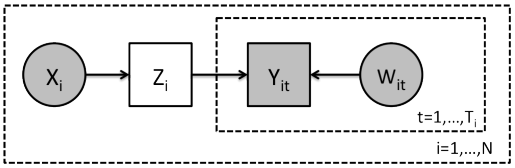


Figure 1: The single-species OD model

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 $Z_i d_{it}$. Note that if $Z_i = 0$, 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.

3 The Multi-Species Occupancy-Detection model

The multi-species OD (MSOD) model consists of observed (Y) and latent binary variables (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 in Equation 1.

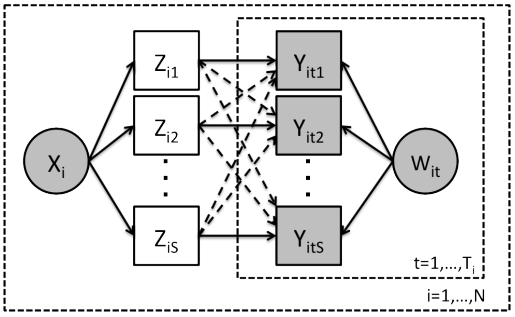


Figure 2: The multi-species OD model

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

Here, $\boldsymbol{Z}_{i\cdot}$ refers to all the S latent occupancy variables at site i. The species-specific occupancy models ($P(Z_{is}|\boldsymbol{X}_i)$ for each s) are parameterized as in the OD model, where $Z_{is} \sim Bernoulli(o_{is})$ and $o_{is} = \sigma(\boldsymbol{X}_i \cdot \boldsymbol{\alpha}_s)$. The detection probabilities ($P(Y_{its}|\boldsymbol{Z}_{i\cdot}, \boldsymbol{W}_{it})$ for each s) depend on the occupancy status of species s (Z_{is}) and the occupancy status of other species s' ($Z_{is'}$, $s' \neq s$) that are parent nodes of the observation variable of species s (Y_{its}).

3.1 Parameterization

We model the detection process based on the noisy-or parameterization of the QMR-DT network for medical diagnosis [4, 13, 6]. 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.

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, $d_{itrs} = P(Y_{its} = 1|Z_{ir} = 1) = \sigma(\boldsymbol{W}_{it} \cdot \boldsymbol{\beta}_{rs})$. 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. Thus

the probability of species s being reported during visit t at site i is given in Equation 2.

$$P(Y_{its} = 1|\boldsymbol{Z}_{i\cdot}, \boldsymbol{W}_{it}) = 1 - P(Y_{its} = 0|\boldsymbol{Z}_{i\cdot}, \boldsymbol{W}_{it}) = 1 - (1 - d_{0s}) \prod_{r=1}^{S} (1 - d_{itrs})^{\gamma_{rs} Z_{ir}} \quad (2)$$

3.2 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 $\gamma_{rs} \in \{0, 1\}$ to $\gamma_{rs} \in [0, 1]$, turning the integer program into a linear program. With this linear relaxation, we then estimate the MSOD model parameters using Expectation Maximization [3]. In the E-step, EM computes the expected occupancies $\boldsymbol{Z}_{i\cdot}$ for every site i using Bayes rule. In the M-step, we use L-BFGS-B [1] to re-estimate the model parameters $\{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}\}$ that maximize the expected log-likelihood in Equation 3.

$$\begin{aligned} Q &= E_{\boldsymbol{Z}|\boldsymbol{Y},\boldsymbol{X},\boldsymbol{W}}[\log(P(\boldsymbol{Y}, \boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{W}))] \\ &= \sum_{i=1}^{N} \sum_{s=1}^{S} \left[E_{\boldsymbol{Z}|\boldsymbol{Y},\boldsymbol{X},\boldsymbol{W}}[\log(P(Z_{is}|\boldsymbol{X}_i))] + \sum_{t=1}^{T_i} E_{\boldsymbol{Z}|\boldsymbol{Y},\boldsymbol{X},\boldsymbol{W}}[\log(P(Y_{its}|\boldsymbol{Z}_{i\cdot}, \boldsymbol{W}_{it}))] \right] \end{aligned} \quad (3)$$

In the learned adjacency matrix, γ_{rs} specifies the probability of species r being confused for species s. We sort pairs of misidentified species according to their probability of misidentification in $\boldsymbol{\gamma}$ and greedily add edges into the model structure until the log-likelihood on a validation set does not improve. Once we determine the structure, we re-estimate the MSOD model ($\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$) with a fixed structure. Exact computation of the expectations in Equation 3 is computationally expensive with large S; we will investigate speedups using variational approximations [6, 10, 12].

4 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 "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. We also show the model structures learned for two case studies using sets of species known to be confused for each other in eBird.

4.1 Synthetic dataset

We generate synthetic data ("Syn") for 500 sites and 3 visits per site,with 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 misidentified 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. We also generate two harder synthetic datasets by allowing species occupancy interactions ("Syn-I") and non-linear occupancy components ("Syn-NL"). For each synthetic dataset, a training, validation and test dataset are generated following the MSOD model, and this entire process is repeated 30 times to generate 30 datasets.

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). We report the AUC and accuracy averaged over 30 datasets in Table 1. 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 from misidentified species and observation even with species occupancy interactions and non-linear terms in the occupancy process.

We also compute a "structural AUC" to compare the learned model structure to the true model structure. The structural AUC value, averaged over 30 datasets, specifies the probability of ranking a true cross edge over an incorrect cross edge in the learned adjacency matrix. The MSOD model

		Occupancy (Z)		Observation (Y)	
		AUC	Accuracy	AUC	Accuracy
Syn	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	$\mathbf{0.935 \pm 0.005}^{\star\dagger}$	$\mathbf{0.872 \pm 0.006}^{\star\dagger}$	$\mathbf{0.776 \pm 0.004}^{\star\dagger}$	$\mathbf{0.750 \pm 0.004}^{\star\dagger}$
Syn-I	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	$\mathbf{0.925 \pm 0.004}^{\star\dagger}$	$\mathbf{0.862 \pm 0.006}^{\star\dagger}$	$\mathbf{0.763 \pm 0.004}^{\star\dagger}$	$\mathbf{0.755 \pm 0.006}^{\star\dagger}$
Syn-NL	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	$\mathbf{0.903 \pm 0.006}^{\star\dagger}$	$\mathbf{0.842 \pm 0.007}^{\star\dagger}$	$\mathbf{0.755 \pm 0.004}^{\star\dagger}$	$\mathbf{0.751 \pm 0.007}^{\star\dagger}$

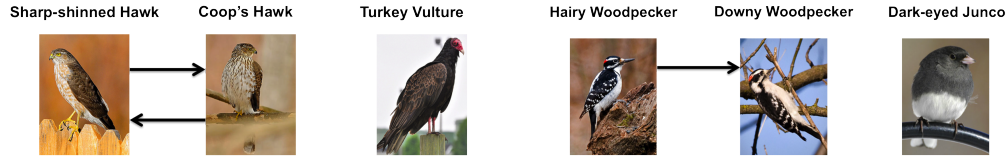
Table 1: The AUC and accuracy of occupancy and observation prediction (with standard error) over 30 synthetic datasets. These metrics are computed per species and averaged across species. $\star$ and $\dagger$ indicate the MSOD model is statistically better than the OD model and the ODLP model.

archives AUC value of more than 0.97 in all three synthetic data, indicating that the MSOD model almost always discovers the correct species confusions.

4.2 eBird dataset

We also test the ability of the MSOD methods to recover sensible structures on two case studies involving real-world eBird data, which was selected by consulting with experts at the Cornell Lab of Ornithology. We evaluated MSOD 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 **Hawks** case study, we consider the Cooper's Hawk and Sharp-shinned Hawk, and Turkey Vulture as a distractor species in California. In the **Woodpeckers** case study, we consider the Hairy Woodpecker and Downy Woodpecker, and Dark-eyed Junco as a distractor species in California. We show the learned model structures in Figure 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, and Hairy Woodpecker is likely to be confused for Downy Woodpecker. For both 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 one due to their lack of birding skills.



(a) Hawks case study (b) Woodpecker case study

Figure 3: The arrows specify the species confusions recovered by the MSOD model.

5 Conclusion

We introduce the multi-species OD model to identify species confusions and show promising preliminary results on both synthetic and eBird data. We plan to apply this model to discover species confusions in eBird that are not already known.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grants No. 1215950, 0941748 and 1209714.

References

[1] 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.

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

[3] 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.

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

[5] 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.

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

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

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

[9] 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.

[10] A. Y. Ng and M. I. Jordan. Approximate inference algorithms for two-layer bayesian networks. volume 12, 1999.

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

[12] J. C. Platt, E. Kiciman, and D. A. Maltz. Fast variational inference for large-scale internet diagnosis. In *Advances in Neural Information Processing Systems*, volume 20, 2007.

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

[14] 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.

[15] 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.

[16] 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.