# Species Distribution Modeling of Citizen Science Data as a Classification Problem with Class-conditional Noise

**Rebecca A. Hutchinson**[1,2] and **Liqiang He**[1] and **Sarah C. Emerson**[3]
[1]School of Electrical Engineering and Computer Science; [2]Dept. of Fisheries and Wildlife; [3]Statistics Dept.
Oregon State University
Corvallis, OR 97331
{rah,heli,sarah.emerson}@oregonstate.edu

## Abstract

Species distribution models relate the geographic occurrence pattern of a species to environmental features and are used for a variety of scientific and management purposes. One source of data for building species distribution models is citizen science, in which volunteers report locations where they observed (or did not observe) sets of species. Since volunteers have variable levels of expertise, citizen science data may contain both false positives and false negatives in the location labels (present vs. absent) they provide, but many common modeling approaches for this task do not address these sources of noise explicitly. In this paper, we propose to formulate the species distribution modeling task as a classification problem with class-conditional noise. Our approach builds on other applications of class-conditional noise models to crowdsourced data, but we focus on leveraging features of the noise processes that are distinct from the class features. We describe the conditions under which the parameters of our proposed model are identifiable and apply it to simulated data and data from the eBird citizen science project.

## Introduction

A species distribution model (SDM) describes the geographic occurrence pattern of a species in terms of environmental features like climate and habitat variables. SDMs are useful for a variety of scientific and management purposes, including making inferences about the habitat requirements for a species and making predictions about whether a species is likely to persist in new locations. In this paper, we focus on a variant of species distribution modeling in which individual species are modeled using binary labels indicating detection or non-detection at a set of locations; that is, we use so-called 'presence-absence' data rather than 'presence-background' or 'occupancy-detection' data (Guillera-Arroita et al. 2015).

Citizen science projects engage volunteers in the scientific process. In the realms of biology and ecology, the role of citizen scientists is often to collect data about which species occur at various locations. In the eBird project, bird watchers can report their observations to a database at the Cornell Lab of Ornithology using an online checklist system (Sullivan et al. 2014). Participants may make observations at any time and location they wish. Each checklist includes the species that were observed, whether or not the observer is reporting all species they detected, measurements indicating the effort expended during the observation, and the time, date, and location of the observation. Note that if the observer reports all the species they detected, we can infer the absence of the complement of this set of species. Using these types of records, the eBird database provides a set of instances for each species (one for each location) with binary labels, which can be linked with environmental features.

In this paper, we propose to formulate the problem of species distribution modeling using citizen science data as a classification problem subject to class-conditional label noise. For the class *present*, label noise arises from *imperfect detection* of secretive or cryptic species, a problem long-recognized in ecological studies (MacKenzie et al. 2002). For the *absent* class, label noise arises due to mistaken identifications by observers (Royle and Link 2006). Some popular species distribution modeling approaches assume the absence of these false positives (MacKenzie et al. 2002) because field studies often employ highly trained observers, but the variable skill levels of volunteers make this a tenuous assumption for citizen science data. False positives have received less attention than false negatives in this domain, though some work has used auxilliary information about surveys to account for them (Miller et al. 2011; Ruiz-Gutierrez, Hooten, and Campbell Grant 2016).

Since we expect the class-conditional noise rates in this setting to be asymmetric, this is a challenging learning problem (Scott, Blanchard, and Handy 2013). In this work, we propose and investigate a model for classification with class-conditional label noise that leverages features of asymmetric label noise without requiring multiple labels per example. The only auxilliary information we require is features of the noise models, such as are available in the information collected in the eBird protocol on the observation conditions for each instance. We analyze the identifiability of the model, use simulated data to explore model behavior in a variety of settings, and apply the model to the eBird data to compare against alternative approaches.

### Related Work

Our work builds on research into classification settings with label noise (Menon et al. 2015; Natarajan et al. 2013; Man-

wani, Sastry, and Member 2013; Li et al. 2007; Lawrence and Scholkopf 2001). Frénay and Verleysen (2014) provide a helpful review of types of label noise and methods for dealing with them. We focus on asymmetric label noise (Scott, Blanchard, and Handy 2013), as opposed to methods that assume equal rates of false positives and false negatives. Our framework is most closely related to the 'noisy not at random' (NNAR) model since we treat label noise as feature-dependent (Frénay and Verleysen 2014). In contrast to the typical view of NNAR noise, the set of features on which the noise depends are fully or partially disjoint from the features on which the true class label depends in our work.

Citizen science has both similarities and differences from crowdsourcing paradigms for generating labeled data, in which workers are presented with a set of instances to label. Some similarities are obvious; for example, the labels are being generated by a pool of workers with variable skill levels. Errors are likely to occur in both settings, causing both false negatives and false positives. In both paradigms, the difficulty of the instance as well as the skill and effort levels of the labeler may affect the probability of mislabeling an instance (Bi et al. 2014). However, there are some important distinctions between citizen science and crowdsourcing as well. In crowdsourcing, the set of instances to be labeled and their assignments to labelers are controlled by the task setting. In contrast, citizen scientists effectively choose their own instances to label from an infinite set of possibilities, since volunteers choose the times and places to make their observations. A consequence of this freedom is that we do not in general have multiple labels of each instance. Some analyses of citizen science data have grouped checklists that are close in space and time to construct multiple label structure (Yu, Hutchinson, and Wong 2014; Hutchinson, Liu, and Dietterich 2011; Yu, Wong, and Hutchinson 2010), but this requires assumptions about the granularity with which to aggregate, which we seek to avoid in this work. In addition, there are sources of variability in mislabeling probabilities in citizen science that go beyond labeler effort and task difficulty: the observation conditions under which the labels were generated (e.g. time of day, weather). Work by Raykar *et al.* (2010) takes a similar approach to crowdsourced data as the model we propose herein, except that it relies on multiple labels per instance and does not include features in the noise models.

Our approach also builds on statistical models in ecology. MacKenzie *et al.* (2002) popularized *occupancy models* to account for imperfect detection in ecological studies. This approach requires multiple observations during a period when the true status of the species (the true class label) is unchanging, and it incorporates an explicit model of the observation process to correct for underreporting. However, it assumes that there are no false positives in the data. Further work on this family of models added the possibility of false positives (Royle and Link 2006), but the resulting model had limited application in practice due to identifiability issues (Miller et al. 2011). Extensions of it improved model behavior, but they required additional information (e.g. multiple survey methods; (Miller et al. 2011)). More recent work has also explored variants of occupancy models requiring only a single observation at each location (Lele, Moreno, and Bayne 2012). There is also discussion in the literature about the identifiability of single-visit occupancy models (without false positives) (Knape and Korner-Nievergelt 2014; Sólymos and Lele 2015); in particular, identifiability for these models relies on the fidelity of the link function.

Our proposed approach unifies several of these frameworks. It can be seen as a novel use of the model from Royle and Link (2006), applied to single-observation data rather than multiple-observation data. Alternatively, it can be seen as an extension of the single-observation occupancy model of Lele, Moreno, and Bayne (2012) that allows for false positives. In the crowdsourcing context, our approach can be viewed as an extension of the Raykar *et al.* (2010) model that uses only a single label per instance and introduces features to describe the noise processes. The contributions of this paper are: to connect literature addressing this problem across several research communities (crowdsourcing, ecology, machine learning); to propose a generalization of the existing approaches described above; to describe conditions for identifiability of the proposed model; to explore performance of the proposed approach in simulation as compared with existing approaches; to evaluate whether model selection identifies the features associated with each sub-model of the approach correctly; and to compare predictions from the proposed approach to existing approaches.

## Model Framework

Our notation follows that of Frénay and Verleysen, 2014. Let $Y$ be the true class, $\tilde{Y}$ be the observed label, and $E$ be a binary variable indicating whether a labeling error is present (i.e. $Y \neq \tilde{Y}$). In this paper, we address binary classification ($Y \in \{0, 1\}$) and leave the multi-class setting to future work. Let $X$ be a vector of features related to $Y$. The data-generating model is that the true class $Y$ is chosen based on the features $X$. Then, if an error occurs ($E = 1$), the label is flipped to produce $\tilde{Y}$.

The key difference between our proposed framework and previous work is that we allow errors to depend on an additional set of features $W$, distinct from the features $X$ that influence the true class. (Below, we explore cases with overlap between these feature sets, but they play a conceptually different role in the framework.) For clarity, we will refer to $X$ as the *class features* and $W$ as the *noise features*. We define the following quantities:

$$\psi := P(Y = 1|X) = \sigma_f(f(X; \alpha))$$
$$\rho := P(E = 1|Y = 0, W) = \sigma_g(g(W^{(\rho)}; \beta))$$
$$\eta := P(E = 1|Y = 1, W) = \sigma_h(h(W^{(\eta)}; \gamma)),$$

where $W^{(\rho)}$ is the subset of the noise features that impact the probability of a false positive ($\rho$), $W^{(\eta)}$ is the subset of the noise features that impact the probability of a false negative ($\eta$), and $\sigma$ denotes the link function for each component (e.g. logistic). We use the following vocabulary conventions to designate the different pieces of this model: the functions $\psi, \rho$, and $\eta$ will be referred to as the *fundamental parameters*; the functions $\sigma_f, \sigma_g$, and $\sigma_h$ will be referred to as the *link functions*; the functions $f, g$, and $h$ will be referred to as

the *feature functions*; and the coefficients $\alpha, \beta$, and $\gamma$ will be referred to as the *coefficient parameters*. Letting the probability of a positive observation for instance $i$ be

$$p_i = \psi_i(1 - \eta_i) + (1 - \psi_i)\rho_i,$$

the likelihood function for the model is

$$L = \prod_{i=1}^{N} p_i^{Y_i}(1 - p_i)^{1-Y_i}.$$

In the experiments below, we fit model parameters using maximum likelihood estimation. We used an existing implementation of the Royle and Link (2006) model available in the R package *unmarked* (R Core Team 2015; Fiske and Chandler 2011), since it was sufficiently general to allow for features of the noise models and lack of replicate surveys. Since the likelihood is not convex, we select the fit with the greatest likelihood from a set of random restarts. Occasionally the optimization does not converge, so we allow 50 random restart attempts to get 40 successful fits.

## Identifiability

The basic question we would like to answer is: Under condition $X$, will occupancy probability be increased or decreased (and by how much) compared to some baseline condition $X_0$? We cannot answer this most general form of the question without making some assumptions (constraints). As a demonstration of this fact, suppose that we observe that for larger values of $X$, $\tilde{Y}$ is more likely to be one. Unless we make additional constraining assumptions, we cannot tell whether this is because $Y$ is more likely to be one in this setting, $E$ is more likely to be one when $Y$ is zero in this setting, or $E$ is more likely to be zero when $Y$ is one in this setting (or some combination of these possibilities).

Without significantly constraining the problem, we are unable to distinguish between these different explanations. This is the same concern that was discussed by Knape and Korner-Nievergelt (2014), and addressed further with some constraint recommendations by Sólymos and Lele (2015). Specifically, we are concerned with the identifiability of the fundamental parameters $\psi, \rho$, and $\eta$, and also with the identifiability of the coefficient parameters $\alpha, \beta$, and $\gamma$. Clearly the coefficient parameters are not identifiable if the fundamental parameters are non-identifiable. If the fundamental parameters *are* identifiable, then the identifiability of the coefficient parameters depends on the form and parameterization of the covariate functions. Our concern is how to make both the the fundamental parameters *and* the coefficient parameters identifiable, since it is the coefficient parameters that address the primary scientific questions of interest: namely, if a covariate value changes, what are the effects on the probability that the species of interest is present and, secondarily, on the probabilities of observation errors.

Starting with complete generality, letting $Z = (X, W)$ and $\theta = (\psi, \eta, \rho)$, we need:

$$P(\tilde{Y} = 1|Z = z, \theta) = \psi(z)\left[1 - \eta(z)\right] + \left[1 - \psi(z)\right]\rho(z)$$

to satisfy

$$\sup_z \left| P(\tilde{Y} = 1|z, \theta) - P(\tilde{Y} = 1|z, \theta^*) \right| > 0$$

whenever $\theta \neq \theta^*$. First note that there are an infinite number of solutions since there are three unknowns in this single equation. The solutions are naturally constrained by the fact that $\psi, \eta$, and $\rho$ are all probabilities, and therefore must be in the range $[0, 1]$. The solution space is further reduced if we require that $\psi$, $\eta$, and $\rho$ have certain functional forms as a function of $Z$, but of course then the solution is dependent upon the form chosen. We focus here on the logistic link function, as one of many link functions that would work to assist identifiability. In some settings, it can be further useful to require that at least one of the active covariates is continuous, since in the case of discrete covariates handled with indicator functions the functional form may not provide as much structure: if a different value of $\psi(z), \rho(z)$, and $\eta(z)$ is allowed for every possible discrete value of $Z = z$, we are back to the most general case and have lost identifiability. However, in more restricted settings such as when distinct non- or minimally-overlapping sets of covariates affect the different fundamental parameters $\psi$, $\rho$, and $\eta$, we have identifiability without requiring that any of the covariates be continuous.

Note that even specifying the functional form (link function) is not enough to provide identifiability if the model families for $\eta$ and $\rho$ are the same and include 1 - $\sigma$ whenever $\sigma$ is a valid link. This is because the following two complementary models give identical probabilities: $\theta_1 = (\psi, \eta, \rho)$ and $\theta_2 = (1 - \psi, 1 - \rho, 1 - \eta)$. To address this complementary model source of non-identifiability, we must constrain the problem to allow only one of these results (i.e., define a way to select between these two equal-likelihood solutions). This can be accomplished in several ways: either by selecting the solution that gives values of $\psi$, $\eta$, or $\rho$ in a certain range (e.g., choose the solution with the minimum value of $\eta$), or by specifying that, for instance, the function $\eta$ depends on a certain set of covariates, while $\rho$ depends on a distinct set. Constraining the covariates that appear in the different noise models to be distinct sets will work to eliminate the issue of complementary model solutions since if we know that $\rho = \sigma_g(g(W^{(\rho)}; \beta))$ and $\eta = \sigma_h(h(W^{(\eta)}; \gamma))$ with distinct covariate sets $W^{(\rho)}$ and $W^{(\eta)}$, then $\rho_2 = 1 - \eta = 1 - \sigma_h(h(W^{(\eta)}; \gamma))$ is not a valid solution because it depends on the wrong covariates (and likewise for $\eta_2 = 1 - \rho$).

The two complementary equally optimal solutions are also the reason that this model does not perform well when the noise models do not depend on covariates (the 'constant noise' settings in the experiments below): the resulting lack of identifiability means that there are two complementary solutions that fit the observed data the exact same, so the parameter estimates are not unique (and thus have very large variances). Likewise, if the covariates that affect the noise models are effectively close to constant, identifiability will be very fragile as we will be very close to the constant noise setting. Thus we have a spectrum of identifiability that depends on the chosen form of the model, and the roles and distributions of the relevant covariates.

| Number | Name | Feature overlap | Feature distributions | Link |
|---|---|---|---|---|
| 1 | *none-cont-logit* | none | all $N(0,1)$ | *logit* |
| 2 | *none-mix-logit* | none | odd $N(0,1)$, even $Bern(0.5)$ | *logit* |
| 3 | *none-cat-logit* | none | all $Bern(0.5)$ | *logit* |
| 4 | *noise-cont-logit* | noise: $W_2 = W_4$ | all $N(0,1)$ | *logit* |
| 5 | *noise-mix-logit* | noise: $W_2 = W_4$ | odd $N(0,1)$, even $Bern(0.5)$ | *logit* |
| 6 | *class-cont-logit* | class and noise: $X_1 = W_1$ and $X_3 = W_3$ | all $N(0,1)$ | *logit* |
| 7 | *class-mix-logit* | class and noise: $X_1 = W_1$ and $X_3 = W_3$ | odd $N(0,1)$, even $Bern(0.5)$ | *logit* |
| 8 | *noise-mix-probit* | noise: $W_2 = W_4$ | odd $N(0,1)$, even $Bern(0.5)$ | *probit* |
| 9 | *noise-mix-scale* | noise: $W_2 = W_4$ | odd $N(0,1)$, even $Bern(0.5)$ | *scale* |

Table 1: Simulated data settings. Names reflect the overlap setting, feature distributions, and link functions. Feature overlap refers to the extent to which the features of each feature function were shared. Feature distributions were either Gaussian or Bernoulli, in some cases depending on whether the index of the feature was odd or even. The link function listed was used to generate data, though the *logit* link was always used in fitting.

# Simulation Experiments

## Data Generation

We designed simulation experiments to explore model performance and identifiability under scenarios with varying feature overlap between feature functions (i.e. $f$, $g$, and $h$), feature distributions, the true link function, class balance, noise rates, and the number of training instances. We considered models of the following form.

$$f = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3$$
$$g = \beta_0 + \beta_1 W_1 + \beta_2 W_2$$
$$h = \gamma_0 + \gamma_1 W_3 + \gamma_2 W_4$$

In the first set of simulations, all seven features were generated independently (called *none*, for no overlap). We created three variants of this setting (#1-3 in Table 1): 1) *cont*: all continuous features, from $N(0,1)$, 2) *mix*: odd-indexed continuous and even-indexed categorical features, from $N(0,1)$ and $Bern(0.5)$, and 3) *cat*: all categorical features, from $Bern(0.5)$. In the second set of simulations, we set $W_2 = W_4$ to explore the effect of overlapping noise features (called *noise*, for overlap in the noise features; #4-5 in Table 1). In this setting, we created variants with both continuous and mixed features. In the third set of simulations, we generated overlap between the class and noise models by setting $X_1 = W_1$ and $X_3 = W_3$ (called *class*, for overlap in the class and noise features; #6-7 in Table 1). Similarly to the *noise* setting, we used the *cont* and *mix* feature settings.

In addition to varying the feature overlap and distributions, we explored the effect of a misspecified link function. As explained above, the identifiability of the model parameters depends on the link function. To explore sensitivity to this condition, we generated datasets from the *noise-mix* setting using two incorrect link functions (the model was always fit with a *logit* link): the *probit* link and a *scale* link that scaled and shifted the real values linearly to transform them into probabilities (#8-9 in Table 1). The feature overlap, distribution, and link settings are summarized in Table 1. Note that all settings have at least one distinct feature in each feature function, so the coefficient parameters are identifiable.

For each of these settings, we varied $\alpha$, $\beta$, and $\gamma$ to explore the effects of class balance and noise rates. For all scenarios, we set the non-intercept coefficients to 1 and varied the intercepts to explore the effects of class balance and noise levels on performance. For class balance, the low, medium, and high values were $\bar{\psi} \in \{0.25, 0.5, 0.75\}$. For the noise levels, the low, medium, and high levels were $\bar{\eta} \in \{0.1, 0.25, 0.4\}$ and $\bar{\nu} \in \{0.1, 0.25, 0.4\}$. For each combination of feature settings (9) and parameter settings (27), we simulated 30 training sets each of sizes 200, 400, 800, 1600, 3200, 6400, and 12800 and one test set of 5000 examples. Additional details of the simulated data are available in supplemental material.

## Method Comparison and Evaluation

For each training set, we compared five methods:

1. *LR1*: Logistic regression ignoring label noise (using just the class features to predict the noisy labels).
2. *LR2*: Logistic regression ignoring the label structure but using the noise features (using the class and noise features all together to predict the noisy labels).
3. *OCC* (for OCCupancy): Single-visit occupancy model, using all noise features for false negatives and ignoring false positives (MacKenzie et al. 2002).
4. *CN* (for Constant Noise models): Proposed model with false positives and false negatives but without noise features (using a constant/intercept-only model for both); similar to Raykar *et al.* (2010).
5. *FP*: (for False Positive models with features) Proposed model with noise features for both false positives and false negatives. Sets of noise features in feature functions $g$ and $h$ were consistent with the data-generating models, so the fundamental parameters were identifiable.

For all methods, we measured the quality of predictions of the true class labels using mean squared error (MSE) in the class probability predictions.

## Simulation Results

Here, we present results from the intermediate setting for class balance and noise rates as a representative example,

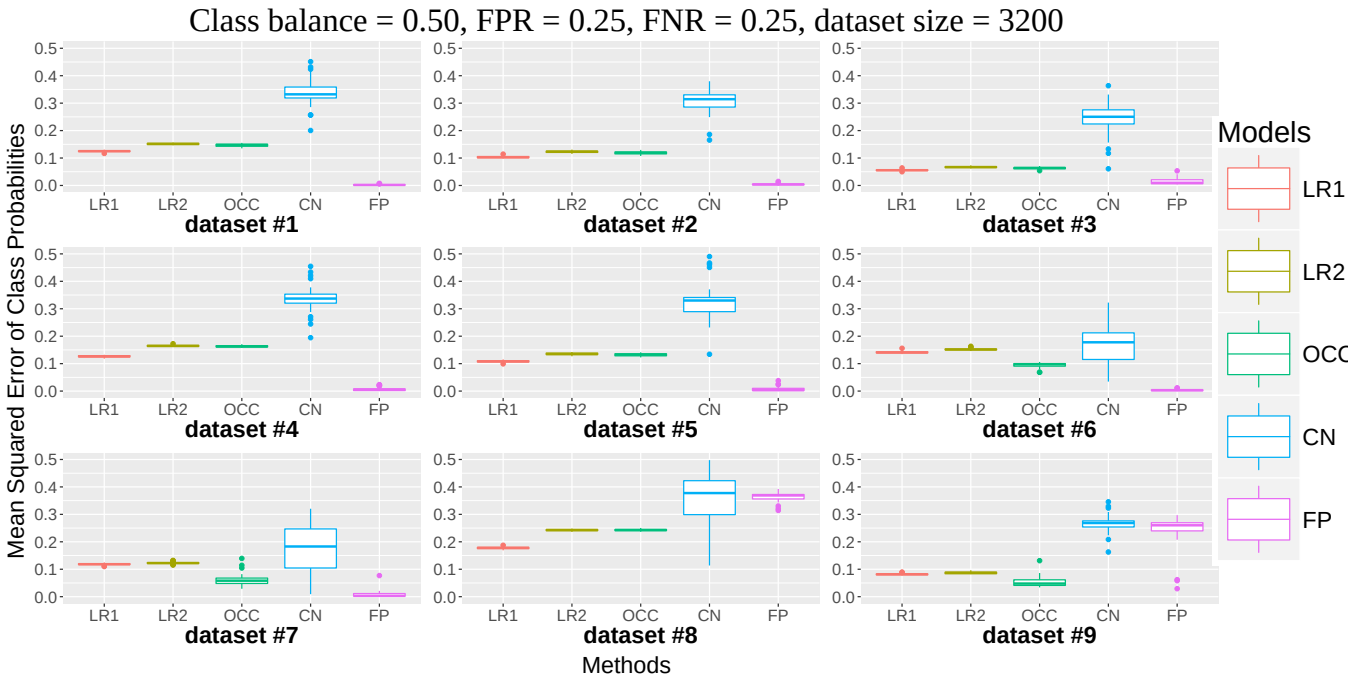


Figure 1: Mean squared error in the class probabilities ($\psi$) for each method on each of the data-generating models (Table 1). All datasets had 3200 training instances, and each boxplot represents 30 simulated datasets.

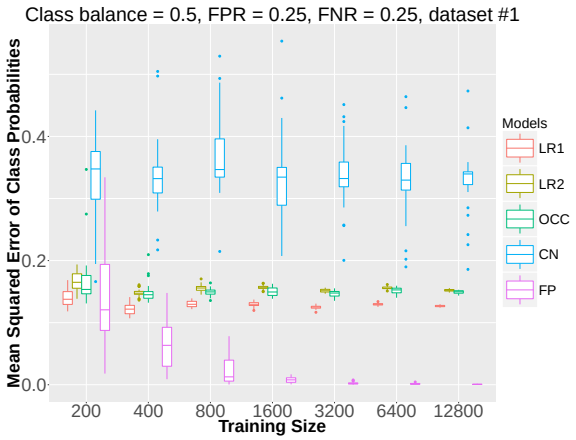


Figure 2: Mean squared error in the class probabilities for dataset #1 with mid-level class balance and noise rates. Each boxplot represents 30 simulated datasets.

since the results are relatively consistent across settings (results for all settings and datasets are in the supplement; Figures S1-S27). In this ideal setting with identifiable, well-specified models and enough data, the types of features and their overlap in the feature functions is not critical (Figure 1, #1-7). The *FP* model predicts the class probabilities better than the alternative methods, as expected. Exceptions occur when the link function is misspecified (Figure 1, #8-9). The *CN* model tends to perform worst, which is likely due to lack of identifiability in the noise parameters, since the *CN* model does not have features specified for $g$ and $h$ to distinguish between the two symmetric solutions discussed above.

However, these results only hold if there is sufficient data for fitting the *FP* model. The sample complexity requirements of the *FP* model are greater than those of simpler models, so for small datasets, even the simple logistic regressions predict class probabilities better (Figure 2). The specific threshold for 'enough' data to support the *FP* approach will vary depending on other characteristics of the problem (e.g. number of features).

## Empirical Experiments

### eBird Data

The eBird Reference Dataset consists of checklists indicating which species were observed during birding events in which citizen scientists report all of the species they observed (Munson et al. 2012). We used data collected in 2012. Following previous analyses of eBird data (Yu, Wong, and Hutchinson 2010; Hutchinson, Liu, and Dietterich 2011; Yu, Hutchinson, and Wong 2014), we chose to focus on stationary and traveling counts from California and New York in May, June, and July, during which time habitat associations are relatively stable.

The eBird Reference Dataset is distributed with both environmental (class) features and observation (noise) features. The features we considered included 11 class features (3 real, 2 categorical, 6 principal components of land cover measurements) and 5 noise features (all real-valued) and are listed in Table 2. We scaled continuous features to $N(0, 1)$. After removing records with missing values and outliers for

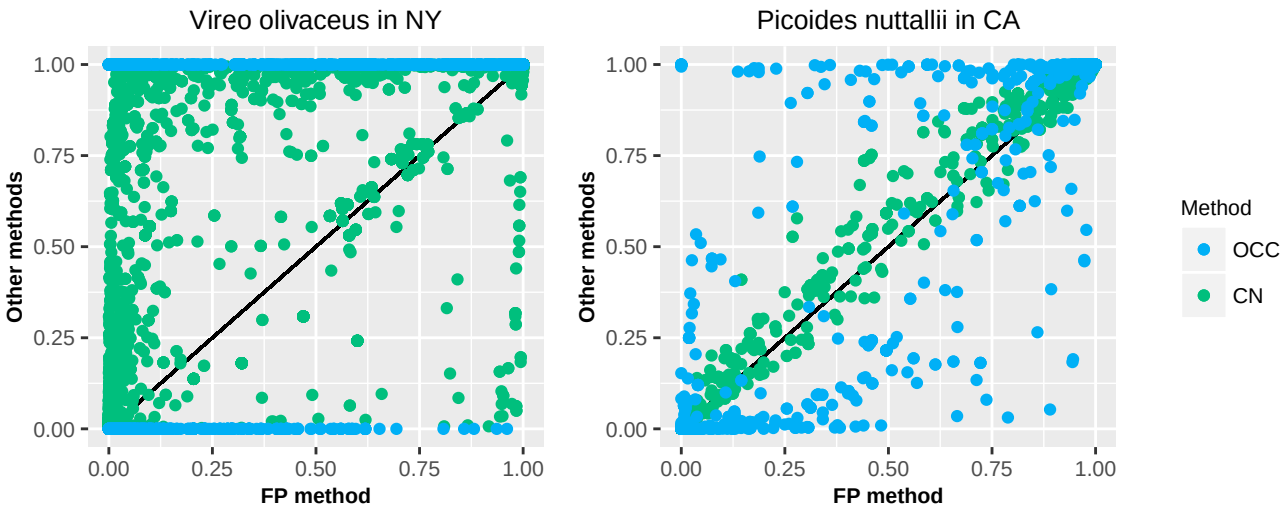


Figure 3: Comparison of predicted class probabilities from the *FP* method on the x-axis versus the *OCC* and *CN* methods on the y-axis. The estimated average false positive rates are 0.095 and 0.0058 for *Vireo olivaceus* and *Picoides nuttallii* respectively.

| Habitat Features | |
|---|---|
| Feature | Type |
| Human population from 2000 census | real |
| Housing percent vacant | real |
| Elevation | real |
| Average temperature | categorical |
| Precipitation | categorical |
| Percent of surrounding area covered in 15 land cover types | real |
| **Observation Features** | |
| Feature | Type |
| Day of year | real |
| Time of day | real |
| Effort in hours | real |
| Effort in distance travelled | real |
| Number of observers | real |

Table 2: Features of models fit to the eBird data, taken from the eBird Reference Dataset.

the features, the California data contained 16,742 checklists and the New York data contained 11,982 checklists. For each state, we randomly selected 4000 checklists as a test set, 4000 checklists as a validation set, and used the remaining checklists for training.

The eBird Reference Dataset is also distributed with information about the species it contains, including whether or not one species is often confused with another. We selected species with this property for this analysis, since species confusions are a potential source of false positives in the eBird data (Yu, Hutchinson, and Wong 2014). In addition, we limited the species pool to species observed in at least 10% of the checklists in the state. The pool of species is listed in the supplement (Table S4).

We also created two simulated species using the eBird features, which are more realistic than standard normal or binary features (see supplement, Table S2). The class models for both species depended on two real-valued class features. The noise models for the first simulated species (*Sim1*) had disjoint sets of features (two for false negatives and one for false positives). The noise models for the second simulated species (*Sim2*) shared one feature (two features for each model with one overlapping). Both species had true occupancy rates near 55%. The average false negative rates were 20% and 40%, and the average false positive rates were 5% and 10%, for *Sim1* and *Sim2* respectively.

## eBird Experiments

In the simulated experiments, we achieved identifiability by specifying each feature function in accordance with the data-generating mechanisms. In the eBird data, we faced a model selection problem in assigning features to feature functions, in particular for the noise models. We fit 21 different models to each of the real and simulated eBird species, all of which met the identifiability conditions discussed above (at least one unique feature in each feature function). In each model, the class feature function included all of the habitat features. The false positive and false negative feature functions partitioned the noise features differently. Models 1-16 assigned each noise feature to exactly one of the noise models. Models 17-21 included all noise features in one noise model and all noise features except one in the other noise model (see supplement, Table S3). Each of these 21 models has a symmetric analog, so in each case we chose between the pair of symmetric models by selecting the one in which the false positive rate ($\rho$) was less than the detection rate ($1-\eta$). That is, we used the constraint that observers are more likely to detect the correct species than to misidentify it.

We compared against the same set of methods as in the simulated data experiments. For the simulated species, we can refer to the data-generating models to evaluate the mod-

els. For the real species, we do not have access to 'ground truth' about the species true presence or absence while a checklist was collected, but we can examine the differences in the class probabilities predicted by different methods.

### eBird Results

For *Sim1*, the models selected in each state based on the validation sets were not congruent with the data-generating mechanisms. None of the models were fully correct, so by 'congruent' we mean that the correct features were included in the noise models in combinations such that either the model as written or the symmetric analog could represent the generating model if the irrelevant features were given coefficients of 0. Interestingly though, on the test sets, the selected models had lower MSE on $\psi$ compared with most other *FP* models and the *OCC* model. The selected models also had lower MSE on $\psi$, $\rho$, and $\eta$ than the *CN* model (see supplement, Tables S5 and S7). Therefore, despite lack of a perfect representation of the data-generating model, the predictions were superior to alternative methods. For *Sim2*, the models selected in each state were congruent with the data-generating mechanisms (see supplement, Tables S6 and S8). Again, many of the 21 *FP* models, including those unable to represent the data-generating mechanism, outperformed *OCC* and *CN*. In California, the vast majority of the *FP* models outperform the *OCC* and *CN* alternatives, whereas only a subset of the *FP* models in New York are clearly superior to the alternatives. This may be due to greater sample size; the training dataset for California is roughly twice as large as that for New York.

For the real species, estimated false positive rates ranged from 0.0058 to 0.095 (see supplement, Table S4). These low rates are likely due to eBird's quality control measures. Given the low rates, some *FP* models made predictions similar to the *OCC* models that ignored false positives, like *Picoides nuttallii* in California. For other species, like *Vireo olivaceus* in New York, the predictions from *FP* and *OCC* models differ more, suggesting overprediction of the species by the *OCC* model (Figure 3). Some models, like the *OCC* model for *Vireo olivaceus* in New York, also showed signs of overfitting in the form of boundary estimates for the class probabilities. Figure 3 also compares the *FP* predictions to the *CN* predictions for the class probabilities; the degree of correlation between these predictions varied across species. We expect this variation is due to differences in the importance of the noise features and to identifiability issues with the *CN* method.

## Discussion

In this paper, we have explored the idea of treating species distribution modeling of citizen science data as a classification problem with class-conditional label noise. Our approach stands in contrast to methods that ignore labeling errors or only address the more common case of false negatives while ignoring false positives. Failing to account for observation error can have severe consequences; not only can the distribution of a species be underestimated, but its relationship to features of interest can be estimated arbitrarily poorly (MacKenzie et al. 2006). While we are motivated by the citizen science domain, we note that the problem we address is more general. Our work is applicable for settings in which the sensors providing labels fail not at random, nor based solely on the features related to the true class label, but based on some combinaton of features describing the observation conditions, the sensors themselves, and the instances they are labeling. Other domains that use human sensors or labelers may fit this description; e.g. if available, characteristics of workers in the Amazon Mechanical Turk system (Yuen, King, and Leung 2011) like skill levels or time spent on task could be used as noise features, separately from class features for predicting the label itself. For labels provided by mechanical sensors, noise features might include exposure of the sensor to potential sources of damage (e.g. weather) or the sensor's inherent failure rates.

Our simulation studies elucidate the conditions under which the proposed approach is most promising. Since the model is more complex than simpler alternatives, it is important that enough data be available for fitting to realize benefits. For identifiability, each feature function should have at least one unique feature. As with similar models in the ecology literature, identifiability also relies on the fidelity of the link functions.

The eBird case study explored model selection among a variety of feature combinations for the noise feature functions. For the simulated species with low false positive rates, the selected model was not congruent with the data-generating mechanism, though it gave better predictions of the true class probabilities on the test set than most alternatives. We hypothesize that the low FP rates (5%) contributed to selecting an inconsistent model, though perhaps these models could be recovered with more data. For the simulated species with higher FP rates (10%), a congruent model was selected and outperformed other *FP* models as well as *OCC* and *CN*. The real species in the eBird case study were estimated to have low false positive rates, leading to substantial consistency with the *OCC* models and in some cases the *CN* models. Given the uncertainty around model selection for the simulated species with the lowest false positive rates, caution is warranted in interpreting the selected models for the real species.

In future work, we plan to explore better strategies for optimization and avoiding local optima than the current use of random restarts in the proposed method. We will experiment with regularization to penalize model complexity, which may also simplify the optimization procedure. We will also evaluate the method with additional noise-injection strategies and apply it to other domains in which the noise features are distinct from the class features. Finally, our next steps will include more thorough ecological evaluation of the model results to gauge the practical relevance of the method for species distribution modeling.

**Acknowledgements** We are grateful for helpful comments from three anonymous referees, whose input substantially improved this paper. This work was supported by the National Science Foundation under Grant No. CCF-1215950.

## References


Bi, W.; Wang, L.; Kwok, J. T.; Tu, Z.; Kong, H.; States, U.; and States, U. 2014. Learning to Predict from Crowdsourced Data. In *Uncertainty in Artificial Intelligence: Proceedings of the Thirtieth Conference*.

Fiske, I., and Chandler, R. 2011. unmarked: An R package for fitting hierarchical models of wildlife occurrence and abundance. *Journal of Statistical Software* 43(10):1–23.

Frénay, B., and Verleysen, M. 2014. Classification in the Presence of Label Noise: a Survey. *IEEE Transactions on Neural Networks and Learning Systems* 25(5):845–869.

Guillera-Arroita, G.; Lahoz-Monfort, J. J.; Elith, J.; Gordon, A.; Kujala, H.; Lentini, P. E.; McCarthy, M. A.; Tingley, R.; and Wintle, B. A. 2015. Is my species distribution model fit for purpose? Matching data and models to applications. *Global Ecology and Biogeography* 24(3):276–292.

Hutchinson, R. A.; Liu, L.-P.; and Dietterich, T. G. 2011. Incorporating Boosted Regression Trees into Ecological Latent Variable Models. In *Proceedings of the Twenty-fifth Conference on Artificial Intelligence*.

Knape, J., and Korner-Nievergelt, F. 2014. Estimates from non-replicated population surveys rely on critical assumptions. *Methods in Ecology and Evolution* 6:298–306.

Lawrence, N. D., and Scholkopf, B. 2001. Estimating a Kernel Fisher Discrimant in the Presence of Label Noise. *Proceedings of the 18th International Conference on Machine Learning* 306–313.

Lele, S. R.; Moreno, M.; and Bayne, E. 2012. Dealing with detection error in site occupancy surveys: what can we do with a single survey? *Journal of Plant Ecology* 5(1):22–31.

Li, Y.; Wessels, L. F.; de Ridder, D.; and Reinders, M. J. 2007. Classification in the presence of class noise using a probabilistic Kernel Fisher method. *Pattern Recognition* 40(12):3349–3357.

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.

MacKenzie, D. I.; Nichols, J. D.; Royle, J. A.; Pollock, K. H.; Bailey, L. L.; and Hines, J. E. 2006. *Occupancy estimation and modeling: inferring patterns and dynamics of species occurrence*. Elsevier, San Diego, USA.

Manwani, N.; Sastry, P. S.; and Member, S. 2013. Noise Tolerance Under Risk Minimization. *IEEE TRANSACTIONS ON CYBERNETICS* 43(3):1146–1151.

Menon, A. K.; van Rooyen, B.; Ong, C. S.; and Williamson, R. C. 2015. Learning from Corrupted Binary Labels via Class-Probability Estimation. *Journal of Machine Learning Research* 37.

Miller, D. A.; Nichols, J. D.; McClintock, B. T.; Grant, E. H. C.; Bailey, L. L.; and Weir, L. A. 2011. Improving occupancy estimation when two types of observational error occur: Non-detection and species misidentification. *Ecology* 92(7):1422–1428.

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. 2012. The ebird reference dataset, version 4.0. Technical report.

Natarajan, N.; Dhillon, I. S.; Ravikumar, P.; and Tewari, A. 2013. Learning with Noisy Labels. *Advances in neural information processing systems* 1196–1204.

R Core Team. 2015. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

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.

Ruiz-Gutierrez, V.; Hooten, M.; and Campbell Grant, E. H. 2016. Uncertainty in biological monitoring: a framework for data collection and analysis to account for multiple sources of sampling bias. *Methods in Ecology and Evolution* (Jones 2011):in press.

Scott, C.; Blanchard, G.; and Handy, G. 2013. Classification with Asymmetric Label Noise : Consistency and Maximal Denoising. *JMLR: Workshop and Conference Proceedings* 30:1–23.

Sólymos, P., and Lele, S. R. 2015. Revisiting resource selection probability functions and single-visit methods: Clarification and extensions. *Methods in Ecology and Evolution* 7:196–205.

Sullivan, B. L.; Aycrigg, J. L.; Barry, J. H.; Bonney, R. E.; Bruns, N.; Cooper, C. B.; Damoulas, T.; Dhondt, A. A.; Dietterich, T.; Farnsworth, A.; Fink, D.; Fitzpatrick, J. W.; Fredericks, T.; Gerbracht, J.; Gomes, C.; Hochachka, W. M.; Iliff, M. J.; Lagoze, C.; La Sorte, F. A.; Merrifield, M.; Morris, W.; Phillips, T. B.; Reynolds, M.; Rodewald, A. D.; Rosenberg, K. V.; Trautmann, N. M.; Wiggins, A.; Winkler, D. W.; Wong, W. K.; Wood, C. L.; Yu, J.; and Kelling, S. 2014. The eBird enterprise: An integrated approach to development and application of citizen science. *Biological Conservation* 169:31–40.

Yu, J.; Hutchinson, R. A.; and Wong, W.-k. 2014. A Latent Variable Model for Discovering Bird Species Commonly Misidentified by Citizen Scientists. *Proceedings of the 29th National Conference on Artificial Intelligence* 500–506.

Yu, J.; Wong, W.-K.; and Hutchinson, R. A. 2010. Modeling Experts and Novices in Citizen Science data for Species Distribution Modeling. In *Proceedings of the 10th IEEE International Conference on Data Mining (ICDM 2010)*.

Yuen, M. C.; King, I.; and Leung, K. S. 2011. A survey of crowdsourcing systems. In *Proceedings - 2011 IEEE International Conference on Privacy, Security, Risk and Trust and IEEE International Conference on Social Computing, PASSAT/SocialCom 2011*, 766–773.