Rempel, R. S., J. M. Jackson, S. L. Van Wilgenburg, and J. A. Rodgers. 2019. A multiple detection state occupancy model using autonomous recordings facilitates correction of false positive and false negative observation errors. *Avian Conservation and Ecology* 14(2):1. <https://doi.org/10.5751/ACE-01374-140201>

Copyright © 2019 by the author(s). Published here under license by the Resilience Alliance.

Methodology

A multiple detection state occupancy model using autonomous recordings facilitates correction of false positive and false negative observation errors

Robert S. Rempel¹, Janet M. Jackson¹, Steven L. Van Wilgenburg² and Jennifer A. Rodgers¹

¹Ontario Ministry of Natural Resources and Forestry, Centre for Northern Forest Ecosystem Research, ²Environment and Climate Change Canada, Canadian Wildlife Service

ABSTRACT. Bird surveys have relied upon acoustic cues for species identification for decades; however, errors in detection and identification can lead to misclassification of the site occupancy state. Although significant improvements have been made to correct for false negative (FN) error, less work has been done on identifying and modeling false positive (FP) error. In our online survey we found misidentification can occur even among highly skilled observers, thus methods are required to correct for FP error. In this study we model both FP and FN error in bird surveys using a multiple detection state model (MDSM), and found that modeling both types of error lowered occupancy (ψ) relative to the FN only models in 84% of the observation data sets, and this suggests significant bias in ψ can occur in studies that do not correct for both FN and FP error. In our autonomous recording units (ARU) data we had two detection states, “confirmed” and “unconfirmed,” where confirmation was based on agreement of two interpreters, and through simulation evaluated performance of the MDSM using this type of ARU data. We found that MDSM can effectively correct for both FN and FP error across a broad range of survey observation rates and detection rates (d) and is appropriate for data using “confirmed detections.” We developed a binary classification model to assign risk of bias to field observation sets based on survey and model parameters, and found that lower risk of bias cannot be predicted by a single variable or value, but rather occurs under certain combinations of low naïve occupancy rate ($< \sim 0.2$), detection rate ($< \sim 0.2$), number of confirmed recordings ($< \sim 20$) and high FP rate ($> \sim 0.07$). Our approach to interpreting ARU data along with our analysis guidelines should help reduce potential inflation of ψ resulting from FP error.

L'emploi d'un modèle d'occupation à états multi-détections utilisant des enregistrements autonomes facilite la correction d'erreurs d'observations liées aux faux positifs et aux faux négatifs

RÉSUMÉ. La détection de vocalisations pour identifier les espèces est intrinsèque aux relevés d'oiseaux depuis des décennies; toutefois, les erreurs de détection et d'identification peuvent mener à des erreurs quant à l'occupation d'un site. Bien que d'importantes améliorations ont été proposées pour corriger les erreurs liées aux faux négatifs (FN), moins de travaux ont porté sur la détermination et la modulation des erreurs liées aux faux positifs (FP). Dans notre enquête en ligne, nous avons constaté que les mauvaises identifications peuvent même être le fait d'observateurs chevronnés, de sorte qu'il est nécessaire d'élaborer des méthodes pour corriger les erreurs liées aux FP. Dans la présente étude, nous avons modélisé les erreurs liées aux FP et aux FN de relevés d'oiseaux au moyen d'un modèle à états multi-détections (MEMD). Nous avons trouvé que la modélisation des deux types d'erreurs diminuait le taux d'occupation (ψ) comparativement aux modèles ne tenant compte que des FN dans 84 % des jeux de données d'observations, ce qui indique que ψ peut comporter un biais important dans le cas d'études ne corrigeant pas les deux types d'erreurs. Nos données tirées d'enregistreurs autonomes (EA) présentaient deux états de détection, « confirmé » et « non-confirmé »; la confirmation reposait sur l'accord de deux interprètes et la simulation de performance du MEMD utilisant ce type de données issues d'EA. Nous avons constaté que le MEMD peut corriger correctement les erreurs liées aux FN et aux FP dans une grande étendue de taux d'observation et de taux de détection (d), et est approprié pour les données présentant des « détections confirmées ». Nous avons élaboré un modèle de classification binaire assignant le risque de biais de jeux d'observations fondé sur les paramètres de relevé et de modèle, et avons trouvé qu'un faible risque de biais ne peut pas être prédit par une seule variable ou valeur, mais se présente plutôt sous certaines combinaisons faiblement naïves du taux d'occupation ($< \sim 0.2$), du taux de détection ($< \sim 0.2$), du nombre de mentions confirmées ($< \sim 20$) et du taux élevé de FP ($> \sim 0.07$). Notre approche d'interprétation de données d'EA couplée à nos recommandations d'analyse devraient contribuer à réduire l'inflation possible de ψ résultant d'erreurs liées aux FP.

Key Words: ARU; false negative; false positive; misclassification; occupancy modeling; songbird survey; survey design

INTRODUCTION

Ecologists and conservation biologists often rely upon occupancy or count based surveys to monitor species population status and trends, determine habitat associations, and examine species responses to environmental change (MacKenzie et al. 2006). In bird studies these surveys are commonly based largely upon the detection of acoustic cues (Matsuoka et al. 2014), which often account for greater than 90% of the detections in the data (e.g., Alquezar and Machado 2015). In particular, point count surveys (during which all birds are recorded that are heard or seen during a specified duration of time) are one of the most frequently used survey techniques in ornithology (Matsuoka et al. 2014), and autonomous recording units (ARUs) are increasingly being used for such studies (Shonfield and Bayne 2017). Despite their common use, there is fairly widespread recognition that point counts are prone to detection errors and there has been a concerted effort to adopt field and statistical methods to deal with false absences in point count data (e.g., MacKenzie et al. 2002, 2006, Royle and Link 2006). More recently increased recognition that species misidentifications are also relatively common in acoustic survey data (Simons et al. 2007, 2009, Farmer et al. 2012, Miller et al. 2012, Chambert et al. 2015, Banner et al. 2018) has prompted the development of statistical and sampling methods to account for misidentification (false positives; Royle and Link 2006, Miller et al. 2011, 2012, Ferguson et al. 2015, Clement 2016, Banner et al. 2018).

Both false negative (FN) and false positive (FP) error lead to misclassification of site occupancy (Nichols 2019), where FN leads to a misclassification of the site as unoccupied, and FP error leads to both misclassification of the site as unoccupied for the species that was misidentified, and occupied for the species for which it was incorrectly identified. These are important errors, and accurate estimation of site occupancy (ψ) in the presence of both FP and FN errors requires additional data beyond those normally collected in occupancy studies (Miller et al. 2011, 2012, 2013). Specifically, additional data are required to allow explicit modeling of the FP detection process (Miller et al. 2011, 2012, 2013, Chambert et al. 2015). Occupancy models designed to account for species identification error hold promise for correcting biases in point count surveys (Royle and Link 2006, Simons et al. 2007, 2009, Miller et al. 2011, Farmer et al. 2012, Ferguson et al. 2015), although there is still discussion about the best approach. For example, Ferguson et al. (2015) discuss estimation issues related to both the Miller et al. (2011) and Royle and Link (2006) models, and suggest further testing is required. The multiple detection state model (MDSM) of Miller et al. (2011) relies on confirmation of species occupancy for all or a subset of sites using more intensive surveys and is the approach used for site-level confirmation designs (Chambert et al. 2015). The MDS model assigns observations into 2-states, “certain detections” and “uncertain detections.” This approach allows the explicit assignment of data into different detection states (0 = no detection, 1 = uncertain, and 2 = certain), and is potentially appropriate for ARU-based songbird studies. This approach does not conform strictly to the MDS model requiring “certain detection,” so the validity of applying the MDS model where identification certainty is based on confirmation by independent interpreters requires investigation of bias through simulation of known occupancy, detection, and FP rates. It is also necessary to

test for bias in the model estimates across a broad range of observation, FP, and detection rates.

Multiple visits and more intensive methods to account for FN and FP errors, respectively impose additional logistical sampling constraints and costs. Furthermore, recent design optimization studies by Clement (2016) suggest survey/analysis strategies that can efficiently account for both FP and FN error using additional surveys, but if FP errors are not present they will unnecessarily increase cost. Thus, efficient estimation of FP and FN error will have important implications for optimal species occupancy monitoring. The relatively recent advent of commercially available programmable ARUs provide the opportunity to obtain sufficient numbers of repeat samples while only requiring field staff to physically access a site twice, i.e., deployment and retrieval of the ARUs, and thus have relatively low marginal costs for adding additional samples, i.e., recordings (Rempel et al. 2014). In addition, recordings from ARUs can be archived, allowing confirmation of identifications through use of multiple interpreters, and can provide additional confirmation aids such as spectrograms and digital signatures (Hobson et al. 2002, Acevedo et al. 2009, Rempel et al. 2014, Knight et al. 2017). Such recordings are conducive to implementing statistical methods to simultaneously model both FN and FP detection errors (Chambert et al. 2015).

In our study we first simulated songbird data that would be collected from an ARU and interpreted by two people listening to the same species song/call, then mimicked the introduction of detection error, misidentification error, and species confirmation. We then evaluated the validity of the MDS model for this type of data by evaluating bias and precision across a range of model estimates for occupancy (ψ), detection rate (d), and FP rate. We then investigated the relative effects of correcting for only FN error in modeling occupancy (ψ) versus correcting for both FN and FP error using ARU recordings of forest birds in three ecoregions across northern Ontario. We explored the relationship between the % change in modeled occupancy rates across survey conditions as a function of detection rate, number of confirmed observations, and naïve occupancy rate. Finally, we discuss ARU-based monitoring and analysis strategies for cost-effectively correcting for these errors.

METHODS

Study area

We conducted our study in the Boreal Shield ecozone of Ontario, Canada (Fig. 1). Study area 1 (SA1) lies within the Lake Abitibi ecoregion (ecoregion 3E) of eastern Ontario whereas study areas 2 (SA2) and 3 (SA3) are both partly in the Lake Wabigoon ecoregion with SA2 also straddling the Lake St. Joseph ecoregion (ecoregion 4S3S) and study area 3 (SA3) straddling the Pigeon River ecoregion (ecoregion 4W; Crins et al. 2009). SA3 and the southern portions of SA 1 and SA2 fall within Bird Conservation Region (BCR) 12 (Boreal Hardwood Transition), the northern portions of SA1 and SA2 fall within BCR 8 (Boreal Softwood Shield; North American Bird Conservation Initiative 2000). SA2 and SA3 are generally drier with more extensive areas of exposed bedrock while SA1 has extensive wet, low-lying areas with organic soils and lowland forest species, as well as a more variable fire cycle and more precipitation (Crins et al. 2009). The forest cover

of all study areas is dominated by mixed, coniferous and sparse forest (more sparse forest in SA2) with some deciduous forest, mainly white spruce (*Picea glauca*), black spruce (*Picea mariana*), jack pine (*Pinus banksiana*), balsam fir (*Abies balsamea*), tamarack (*Larix laricina*), eastern white cedar (*Thuja occidentalis*), trembling aspen (*Populus tremuloides*), balsam poplar (*Populus balsamifera*), and white birch (*Betula papyrifera*); SA3 has a mixture of Boreal and Great Lakes-St. Lawrence Forest tree species, e.g., red pine (*Pinus resinosa*), white pine (*Pinus strobus*), red maple (*Acer rubrum*), yellow birch (*Betula alleghaniensis*), black ash (*Fraxinus nigra*), white elm (*Ulmus americana*), and American beech (*Fagus americana*). SA1 was sampled in 2013, SA2 in 2014, and SA3 in 2015.

Fig. 1. Study area 1 (SA1, n = 117) in the Lake Abitibi ecoregion (within ecoregion 3E), study area 2 (SA2, n = 115) in northwestern Ontario straddling both the Lake Wabigoon and Lake St. Joseph ecoregions (within ecoregion 4S3S), and study area 3 (SA3, n = 114) in northwestern Ontario straddling both the Lake Wabigoon and Pigeon River ecoregions.

Field data

Audio recordings were collected using a repeat sampling design. Automated recording units were scheduled to capture five 10-minute recordings per day including three times during the dawn chorus (half an hour before sunrise, at sunrise, and half an hour after sunrise) and twice in the evening (half an hour before sunset and one hour after sunset) within each study area. A combination of Song Meter™ SM1, SM2, and SM2+ recording units, set as similarly as possible (at 22050Hz, 45dB microphone gain and low pass filter at ~160Hz for SM1 and +48dB gain and ~180Hz low pass filter for SM2/2+), were used to capture recordings in Wildlife Acoustics Audio Compression (.WAC) format. Recordings were selected for interpretation across a 4- to 6-day window within the songbird breeding season (late May to early July) assuming that no new territories would be established or abandoned within the analysis period and that the birds present and vocalizing were within their home range. We randomly selected six good quality

(i.e. little wind, rain, or traffic noise) dawn chorus recordings for each site and two additional evening recordings for the wetland sites only (one from each evening time period) for a total of 117, 115, and 114 sites in SA1, SA2 and SA3, respectively. Each recording was independently interpreted twice; interpreter A processed all of the recordings for all study areas, while the duplicate interpretation of recordings in SA1 was split between two interpreters (B₁ and B₂) and duplicate interpretation for SA2 and SA3 was completed by a single interpreter (B₁). Because interpreters and their respective identification skills differed between the two study areas, we treated models independently of study area, and did not average model results by species across study areas. We term these groupings of species by study area “observation sets.”

We instructed interpreters to listen to 10-minute recordings and record all detectable and identifiable species. If species identification for a particular audio cue was uncertain, the interpreter was instructed not to record that instance but instead continue with the interpretation, assuming that if a species truly occupies a site then it is likely it would be detected again at another time during the recording sequence.

Interpretations

FP modeling requires assignment of certain detections, i.e., observations where there is little doubt that a true detection occurred given the site was occupied. Identification was based on interpretation of the recorded acoustic cues, and in the MDS model we used interpretation from a second interpreter to “confirm” species presence at a site. Species presence was assigned a “0” if not recorded, a “1” if only the primary interpreter (A) recorded an observation and a “2” if the both the primary and secondary interpreters (B) confirmed hearing that species. This analysis was repeated so that the secondary interpreter (B) was treated as the primary observer. For each site one interpreter was randomly selected as the primary observer, and those results used to calculate naïve occupancy estimates and to estimate model-averaged ψ . We used this approach to avoid double counting where interpreters did not agree. For example, if interpreter A was selected as primary observer for the site, and identified ALFL, BWWA, and CHSP, whereas interpreter B identified ALFL, BWWA and DEJU, then this would indicate the presence of 3 species with ALFL and BWWA confirmed, and the third species, CHSP, as unconfirmed. Alternatively, combining results from the two interpreters rather than assigning a primary interpreter would incorrectly indicate the presence of four species, with ALFL and BWWA as confirmed, but now with two species, CHSP and DEJU, as unconfirmed.

The MDS model requires a subset of observations to be certain, but unless the bird is in-hand the identification is always subject to error, so although confirmation by two interpreters does not guarantee certain identification it can be used as an estimate. Our approach of using 2-interpreter confirmation to establish certain detection may introduce bias in the model, so we evaluated potential bias using simulated observations that reflect that process by which error is introduced into empirical data through interpretation of recorded audio cues. We simulated true occupancy, FN detection error, and FP identification error (see Appendices 1 and 2 for spreadsheets with simulation code). Based on analysis of our empirical data, we found detection history from

interpreters 1 and 2 were correlated with an R of 0.88, so we simulated this level of correlation when introducing the FN detection error, i.e., species was not detected in a recording because of an unclear audio cue, even though the species occupies the site. We then introduced FP identification error where a species was detected in a recording even though it did not occupy the site, and where misidentification is related to clarity of the audio cue. Conceptually, an audio cue may be unclear because of ambient background noise, it is a rare variant of song used by the species, the volume of the song is very low, or the recording only caught a piece of the full song.

Occupancy modeling

Correction for FP error was modeled using the MDS model for 2 detection states (Hines 2006, Miller et al. 2011), which corrects ψ estimates for both detection and misidentification error (where the species is detected but the site is unoccupied). Occupancy analyses were completed in R (v.3.4.2) using the package unmarked (v.0.12-2; Fiske and Chandler 2011) using the occuFP function. Data was coded as 0, 1, and 2 (type 3 data only in occuFP) to represent no detection, uncertain detection, and confirmed detection, respectively, triggering the model to run the multiple detection state model, creating estimates for occupancy (ψ), true positive detection probability (model parameter $p11$, hereafter termed d), false positive detection probability (model parameter $p10$, hereafter termed FP), and the probability a true positive detection was designated as confirmed (model parameter b).

For field data we generated three estimates of site occupancy: the raw naïve estimate, estimate from the model accounting for FN error only (occu function in the unmarked R package), and estimate from the model accounting for both FN+FP error (MDSM). Based on previous studies (Rempel et al. 2014) we selected four detection covariates that we hypothesized would influence diurnal, seasonal, noise, and sound transmission effects on detection rate: time since sunrise/sunset (TSSR), days since spring (15 May; DSS), recording quality (RQ), and percent hardwood (PHW), respectively; RQ was ordinal, and varied from 1 to 4 (with 1 representing the lowest and 4 representing the highest quality recordings, respectively). These detection covariates were included because some birds adjust singing frequency as the morning progresses (TSSR; Sólymos et al. 2018), some birds adjust singing frequency as the season progresses (DSS; Sólymos et al. 2018), sound can be impeded in stands with heavy deciduous cover (PHW; Yip et al. 2017), and the interpreter can miss hearing a song where heavy background noise occurs in the recording (RQ). AIC model selection was used to select the best supported model from our a priori detection hypotheses (Appendix 3). All covariates were modeled as linear functions.

For simulated data we evaluated bias and precision of model estimates using assigned levels of true occupancy rate, detection rate (d), FP rate, and probability of confirmed detection (b) that reflected the range of conditions found in our empirical data. We also compared three versus six repeat recordings (K). For the simulated data we evaluated bias and precision by producing side by side box plots to compare the true values to the modeled estimates, and by calculating % bias as:

$$bias = \left(\frac{\psi - \text{true occupancy}}{\text{true occupancy}} \right) \quad (1)$$

where ψ is the estimated occupancy under the MDS model, and true occupancy is based on the known simulated occupancy values.

For our field recordings we derived ψ estimates for a total of 166 observation sets, comprising 64 species across the 3 study areas (54, 54, and 56 unique species in SA1, SA2, and SA3, respectively). Two of the 64 species did not have sufficient data in some study areas to model ψ and were removed from the analysis. For field data we calculated % change in the occupancy estimate as:

$$\% \text{ change} = 100 * \left(\frac{\psi - \text{naïve}}{\text{naïve}} \right) \quad (2)$$

where ψ is the occupancy estimate under the MDS model and naïve is the uncorrected occupancy rate. After graphical inspection of % change versus detection rate (d), it was clear that a major change in slope and variance occurred near the x-axis origin, so we fit piecewise (segmented) regression (Muggeo 2003), using the R package Segmented (Muggeo 2008), to determine if the relationship between % change and d showed evidence for breakpoints, and to estimate separate (segmented) slopes if confirmed. Some % change values were > 1000 , so % change was first scaled to its maximum value using the R function rangeScale, and then transformed using the arcsine of the square root.

Estimation of false positive rates from an online survey

To better understand the frequency with which interpreter error occurs among a broad range of interpreter skill levels, an online identification survey was created to document FP rates among species and interpreter expertise. Respondents were asked to identify the correct species using only acoustic cues, where for each of the 20 questions, interpreters were asked to identify the species from a multiple choice (four to five similar sounding species) answers set. We only used verified song clips obtained from the Macaulay Library at the Cornell Lab of Ornithology in our identification quiz. We solicited 225 participants for the survey to sample a wide range in years of experience and skill in identifying species by audio cues. For each audio interpretation question in the survey, the respondents were asked to make their best educated guess as to the identity of the species in the recording; however, if they responded that “he or she was unable to make an educated guess of the species identity,” then this was coded as a FN, and excluded from the estimation of FP rate. If the respondent selected the correct species, this was coded as true positive (TP). Identification accuracy was hypothesized to be a function of experience and ability, so we created three categorical variables representing answers to the yes/no questions “have you had greater than 5 years’ experience identifying birds by song,” “have you had your interpretation ability tested,” and “do you believe that your interpretation ability is sufficient for providing valid identifications” that were treated as categorical factors. We also asked the respondents to rate their confidence for identifying species commonly found in the ecoregions associated with the test (Boreal and Great-Lakes St. Lawrence) from 1 to 5, and this variable was treated as continuous covariate. Models were estimated using the Generalized Linear Models procedure in SPSS, with the Poisson loglinear option selected.

Optimizing the number of listeners needed to estimate FP rate

Similar to determining the optimal number of repeat visits (K) needed to achieve a desired probability of detecting a species given its presence ($p^* = 1 - (1-p)^K$), we can reformulate this expression and solve for the number of interpreters that should be used to reach a desired probability of “confirmed detection” (b^*), where $b^* = 1 - (1-b)^L$ (i.e., the probability of a confirmed detection given L additional interpreters), L is the number of additional interpreters, and b is the probability that a detection is classified as confirmed given that the site is occupied and the species was detected. Note that where only one additional interpreter is used (i.e., $L = 1$), $b^* = b$, and our “confirmed detection” is an approximation of “certain detection.” We can then reformulate the equation for b^* to solve for L under a target value of b (e.g., Sewell et al. 2012), as follows:

$$L = \frac{\log(1-b^*)}{\log(1-b)} \quad (3)$$

where L is the number of interpreters needed to achieve a target b^* , and b is the estimated probability of “confirmed detection” from the occupancy model correcting for both false positives and false negatives. The parameter b was calculated using the occuFP routine in the package unmarked. Here, we estimate L by setting b^* to achieve an 85% chance that species identification at a site will be “confirmed” by the primary interpreter using L additional interpreters.

RESULTS

Simulated recordings

Our simulations revealed that the model estimates from the MDS model had low bias, with estimated ψ and associated variance similar to true (known) occupancy rates across a broad range of occupancy and FP rates (Fig. 2 C-F; Table 1). As well, across this same range the model accurately estimated the survey parameters d , FP rate, probability of “confirmed detection” (b), and associated variances. However, at low occupancy of 0.2, and with higher FP ($p10$) rates of about ≥ 0.02 , the model produced biased estimates of ψ , with modeled ψ about twice the true occupancy rate (Fig. 2 A and B). Slightly biased results also occurred with an occupancy rate of 0.5 but where FP was higher at 0.07 (Fig. 3 D); however, bias disappeared at the lower FP rate of 0.02 (Fig. 3 C). For all other simulations where true occupancy was ≥ 0.2 , and where FP was < 0.02 the modeled estimates were unbiased (Table 1, Appendix 4). The other survey parameters, d , FP, and b were relatively unbiased across the full spectrum of simulation conditions.

In addition to effects from FP and occupancy rate, we also found a strong effect of detection rate on bias and precision of ψ , where at low detection rates ($d = 0.1$) estimates of ψ that were orders of magnitude higher than true occupancy rate (Fig. 3 A, C and E). Bias was also found with a d of 0.2 where true occupancy rate was only 0.1, but at this same detection rate bias largely disappeared when true occupancy increased to 0.5 (Fig. 3 F; Appendix 4)

Fig. 2. Comparison of parameter values for the known simulation versus estimated values, where true occupancy rate (OR) was 0.2 or 0.5, and false positive (FP) rate from about 0.005 to 0.07. Horizontal lines are mean values, and boxes represent the interquartile range; true OR is based on the known simulation data, ψ is the estimate from the multiple detection state model (MDSM), and b is probability of certain detection. Model is unbiased when true occupancy overlaps with ψ .

Fig. 3. Comparison of known parameter values for the simulation versus estimated values. Low detection rates (d) of 0.1 and 0.2 were evaluated at occupancy rates (OR) of 0.1, 0.2, and 0.5. Horizontal lines are mean values, and boxes represent the interquartile range; true OR is based on the known simulation data, ψ is the estimate from the multiple detection state model (MDSM), and b is probability of certain detection. Model is unbiased when true occupancy overlaps with ψ .

Table 1. Simulation parameters for true (known) condition and model (MDSM) estimates.

True Occ.	Model $\Psi^{\dagger}$	STD	% Bias	Bias Class [‡]	Naïve	True $d^{\S}$	Model d	True FP	Model FP	True $b^{\P}$	Model b	# confirmed
0.500	0.499	0.038	-0.11	0	0.489	0.398	0.399	0.005	0.004	0.639	0.639	152.4
0.798	0.795	0.034	-0.35	0	0.765	0.403	0.399	0.002	0.009	0.648	0.653	248.7
0.801	0.807	0.039	0.77	0	0.787	0.402	0.400	0.021	0.016	0.645	0.646	250.3
0.800	0.793	0.037	-0.97	0	0.769	0.400	0.403	0.007	0.011	0.644	0.649	248.6
0.503	0.497	0.037	-1.22	0	0.485	0.410	0.414	0.001	0.003	0.647	0.648	159.6
0.200	0.207	0.029	3.45	0	0.221	0.402	0.393	0.006	0.005	0.627	0.633	61.9
0.794	0.827	0.045	4.13	0	0.828	0.399	0.391	0.069	0.044	0.646	0.649	251.3
0.495	0.521	0.045	5.20	0	0.543	0.403	0.391	0.024	0.018	0.640	0.647	157.8
0.507	0.554	0.093	9.20	0	0.465	0.227	0.215	0.024	0.014	0.519	0.527	74.0
0.500	0.605	0.041	21.0	0	0.641	0.399	0.361	0.066	0.036	0.645	0.649	169.9
0.196	0.241	0.040	22.6	0	0.292	0.405	0.360	0.023	0.016	0.654	0.652	67.1
0.099	0.417	0.179	322.8	0	0.191	0.227	0.116	0.023	0.006	0.529	0.394	20.2
0.203	0.204	0.030	0.56	1	0.199	0.400	0.406	0.001	0.001	0.642	0.660	65.2
0.506	0.713	0.195	40.95	1	0.315	0.109	0.088	0.022	0.054	0.405	0.418	29.0
0.502	0.713	0.195	42.1	1	0.128	0.023	0.088	0.022	0.054	0.250	0.418	6.8
0.210	0.364	0.079	73.5	1	0.267	0.231	0.171	0.023	0.006	0.521	0.469	33.8
0.198	0.398	0.064	101.4	1	0.467	0.404	0.286	0.069	0.037	0.639	0.626	84.3
0.202	0.685	0.238	239.7	1	0.199	0.106	0.061	0.023	0.090	0.393	0.377	15.6
0.202	0.718	0.281	256.4	1	0.128	0.021	0.029	0.023	0.015	0.211	0.418	7.0
0.099	0.616	0.262	522.7	1	0.166	0.115	0.047	0.022	0.014	0.402	0.447	11.3
0.096	0.665	0.256	595.4	1	0.128	0.021	0.023	0.023	0.029	0.252	0.578	6.9

[†] Ψ is probability that the area is occupied;

[‡] % bias is % difference of the modelled Ψ from true occupancy;

[§] d ($p11_i$ in model output) is true positive detection probability (detection rate);

^{||} FP ($p10_i$ in model output) is false positive detection probability;

[¶] b is probability a true positive detection was designated as confirmed.

With the number of repeat observations (“recordings”) set at six, the estimates of variance were unbiased across this broad range of conditions, with box plots revealing a similar range in the true variance of the simulated data and the estimated variance produced by the models. To explore the effect of number of repeat observations (recordings) on bias and precision, we compared simulations using three versus six repeat observations, and FP rates of 2.4% versus 6.7% (keeping occupancy rate at 0.5 and detection rate at 0.4). Under both FP rates we found little effect on magnitude of bias in Ψ , but the variance of the estimate increased as repeat observations decreased (Fig. 4).

As true or naïve occupancy rate decreases, the number of confirmed observations also decreases, and this depression in sample size possibly contributes to poor model performance; thus, it appears from our simulation that bias is related to naïve occupancy rate, number of confirmed observations, d , and FP rates. To help assign risk of bias to our field evaluation data we developed a simple empirical logistic regression equation to relate these four variables to known bias in the simulation results, where the moderate/high risk class was defined as bias $\geq 40\%$ (Table 1).

$$\text{High risk of bias} = \text{naïve} * -295.8 + d * -359.1 + FP * 2377 + NC * 0.532 + 125.0 \quad (4)$$

This simple model had 100% classification accuracy for assigning observations in Table 1 to the associated high/low bias category (with $p < 0.001$, omnibus test of model coefficients). This model was then applied to the field data to assign a category of low versus high risk of bias.

Fig. 4. Comparison of parameter values for the simulation versus estimated values from the multiple detection state model (MDSM) where three versus six repeat samples (recordings) are compared at false positive (FP) rates of 0.02 and 0.07 (with occupancy rate [OR] = 0.5 and $d = 0.4$). Horizontal lines are mean values, and boxes the interquartile range; true OR is based on the known simulation data, Ψ is the estimate from the MDSM, and b is probability of confirmed detection. Model is unbiased when true occupancy overlaps with Ψ .

Table 2. Model estimates, survey parameters, and risk of bias for observation sets (species by study area). Species codes found in Appendix 1.

Species	Study area #	Risk of bias group [†]	% change naïve to ψ	Naïve occ.	ψ (SE)	d	FP	b	# conf. rec.	$L^{\ddagger}$
ALFL	1	0	-10.37	0.402	0.360 (0.044)	0.689	0.011	0.833	145	1
ALFL	3	0	-10.18	0.294	0.264 (0.043)	0.623	0.008	0.841	102	1
AMBI	1	0	7.80	0.179	0.193 (0.045)	0.334	0.004	0.587	25	2
AMRO	1	0	-17.18	0.598	0.495 (0.049)	0.503	0.045	0.812	142	1
AMRO	2	0	-19.22	0.304	0.246 (0.041)	0.474	0.013	0.802	64	1
AMRO	3	0	-9.38	0.560	0.507 (0.054)	0.416	0.025	0.714	104	2
BAWW	1	0	-23.23	0.368	0.282 (0.046)	0.390	0.025	0.868	67	1
BAWW	2	0	-12.44	0.261	0.228 (0.043)	0.442	0.009	0.676	46	2
BAWW	3	0	-5.83	0.495	0.467 (0.054)	0.429	0.014	0.764	109	1
BBWA	1	0	-18.46	0.214	0.174 (0.037)	0.632	0.009	0.631	51	2
BCCH	1	0	8.62	0.342	0.371 (0.062)	0.289	0.009	0.838	68	1
BHVI	2	0	-9.81	0.304	0.274 (0.050)	0.283	0.014	0.846	48	1
BLBW	2	0	-21.75	0.270	0.211 (0.042)	0.432	0.015	0.688	45	2
BLBW	3	0	-22.71	0.349	0.269 (0.043)	0.544	0.023	0.937	93	1
BRCR	1	0	2.68	0.111	0.114 (0.032)	0.354	0.001	0.881	25	1
BTNW	3	0	1.48	0.138	0.140 (0.036)	0.382	0.001	0.850	31	1
CAWA	1	0	-26.37	0.137	0.101 (0.029)	0.415	0.007	0.853	25	1
CAWA	2	0	-71.41	0.061	0.017 (0.012)	0.796	0.008	0.875	7	1
CAWA	3	0	-6.80	0.174	0.162 (0.036)	0.466	0.003	0.724	37	1
COYE	1	0	12.92	0.359	0.405 (0.059)	0.440	0.013	0.801	111	1
COYE	2	0	-8.13	0.052	0.048 (0.021)	0.377	0.001	0.928	13	1
COYE	3	0	16.19	0.339	0.394 (0.064)	0.384	0.003	0.766	97	1
CSWA	1	0	-19.34	0.154	0.124 (0.036)	0.419	0.007	0.486	17	3
DEJU	2	0	-14.12	0.591	0.508 (0.052)	0.446	0.040	0.772	130	1
DEJU	3	0	0.28	0.248	0.248 (0.049)	0.344	0.004	0.525	31	3
GCKI	1	0	-6.74	0.624	0.582 (0.048)	0.507	0.022	0.885	182	1
GCKI	2	0	-14.95	0.435	0.370 (0.046)	0.508	0.019	0.821	111	1
GCKI	3	0	-2.91	0.275	0.267 (0.047)	0.403	0.005	0.537	41	2
HETH	2	0	-16.96	0.530	0.440 (0.049)	0.567	0.038	0.824	157	1
HETH	3	0	-7.44	0.596	0.552 (0.052)	0.576	0.024	0.825	191	1
LEFL	1	0	-32.97	0.154	0.103 (0.031)	0.553	0.011	0.700	28	2
LEFL	2	0	-16.51	0.148	0.123 (0.031)	0.589	0.006	0.945	53	1
LEFL	3	0	-4.15	0.211	0.202 (0.039)	0.689	0.002	0.859	85	1
LISP	1	0	-17.47	0.137	0.113 (0.030)	0.595	0.005	0.721	34	1
LISP	2	0	-9.02	0.209	0.190 (0.038)	0.568	0.005	0.893	75	1
MAWA	1	0	-4.64	0.769	0.734 (0.044)	0.612	0.037	0.837	268	1
MAWA	2	0	-10.36	0.557	0.499 (0.047)	0.636	0.018	0.837	196	1
MAWA	3	0	-12.36	0.596	0.523 (0.049)	0.586	0.028	0.738	154	1
MOWA	1	0	-7.27	0.222	0.206 (0.039)	0.507	0.005	0.753	56	1
MOWA	2	0	-9.96	0.322	0.290 (0.043)	0.554	0.008	0.840	98	1
MOWA	3	0	-6.72	0.422	0.394 (0.057)	0.436	0.024	0.789	100	1
NAWA	3	0	-0.80	0.899	0.892 (0.030)	0.673	0.013	0.833	349	1
NOPA	2	0	-21.75	0.078	0.061 (0.022)	0.723	0.003	0.909	20	1
NOPA	3	0	-15.23	0.294	0.249 (0.045)	0.378	0.013	0.921	63	1
NOWA	1	0	-8.42	0.137	0.125 (0.032)	0.466	0.003	0.822	33	1
OSFL	2	0	-18.16	0.209	0.171 (0.036)	0.520	0.007	0.694	44	2
OSFL	3	0	-32.64	0.073	0.049 (0.022)	0.418	0.004	0.789	12	1
OVEN	1	0	-8.18	0.504	0.463 (0.046)	0.691	0.016	0.949	226	1
OVEN	2	0	-7.66	0.330	0.305 (0.043)	0.654	0.006	0.955	148	1
OVEN	3	0	-5.72	0.771	0.727 (0.043)	0.708	0.026	0.924	341	1
PAWA	1	0	16.83	0.145	0.170 (0.043)	0.364	0.002	0.756	36	1
PAWA	2	0	-12.48	0.122	0.107 (0.034)	0.354	0.005	0.658	19	2
PHVI	3	0	-22.71	0.073	0.057 (0.023)	0.585	0.003	0.555	10	2
RBNU	3	0	5.02	0.440	0.462 (0.061)	0.299	0.009	0.867	82	1
RCKI	1	0	-6.31	0.453	0.424 (0.049)	0.462	0.013	0.812	116	1
RCKI	2	0	2.01	0.574	0.585 (0.052)	0.441	0.007	0.867	168	1
RCKI	3	0	-17.21	0.339	0.281 (0.048)	0.453	0.019	0.829	77	1

(con'd)

REVI	1	0	-6.95	0.761	0.708 (0.048)	0.622	0.050	0.803	252	1
REVI	2	0	-16.56	0.452	0.377 (0.046)	0.550	0.025	0.897	144	1
REVI	3	0	-10.25	0.734	0.659 (0.046)	0.717	0.057	0.837	273	1
RWBL	3	0	-25.26	0.064	0.048 (0.024)	0.530	0.004	0.865	14	1
SOSP	3	0	-18.26	0.183	0.150 (0.035)	0.546	0.007	0.775	49	1
SWSP	1	0	-8.28	0.179	0.165 (0.035)	0.625	0.003	0.651	47	2
SWSP	3	0	-3.08	0.156	0.151 (0.036)	0.549	0.002	0.855	65	1
SWTH	1	0	-2.14	0.897	0.878 (0.031)	0.645	0.029	0.840	333	1
SWTH	2	0	-5.88	0.739	0.696 (0.044)	0.628	0.033	0.862	277	1
SWTH	3	0	-10.57	0.541	0.484 (0.049)	0.541	0.020	0.820	150	1
TEWA	1	0	4.62	0.231	0.241 (0.050)	0.323	0.010	0.842	46	1
TEWA	2	0	-6.36	0.122	0.114 (0.031)	0.398	0.003	0.882	31	1
VEER	1	0	-5.46	0.248	0.234 (0.041)	0.466	0.004	0.743	57	1
VEER	3	0	-2.40	0.450	0.439 (0.048)	0.525	0.004	0.783	124	1
WIWR	1	0	-5.67	0.581	0.548 (0.047)	0.569	0.016	0.877	193	1
WIWR	3	0	-9.90	0.550	0.496 (0.050)	0.479	0.021	0.865	144	1
WTSP	2	0	-1.78	0.957	0.939 (0.022)	0.792	0.047	0.922	507	1
YBFL	1	0	6.73	0.393	0.420 (0.054)	0.429	0.003	0.773	104	1
YBFL	2	0	-0.36	0.522	0.520 (0.052)	0.447	0.006	0.714	122	2
YBSA	3	0	-14.02	0.321	0.276 (0.055)	0.270	0.018	0.808	42	1
YRWA	2	0	-8.13	0.583	0.535 (0.053)	0.423	0.023	0.596	99	2
YRWA	3	0	-8.18	0.651	0.598 (0.057)	0.476	0.032	0.593	120	2
YWAR	3	0	0.92	0.229	0.231 (0.047)	0.408	0.004	0.108	7	17
BCCH	3	1	35.28	0.248	0.335 (0.089)	0.151	0.007	0.892	31	1
BHVI	1	1	20.65	0.137	0.165 (0.054)	0.168	0.009	0.855	15	1
BLBW	1	1	49.24	0.094	0.140 (0.065)	0.203	0.007	0.501	10	3
BLJA	1	1	154.46	0.128	0.326 (0.203)	0.068	0.006	0.520	8	3
BLJA	3	1	52.69	0.642	0.981 (0.019)	0.176	0.412	0.727	87	1
BOCH	1	1	41.16	0.179	0.253 (0.078)	0.167	0.007	0.890	27	1
BTBW	1	1	261.56	0.120	0.433 (0.121)	0.110	0.010	0.942	33	1
BTBW	3	1	64.72	0.037	0.060 (0.042)	0.260	0.003	0.875	7	1
CEDW	3	1	133.22	0.358	0.834 (0.319)	0.079	0.011	0.635	29	2
CHSP	1	1	271.03	0.222	0.825 (0.040)	0.018	0.291	0.905	11	1
CHSP	2	1	178.62	0.287	0.800 (0.040)	0.030	0.572	0.235	4	7
CHSP	3	1	-5.98	0.138	0.129 (0.051)	0.274	0.012	0.088	2	21
CONI	1	1	164.09	0.103	0.271 (0.096)	0.088	0.003	0.838	12	1
CSWA	3	1	135.02	0.376	0.884 (0.032)	0.126	0.857	0.233	18	7
DEJU	1	1	23.13	0.214	0.263 (0.086)	0.206	0.009	0.182	7	9
GRAJ	1	1	615.95	0.120	0.857 (0.119)	0.013	0.093	0.784	6	1
HAWO	2	1	520.79	0.130	0.810 (0.086)	0.003	0.132	0.507	1	3
HAWO	3	1	428.32	0.183	0.969 (0.030)	0.030	0.316	0.650	13	2
HETH	1	1	-17.53	0.726	0.599 (0.047)	0.607	0.089	0.815	216	1
LISP	3	1	84.39	0.174	0.321 (0.070)	0.228	0.003	0.754	39	1
NAWA	1	1	-6.56	0.906	0.847 (0.035)	0.724	0.106	0.871	381	1
NAWA	2	1	-5.55	0.930	0.879 (0.031)	0.759	0.202	0.925	457	1
NOFL	1	1	426.70	0.188	0.990 (0.010)	0.037	0.458	0.618	16	2
NOFL	2	1	4.46	0.209	0.218 (0.072)	0.145	0.016	0.916	21	1
NOFL	3	1	35.40	0.248	0.335 (0.204)	0.110	0.017	0.523	13	3
OSFL	1	1	62.16	0.085	0.139 (0.052)	0.196	0.003	0.654	14	2
PAWA	3	1	107.16	0.165	0.342 (0.082)	0.158	0.009	0.744	26	1
PISI	2	1	58.50	0.226	0.358 (0.116)	0.101	0.016	0.898	21	1
PIWO	3	1	33.03	0.183	0.244 (0.093)	0.166	0.005	0.413	11	4
RBGR	3	1	566.86	0.138	0.918 (0.031)	0.014	0.357	0.694	6	2
RBNU	1	1	3.60	0.179	0.186 (0.045)	0.252	0.006	0.927	32	1
RBNU	2	1	22.75	0.200	0.245 (0.062)	0.197	0.005	0.909	30	1
RWBL	2	1	-8.61	0.035	0.032 (0.019)	0.258	0.001	0.327	2	5
WIWA	2	1	111.13	0.052	0.110 (0.050)	0.272	0.002	0.544	7	2
WIWA	3	1	1665.67	0.055	0.972 (0.016)	0.004	0.602	0.339	1	5
WTSP	1	1	-4.56	0.872	0.832 (0.035)	0.794	0.082	0.948	448	1
WTSP	3	1	-3.26	0.945	0.914 (0.029)	0.751	0.121	0.903	440	1
YBSA	1	1	63.12	0.316	0.516 (0.113)	0.163	0.020	0.799	52	1
YRWA	1	1	-16.07	0.778	0.653 (0.050)	0.537	0.142	0.777	196	1

† Bias Group 0 is associated with unbiased estimates; Group 1 has lower observation and detection rates and higher false positive (FP) rates, and is associated with conditions that cause biased estimates.

‡ L is the number of interpreters (in addition to the primary interpreter) needed to achieve a target b^* of 85% (probability a true positive detection was designated as confirmed).

Field recordings

Initial analysis of our field recordings revealed a strong breakpoint in % change between our MDS model estimates of ψ and the naïve rate among observation sets with a combination of very low detection rates, occupancy rates, and/or FP rates. To explore these relationships further we used segmented regression analysis and found detection rate explained a significant proportion of the variance in the % change in ψ (adjusted $r^2 = 0.679$), with the regression supporting a breakpoint in the relationship where $d < 0.149$ (SE 0.011; Fig. 5A). As expected the % change in ψ decreased as d increases, suggesting the model is operating properly where d is higher than the 0.149 threshold (Fig. 5A).

Fig. 5. Fit from segmented (piecewise) regression analysis (using arcsine transform of % change between naïve and ψ for (A) detection rate (d), (B) observation rate, i.e., naïve occupancy, and (C) number of confirmed recordings.

Through segmented regression we also found break-points in the relationships between naïve occupancy, i.e., observation rate, and % change in ψ (Fig. 5B), and between number of confirmed observations and % change (Fig. 5C). Naïve occupancy rates < 0.15 and number of confirmed observations < 7.6 were associated

with greatly increased variability in the estimate of ψ ($P < 0.001$ for both models). Evaluation of the simulated data found that bias in the estimate is highest where naïve occupancy < 0.2 (Fig. 3), which is similar to the break-point of 0.15 identified from the field recordings. Low occupancy rates are often associated with a low number of confirmed observations, which may have a direct influence on poor model performance. These results support the simulation assessment that certain combinations of low observation rate, low number of confirmed observations, low d , and low FP result in high risk of bias for ψ , so we applied our simulation-based equation 4 to assign bias risk categories based on combinations of these parameters, where Group 0 had low risk of bias and Group 1 high risk of bias in ψ (Table 2). We assigned Group 2 to those observation data sets where the model was not able to converge or provide valid confidence limits for the estimates. Most Group 2 models also had some combination of low naïve occupancy, number of confirmed observations, and d , and high FP rate.

The mean % change (absolute value) from naïve to ψ for the 80 Group 0 observation sets (low risk of bias) was 11.75% (SE 9.98%; median % change of 9.20%), with about 84% of those changes being negative (Table 3, Fig. 6). The mean (SE) of detection rate (d), FP error rate and certain detection rate (b) were 0.50 (0.12), 0.014 (0.013), and 0.79 (0.13), respectively. As expected, ψ estimates from models accounting for only FN error increased relative to naïve estimates. However, it was less obvious what would happen if both FN and FP errors were modeled using the MDS model, where 84% of ψ estimates were lower than naïve (Fig. 6). Clearly modeling FP errors had a stronger effect on estimate correction than modeling FN errors.

Fig. 6. The relative effect on ψ from modeling FN errors only versus modeling FN and FP errors.

Observation sets in Group 1 were classed as having moderate to high risk of bias, although note that this simply assigns risk and is not a statistical assessment of bias. The % change in ψ for these observation sets was highly variable, skewed, and in some cases extremely large, so we presented only nonparametric median estimates of central tendency. For Group 1 observation sets the median % change (absolute value) from naïve to ψ was 58.5% (Table 3) and ranged from 3.3% to 1666%. Relative to the Group 0 models, these observation sets were associated with a higher false positive rate (FP) and lower detection rate (d), naïve occupancy rate, number of recordings with confirmed observations, and number of sites that were confirmed to have

Table 3. Summary statistics for parameters affecting risk of bias and % change in the occupancy estimate from naïve to the multiple detection state model (MDSM) estimate.

Bias group [†]		% change [‡]	<i>d</i>	FP	Naïve	Confirmed recordings	Recordings with observations	Confirmed sites	Sites with observations
Group 0	average	11.75	0.50	0.014	0.37	105.08	134.53	34.68	41.78
Group 0	SE	9.98	0.12	0.013	0.22	93.45	106.43	24.46	25.42
Group 0	Median	9.20	0.49	0.009	0.33	79.50	99.00	28.50	37.00
Group 1	Median	58.50	0.17	0.016	0.18	16.00	34.00	11.00	21.00
Group 2	Median	72.25	0.14	0.000	0.08	11.00	22.00	5.00	9.00

[†] Bias Group 0 has lowest risk of bias, 1 moderate to high risk of bias, and Group 2 are those observation sets for which models did not converge or had invalid estimates of variance.

[‡] % change is the change in occupancy estimate between modeled and naïve.

the species present (Table 3). For the 47 observation sets in Group 2, where models were unable to converge and/or provide valid confidence limits and standard errors, the median % change (absolute value) from naïve to ψ was 72.5% (Table 3) and ranged from 0.48% to 2699%. Most of these models had very low numbers of confirmed observations, with a median of only 11 confirmed recordings and five confirmed sites, although there may be additional unresolved reasons why the models did not converge or estimate properly. The results from these models were excluded from further evaluation.

The online song identification survey included 225 respondents from across North America, including Alberta, Saskatchewan, Ontario, Quebec, Newfoundland, and northern U.S. States, and provided a broader evaluation of the prevalence of misidentification (FP error) among interpreters. The overall Poisson model was significant (omnibus test of fitted model against intercept only model; LLR $\chi^2_{13} = 169.27$, $P < 0.001$), with four significant factors associated with a higher number of correct answers (TP). Identification confidence with the species group used in the survey, ranging from 1 (not confident at all) to 5 (quite confident) was the strongest predictor ($\text{Exp}(\beta) = 1.151$; $\text{CI} = 0.785\text{--}1.01$; $P < 0.001$). Having greater than five years of experience in identifying birds by sound had the second largest effect size ($\text{Exp}(\beta) = 0.888$, $\text{SE} = 0.0625$) but with P slightly less than significant ($P = 0.059$). Two other factors, having had interpretation ability tested and self-assessing ability as sufficient, were also significant, with $\text{Exp}(\beta) = 0.849$, $P < 0.001$ and $\text{Exp}(\beta) = 0.803$, $P < 0.001$, respectively.

Misidentification rate, i.e., probability the respondent picked the incorrect species, decreased dramatically from the group with the lowest identification confidence (0.749) to the highest (0.308; confidence = 1 and 5, respectively; Fig. 7). Based on these results we created an “expert” subgroup ($N = 28$), classed as those with the highest level (5) of self-assessed confidence. However, even among this expert group misidentification rates could be high for certain species.

We identified confusion groups of species based on the answers selected by the “expert” group in the online survey (Table 4), and found that this group always correctly identified Rose-breasted Grosbeak (*Pheucticus ludovicianus*), but misidentification rates for other species ranged from 0.75 (Yellow Warbler, [*Setophaga petechia*]) to a low of 0.071 (Boreal Chickadee [*Poecile hudsonicus*] and Black-throated Blue Warbler [*Setophaga caerulescens*]; Table

4). Misidentification rates also differed between two American Redstart (*Setophaga ruticilla*) song types, with higher misidentification for the higher pitched song (type 2, syllable pattern more similar to Bay-breasted Warbler [*Setophaga castanea*] or Black-and-white Warbler [*Mniotilta varia*]).

Fig. 7. Relationship between the observers’ self-assessed confidence and misidentification rate for correctly identifying species through the online audio survey. Confidence group 5 was labeled as “expert” ($N = 28$) for subsequent analysis.

From our field study we calculated L to estimate of the minimum number of observers, additional to the primary interpreter that would be required to achieve an 85% probability of confirmed detection (b). For all but two species in our 115 observation sets, L ranged from 1 to 9 (Table 2). Chipping Sparrow (*Spizella passerina*) and Yellow Warbler had low values of b , and would require 21 and 17 additional observers, respectively, to achieve an 85% probability that any detection would be a confirmed detection. Chipping Sparrow, Dark-eyed Junco (*Junco hyemalis*), Yellow Warbler, American Redstart, Chestnut-sided Warbler (*Setophaga pensylvanica*) and Wilson’s Warbler (*Cardellina pusilla*) were among the most inconsistently identified species. Of the 115 observation sets, 90% would require between one and three interpreters (additional to the primary interpreter) to reach 85% probability of confirmed detection, while the remaining would require more than three interpreters (Table 2). Among the most problematic species in the field survey, many of these also had high rates of misidentification in the online survey, including Chestnut-sided Warbler, Yellow Warbler, Dark-eyed Junco, Chipping Sparrow, American Redstart, and Bay-breasted Warbler (Table 4).

Table 4. Misidentification rates[†] for the expert ranked observers based on an online audio survey. Species sp1-sp5 represent the species selected as the correct answer, in order of frequency of selection. Species codes found in Appendix 1.

MisID [†] rate	Correct Answer	Species selected as answer in descending order of frequency				
		sp1	sp2	sp3	sp4	sp5
0.750	YWAR	CSWA	YWAR	MAWA	AMRE	
0.607	DEJU	DEJU	SWSP	PIWA	OCWA	CHSP
0.536	AMRE2 [‡]	AMRE	BBWA	BAWW		
0.536	BLBW	BLBW	GCKI	BBWA		
0.500	REVI	REVI	BHVI	YTVI		
0.393	CHSP	CHSP	DEJU	SWSP	PIWA	OCWA
0.357	GCKI	GCKI	RCKI	BLBW		
0.321	AMRE1 [†]	AMRE	CSWA	YWAR	MAWA	
0.321	CAWA	CAWA	MAWA			
0.286	LEFL	LEFL	YBFL			
0.286	NAWA	NAWA	YRWA	TEWA	WIWA	
0.250	YRWA	YRWA	WIWA	NAWA		
0.179	HETH	HETH	SWTH	WOTH		
0.179	PIWO	PIWO	NOFL			
0.143	BAWW	BAWW	BBWA	BLBW		
0.143	YBSA	YBSA	PIWO	ATTW		
0.107	RBNU	RBNU	WBNU			
0.071	BOCH	BOCH	BCCH			
0.071	BTBW	BTBW	BTNW			
0.000	RBGR	RBGR				

[†] MisID rate is the proportion of times the respondents picked the incorrect species.

[‡] AMRE1 and AMRE2 represent two versions of the AMRE song. AMRE2 is a less common, higher-pitched version more similar to Bay-breasted Warbler (*Setophaga castanea*, BBWA) or Black-and-white Warbler (*Mniotilta varia*, BAWW).

DISCUSSION

Misidentification has been largely ignored as a source of observational error in acoustic based studies, and whereas early occupancy modeling focused on nondetection (FN) error leading to misclassification of a site as being unoccupied, more recent attention has been directed on misidentification (FP) error leading to the misclassification of the state of site occupancy (Nichols 2019). In this study we explored the validity and performance of using an existing model to correct for FP error resulting from misidentification of interpreted songbird recordings that are typical of those collected using ARUs.

Performance of the MDS model to correct for occupancy classification error

Our simulation study evaluated bias and precision of occupancy estimates from using the Miller et al. (2011) MDS model, and our results support the position that this model can effectively correct for both FN and FP errors across a broad range of survey observation and detection rates, and is appropriate for the type of two-interpreter confirmation used in this study. The model also produces unbiased occupancy estimates across the range of three to six repeat recordings, although variance in the estimates increases with only three repeat recordings. In our study we used confirmation by two interpreters to estimate identification “certainty” of observations, and it was not known if this approach was appropriate and met the conditions for valid use of the MDS model. In our simulation of FN and FP error we closely mimic the process by which these errors could be introduced into data sets, and also allowed for situations where identity was confirmed by both interpreters, but where the identification was ultimately incorrect.

There are other models that may also be appropriate for modeling FP error, and further work is needed to evaluate these. Royle and Link (2006) developed the original approach to accounting for false positive detections using a binomial mixture model, but this approach had limitations to its successful implementation (Miller et al. 2011). Cook and Hartley (2018) found that using double and multiple observer occupancy models (Huggins’s closed capture data type) resulted in positively biased occupancy estimates and failed to account for observer error. However, we modeled observer error using the Miller et al. (2011) MDS model and found that under appropriate conditions of observation rate and detection rate this approach corrected for positively biased occupancy estimates by lowering the modeled estimate relative to estimates based on correction of FN error alone.

The MDS model, however, did not perform well under very low observation and detection rate conditions, where there was significant bias and high variability in the estimates. This is a well-known problem among occupancy models; for example, Miller et al. (2011) and McKelvey et al. (2008) found that the largest bias occurred when ψ was low and that as a species becomes rarer, the proportion of FP observations compared to true observations increases. Simulated results reveal that in such cases even the naïve occupancy rate can be significantly biased. We developed a simple model to assign risk of bias to an observation data set based on a combination of naïve occupancy, detection rate, and FP rate. Our analysis, however, also suggested that low number of confirmed observations, not just the rate of unconfirmed observations, may be a more important driver of poor model performance. The relationship between observation rates and model bias needs to be evaluated through more extensive

simulation analysis across a broader range of observation and detection rate conditions.

Relative impact of modeling FN versus FP error

For the field data sets with low risk of bias (Group 0), we found that under the Miller et al. (2011) MDS model, FP error had a greater impact on corrected ψ than FN error. Similar impacts of FP error have been reported by others (McClintock et al. 2010a, Chambert et al. 2015, Ferguson et al. 2015, Clement 2016), suggesting that bias resulting from ignoring even a small number of FP errors can often be very large. There is particular concern regarding observation error when using data collected by the public (McKelvey et al. 2008, Miller et al. 2013); however, results from our online survey provide a coarse evaluation of how experience and ability affect misidentification, and suggest that FP errors occur across a range of expertise, and even with the most skilled and experienced interpreters there was still high possibility of significant FP error and variability between interpreters for certain species. Likewise, Simons et al. (2009) designed a series of field experiments using audio broadcasting to assess the factors affecting detection probabilities on auditory counts and concluded that direct estimates of detection rate, including error related to misidentification, should accompany all analyses of avian point count data because measurement error on auditory point counts is substantial.

As expected, we found corrections for FN always increased ψ relative to the naïve estimate, but that in 84% of the cases corrections that modeled both FP and FN brought ψ lower than the naïve estimate. Although this result may be somewhat study specific, it nonetheless flags the importance of accounting for FP error. After accounting for both FP and FN error, the mean deviation from the uncorrected naïve occupancy rate was about 11.5%. The effect of FP error on ψ was often substantially greater than the effect of FN error. We estimated occupancy from six recordings (i.e. "visits") per site, so FN error is inherently low in our study because of multiple opportunities to identify a species; but if even one audio cue is misinterpreted among the recordings, then an FP error occurs. These results suggest that surveys that correct for FN but not FP error will tend to inflate ψ estimates and introduce significant bias into surveys. There may be an interaction effect between the magnitude of inflation and the number of visits to the site, as the correction for FN error generally decreases with higher numbers of site visits.

Overestimation of species occurrence could have consequences for species conservation; for example, delayed identification of species in decline because listing criteria are based on trends in estimated population size and occupancy (Camaclang et al. 2015, Loehle and Sleep 2015). Thus, overestimating species occurrence risks delaying conservation efforts, potentially leading to longer species recovery and decreased likelihood of success (Taylor et al. 2005).

Improving survey design to reduce misclassification

Thompson et al. (2017) suggests a minimum naïve occupancy rate of > 0.05 for conducting occupancy analysis; however this might be too restrictive in some situations. Given the need for

management/conservation prescriptions for rare species, it is particularly important to account for observational error when occupancy or detection rate is low (McKelvey et al. 2008, McClintock et al. 2010a, Miller et al. 2011). Our model to estimate L , the number of observers required to correct for FP error, could also be considered when selecting focal species for a study using ARU data. However, for species that are important to survey, but where detection and occupancy rates are inherently low or the species has inherently high misidentification rates, changes in the survey protocol may increase detection rate and number of confirmed observations. For example, we found that the time of day and date can affect detection probability, so changes in timing of the survey may increase the opportunities to hear the bird vocalize and identify it accurately.

Stratified sampling to increase the number of sites found in habitat used by rare species or species that vocalize infrequently may increase the number of encounters with the species and provide more opportunities for detection and accurate identification. This could also increase the number of sites confirmed to have the species present. These types of survey efforts may help reduce bias and increase precision in the occupancy estimate. If this is not possible, recognize that even a few misidentifications could inflate the naïve occupancy estimate by orders of magnitude, so exercise prudence in reporting of possibly biased estimates.

Additionally, the improved conditional survey design of Specht et al. (2017) may be useful for cost-effectively generating improved correction factors for both FN error and FP errors. In a conditional survey design applicable to interpretation of ARU recordings, one or two recordings from all sites would initially be interpreted, and if the rare species is detected in these initial recordings by at least one of the two interpreters, then additional recordings are interpreted. Optionally, additional interpreters can be assigned to those sites to improve estimation of the FP error rate. This could similarly be done for species with high rates of species misidentification. Estimates of the optimal number of listeners could be used to select the number of additional interpreters to use, and their efforts could be directed to those sites containing species for which estimates of L were > 2 . To reduce cost, interpreters could generate audio clips of species with high FP error rates, and multiple interpreters or cloud sourcing (Wimmer et al. 2013) could be used to classify only the recording clips. These approaches provide options to cost-effectively improve correction for estimates of FN and FP error by focusing observation effort on those few sites where at least one interpreter has detected rare or problematic species in at least one or two recordings. Banner et al. (2018) have developed an observation confirmation model and workflow that provides flexibility in establishing confirmation at the observation level and could possibly reduce cost and effort required for this type of FP modeling.

Revising training and interpretation protocols to improve identification accuracy

Reducing FP rate may also be accomplished through additional interpreter training, especially for those species identified as having a high FP rate. Note, however, that Miller et al. (2012) found only marginal reduction in errors by using additional instruction or training. Farmer et al. (2012) found that rates of nondetection and FP error vary with species rarity and interpreter

skill, and that interpreter confidence (declaration of certainty) was not a reliable measure of these errors. In our online study, however, we found interpreter confidence a useful predictor of FP error. However, we also found that high misidentification rates occurred for interpreters with years of experience and high confidence, so other approaches such as changes to survey design may be required to sufficiently reduce FP error.

Our interpreters were instructed to only record a species if they felt confident in the identification, as recommended in previous studies (McClintock et al. 2010b, Farmer et al. 2012). Suggestions to reduce errors during data collection include recording detailed detection evidence to indicate observation confidence such as call type, i.e., song versus call (Farmer et al. 2012), and detection method, i.e., visual vs. auditory (Simons et al. 2009, Miller et al. 2011), or by discounting detections beyond specific distance thresholds, depending on species and recording conditions (McClintock et al. 2010b). Additional suggestions for reducing misidentification during recording interpretation include the following: providing knowledge of the habitat in which the recording was made, e.g., field conditions, habitat information, site photos, range maps, etc.; using spectrograms and computer based identification resources (Knight and Bayne 2018); providing examples of species songs; and quizzing potential interpreters to assess their ability (Rempel et al. 2014). In this study FP error was very high for Blue Jay (*Cyanocitta cristata*) and Chestnut-sided Warbler; use of spectrogram aids and digital detection algorithms might be particularly helpful whenever species known or estimated to have high FP rates are encountered. Given the high potential for FP error, study design should include a request for song “type or voucher specimens” so errors can be corrected in later years.

Exploratory approaches to reducing error

Approaches to correcting FP error that include nonauditory data or experimental control data might also be considered. Chambert et al. (2015) devised an occupancy modeling approach that involves confirmation of individual observations using additional biological samples, e.g., DNA, or indirect records such as photographs or acoustic recordings from which identification can be confirmed. Alternatively, Chambert et al. (2015) developed a calibration design where they used controlled experiments to evaluate the detection process and use these data to model the FP detection process within the occupancy model. Van Wilgenburg et al. (2017) integrated the use of automated recording units and human observer point count data to take advantage of the relative merits of each survey type to reduce bias in estimated avian counts and/or density; associated recordings from this approach could also be used to estimate FP rates for human observer counts. Guillera-Aroita et al. (2017) demonstrated a two-stage model requiring at least two sources of extra information, e.g., records from a survey method with no possibility of false detections and a calibration experiment, to reliably estimate occupancy. These authors also developed a method within a Bayesian approach to set bounds on false detections rates based on prior knowledge, and we suggest that data from experiments such as our online interpretation experiment would provide a convenient framework for deriving such informative priors. Recently, progress has been made on using various classification and machine learning approaches for recognition of animal sounds, including Hidden

Markov Models, Neural Networks, Deep Learning, and Support Vector Machines (e.g., Cai et al. 2007, Ranjard and Ross 2008, Weninger and Schuller 2011, Stowell et al. 2019). Although automated recognizers are typically applied to an entire recording, often with less than impressive results (e.g., Venier et al. 2017), recognizers could perhaps be more effectively applied to isolated song clips to provide additional evidence for songbird identity, and this may be particularly useful when using ARU based surveys to cost-effectively reduce misidentification error. Indeed, classification scores (Knight and Bayne 2018) from automated classification algorithms could be used as aids to help interpreters decide between alternative species or songs, and spectrogram based tools for visualizing and listening to recordings that suggest possible species identifications are in development (E. M. Bayne, *personal communication*).

Operational and cost considerations

Study objectives and conditions will of course vary, but the cost for modeling FP error in addition to FN error from data collected using ARUs will in many cases be focused on funding additional interpretations rather than sampling additional sites or collecting more repeat recordings. As discussed above, the quality and effectiveness of our two-observer MDS model design could be improved by adding a third interpreter to confirm identification in only those sites where the initial two interpreters disagreed, and may improve the approximation of “certain identification” required by the MDS model. Regardless of approach, additional resources to reveal major identification problems and to correct for FP error (in addition to FN) should be incorporated into study designs. Identifying and correcting species misidentification should reduce the potential introduction of bias into estimate of ψ that could otherwise impede conservation efforts.

In this study we have evaluated a framework (MDSM) for modeling FN and FP error, and have provided suggestions for survey design and interpretation protocols for ARU data (including analysis guidelines) that should aid in simultaneous modeling of both FP and FN errors to reduce the potential introduction of bias in the estimates of ψ . We encourage others to employ similar approaches over a broader range of ecological conditions to examine the generality of this proposed approach.

Responses to this article can be read online at:

<http://www.ace-eco.org/issues/responses.php/1374>

Acknowledgments:

We would like to thank Stephen Gullage, George Holborn, and Jeff Robinson for their interpretation of songbird recordings. We want to thank the 225 respondents that took the online survey for bird identification. Also we want to thank the reviewers for very constructive comments.

LITERATURE CITED

Acevedo, M. A., C. J. Corrada-Bravo, H. Corrada-Bravo, L. J. Villanueva-Rivera, and T. M. Aide. 2009. Automated classification of bird and amphibian calls using machine learning:

- a comparison of methods. *Ecological Informatics* 4(4):206-214. <https://doi.org/10.1016/j.ecoinf.2009.06.005>
- Alquezar, R. D., and R. B. Machado. 2015. Comparisons between autonomous acoustic recordings and avian point counts in open woodland savanna. *Wilson Journal of Ornithology* 127 (4):712-723. <https://doi.org/10.1676/14-104.1>
- Banner, K. M., K. M. Irvine, T. J. Rodhouse, W. J. Wright, R. M. Rodriguez, and A. R. Litt. 2018. Improving geographically extensive acoustic survey designs for modeling species occurrence with imperfect detection and misidentification. *Ecology and Evolution* 8:6144-6156. <https://doi.org/10.1002/ece3.4162>
- Cai, J., D. Ee, B. Pham, P. Roe, and J. Zhang. 2007. Sensor network for the monitoring of ecosystem: bird species recognition. Pages 293-298 in M. Palaniswami, S. Marusic, and Y. W. Law, editors. *2007 3rd International Conference on Intelligent Sensors, Sensor Networks and Information Processing*. <https://doi.org/10.1109/ISSNIP.2007.4496859>
- Camaclang, A. E., M. Maron, T. G. Martin, and H. P. Possingham. 2015. Current practices in the identification of critical habitat for threatened species. *Conservation Biology* 29 (2):482-492. <https://doi.org/10.1111/cobi.12428>
- Chambert, T., D. A. W. Miller, and J. D. Nichols. 2015. Modeling false positive detections in species occurrence data under different study designs. *Ecology* 96(2):332-339. <https://doi.org/10.1890/14-1507.1>
- Clement, M. J. 2016. Designing occupancy studies when false-positive detections occur. *Methods in Ecology and Evolution* 7:1538-1547. <https://doi.org/10.1111/2041-210X.12617>
- Cook, A., and S. Hartley. 2018. Efficient sampling of avian acoustic recordings: intermittent subsamples improve estimates of single species prevalence and total species richness. *Avian Conservation and Ecology* 13(1):21. <https://doi.org/10.5751/ACE-01221-130121>
- Crins, W. J., P. A. Gray, P. W. Uhlig, and M. C. Wester. 2009. *The ecosystems of Ontario, Part I: Ecozones and ecoregions*. Technical Report. Ontario Ministry of Natural Resources, Peterborough, Ontario, Canada.
- Farmer, R. G., M. L. Leonard, and A. G. Horn. 2012. Observer effects and avian-call-count survey quality: rare-species biases and overconfidence. *Auk* 129(1): 76-86. <https://doi.org/10.1525/auk.2012.11129>
- Ferguson, P. F. B., M. J. Conroy, and J. Hepinstall-Cymerman. 2015. Occupancy models for data with false positive and false negative errors and heterogeneity across sites and surveys. *Methods in Ecology and Evolution* 6(12):1395-1406. <https://doi.org/10.1111/2041-210X.12442>
- Fiske, I., and R. Chandler. 2011. Unmarked: an R package for fitting hierarchical models of wildlife occurrence and abundance. *Journal of Statistical Software* 43(10):1-23. <https://doi.org/10.18637/jss.v043.i10>
- Guillera-Arroita, G., J. J. Lahoz-Monfort, A. R. van Rooyen, A. R. Weeks, and R. Tingley. 2017. Dealing with false positive and false negative errors about species occurrence at multiple levels. *Methods in Ecology and Evolution* 8(9):1081-1091. <https://doi.org/10.1111/2041-210X.12743>
- Hines, J. E. 2006. *PRESENCE- Software to estimate patch occupancy and related parameters*. U.S. Geological Survey, Patuxent Wildlife Research Center, Laurel, Maryland, USA.
- Hobson, K. A., R. S. Rempel, H. Greenwood, B. Turnbull, and S. L. VanWilgenburg. 2002. Acoustic surveys of birds using electronic recordings: new potential from an omnidirectional microphone system. *Wildlife Society Bulletin* 30(3):709-720.
- Knight, E. C., and E. M. Bayne. 2018. Classification threshold and training data affect the quality and utility of focal species data processed with automated audio-recognition software. *Bioacoustics*. <https://doi.org/10.1080/09524622.2018.1503971>
- Knight, E. C., K. C. Hannah, G. Foley, C. Scott, R. M. Brigham, and E. Bayne. 2017. Recommendations for acoustic recognizer performance assessment with application of five common automated signal recognition programs. *Avian Conservation and Ecology* 12(2):14. <https://doi.org/10.5751/ACE-01114-120214>
- Loehle, C., and D. J. H. Sleep. 2015. Use and application of range mapping in assessing extinction risk in Canada. *Wildlife Society Bulletin* 39(3):658-663. <https://doi.org/10.1002/wsb.574>
- MacKenzie, D. I., J. D. Nichols, G. B. Lachman, S. Droege, J. A. Royle, and C. A. Langtimm. 2002. Estimating site occupancy rates when detection probabilities are less than one. *Ecology* 83 (8):2248-2255. [https://doi.org/10.1890/0012-9658\(2002\)083\[2248:ESORWD\]2.0.CO;2](https://doi.org/10.1890/0012-9658(2002)083[2248:ESORWD]2.0.CO;2)
- MacKenzie, D. I., J. D. Nichols, J. A. Royle, K. H. Pollock, L. L. Bailey, and J. E. Hines. 2006. *Occupancy estimation and modeling: inferring patterns and dynamics of species occurrence*. Academic Press, Amsterdam, The Netherlands.
- Matsuoka, S. M., L. C. Mahon, C. M. Handel, P. Sóllymos, E. M. Bayne, P. C. Fontaine, and C. J. Ralph. 2014. Reviving common standards in point-count surveys for broad inference across studies. *Condor* 116:599-608. <https://doi.org/10.1650/CONDOR-14-108.1>
- McClintock, B. T., L. L. Bailey, K. H. Pollock and T. R. Simons. 2010a. Unmodeled observation error induces bias when inferring patterns and dynamics of species occurrence via aural detections. *Ecology* 91(8):2446-2454. <https://doi.org/10.1890/09-1287.1>
- McClintock, B. T., L. L. Bailey, K. H. Pollock, and T. R. Simons. 2010b. Experimental investigation of observation error in anuran call surveys. *Journal of Wildlife Management* 74(8):1882-1893. <https://doi.org/10.2193/2009-321>
- McKelvey, K. S., K. B. Aubry and M. K. Schwartz. 2008. Using anecdotal occurrence data for rare or elusive species: the illusion of reality and a call for evidentiary standards. *BioScience* 58 (6):549-555. <https://doi.org/10.1641/B580611>
- Miller, D. A., J. D. Nichols, B. T. McClintock, E. H. C. Grant, L. L. Bailey, and L. A. Weir. 2011. Improving occupancy estimation when two types of observational error occur: non-detection and species misidentification. *Ecology* 92(7):1422-1428. <https://doi.org/10.1890/10-1396.1>
- Miller, D. A. W., J. D. Nichols, J. A. Gude, L. N. Rich, K. M. Podrutzny, J. E. Hines, and M. S. Mitchell. 2013. Determining occurrence dynamics when false positives occur: estimating the range dynamics of wolves from public survey data. *PLoS ONE* 8 (6):e65808. <https://doi.org/10.1371/journal.pone.0065808>

- Miller, D. A. W., L. A. Weir, B. T. McClintock, E. H. C. Grant, L. L. Bailey, and T. R. Simons. 2012. Experimental investigation of false positive errors in auditory species occurrence surveys. *Ecological Applications* 22(5):1665-1674. <https://doi.org/10.1890/11-2129.1>
- Muggeo, V. M. 2003. Estimating regression models with unknown break-points. *Statistics in Medicine* 22(19):3055-3071. <https://doi.org/10.1002/sim.1545>
- Muggeo, V. M. 2008. Segmented: an R package to fit regression models with broken-line relationships. *R News* 8(1):20-25.
- Nichols, J. D. 2019. Confronting uncertainty: contributions of the wildlife profession to the broader scientific community. *Journal of Wildlife Management* 83(3):519-533. <https://doi.org/10.1002/jwmg.21630>
- North American Bird Conservation Initiative. 2000. *North American Bird Conservation Initiative: bird conservation region descriptions*. U.S. Fish and Wildlife Service, Washington, D.C., USA.
- Ranjard, L., and H. A. Ross. 2008. Unsupervised bird song syllable classification using evolving neural networks. *Journal of the Acoustical Society of America* 123(6):4358-4368. <https://doi.org/10.1121/1.2903861>
- Rempel, R. S., J. M. Jackson, and J. N. Robinson. 2014. *Acoustic monitoring and assessment of forest songbirds: sample design, analysis methods, and observation error*. CNFER Technical Report TR-012. Ontario Ministry of Natural Resources, Centre for Northern Forest Ecosystem Research, Thunder Bay, Ontario, Canada.
- Royle, J. A., and W. A. Link. 2006. Generalized site occupancy models allowing for false positive and false negative errors. *Ecology* 87(4):835-841. [https://doi.org/10.1890/0012-9658\(2006\)87\[835:GSOMAF\]2.0.CO;2](https://doi.org/10.1890/0012-9658(2006)87[835:GSOMAF]2.0.CO;2)
- Sewell, D., G. Guillerá-Arroita, R. A. Griffiths, and T. J. Beebe. 2012. When is a species declining? Optimizing survey effort to detect population changes in reptiles. *PLoS ONE* 7(8):e43387. <https://doi.org/10.1371/journal.pone.0043387>
- Shonfield, J., and E. M. Bayne. 2017. Autonomous recording units in avian ecological research: current use and future applications. *Avian Conservation and Ecology* 12(1):14. <https://doi.org/10.5751/ACE-00974-120114>
- Simons, T. R., M. W. Alldredge, K. H. Pollock, and J. M. Wettroth. 2007. Experimental analysis of the auditory detection process on avian point counts. *Auk* 124(3):986-999. [https://doi.org/10.1642/0004-8038\(2007\)124\[986:EAOTAD\]2.0.CO;2](https://doi.org/10.1642/0004-8038(2007)124[986:EAOTAD]2.0.CO;2)
- Simons, T. R., K. H. Pollock, J. M. Wettroth, M. W. Alldredge, K. Pacifici, and J. Brewster. 2009. Sources of measurement error, misclassification error, and bias in auditory avian point count data. Pages 237-254 in D. L. Thomson, E. G. Cooch, and M. J. Conroy, editors. *Modeling demographic processes in marked populations*. Springer Science + Business Media, New York, New York, USA. https://doi.org/10.1007/978-0-387-78151-8_10
- Sólymos, P., S. M. Matsuoka, D. Stralberg, N. K. Barker, and E. M. Bayne. 2018. Phylogeny and species traits predict bird detectability. *Ecography* 41(10):1595-1603. <https://doi.org/10.1111/ecog.03415>
- Specht, H. M., H. T. Reich, F. Iannarilli, M. R. Edwards, S. P. Stapleton, M. D. Weegman, M. K. Johnson, B. J. Yohannes, and T. W. Arnold. 2017. Occupancy surveys with conditional replicates: an alternative sampling design for rare species. *Methods in Ecology and Evolution* 8:1725-1734. <https://doi.org/10.1111/2041-210X.12842>
- Stowell, D., M. D. Wood, H. Pamula, Y. Stylianou, and H. Glotin. 2019. Automatic acoustic detection of birds through deep learning: the first Bird Audio Detection challenge. *Methods in Ecology and Evolution* 10(3):368-380. <https://doi.org/10.1111/2041-210X.13103>
- Taylor, M. F. J., K. F. Suckling, and J. J. Rachlinski. 2005. The effectiveness of the Endangered Species Act: a quantitative analysis. *BioScience* 55(4):360-367. [https://doi.org/10.1641/0006-3568\(2005\)055\[0360:TEOTES\]2.0.CO;2](https://doi.org/10.1641/0006-3568(2005)055[0360:TEOTES]2.0.CO;2)
- Thompson, S. J., C. M. Handel, and L. B. McNew. 2017. Autonomous acoustic recorders reveal complex patterns in avian detection probability. *Journal of Wildlife Management* 81(7):1228-1241. <https://doi.org/10.1002/jwmg.21285>
- Van Wilgenburg, S. L., P. Sólymos, K. J. Kardynal, and M. D. Frey. 2017. Paired sampling standardizes point count data from humans and acoustic recorders. *Avian Conservation and Ecology* 12(1):13. <https://doi.org/10.5751/ACE-00975-120113>
- Venier, L. A., M. J. Mazerolle, A. Rodgers, K. A. McIlwrack, S. Holmes, and D. Thompson. 2017. Comparison of semiautomated bird song recognition with manual detection of recorded bird song samples. *Avian Conservation and Ecology* 12(2):2. <https://doi.org/10.5751/ACE-01029-120202>
- Weninger, F., and B. Schuller. 2011. Audio recognition in the wild: static and dynamic classification on a real-world database of animal vocalizations. Pages 337-340 in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. <https://doi.org/10.1109/ICASSP.2011.5946409>
- Wimmer, J., M. Towsey, B. Planitz, I. Williamson, and P. Roe. 2013. Analysing environmental acoustic data through collaboration and automation. *Future Generation Computer Systems* 29(2):560-568. <https://doi.org/10.1016/j.future.2012.03.004>
- Yip, D., L. Leston, E. M. Bayne, P. Sólymos, and A. Grover. 2017. Experimentally derived detection distances from audio recordings and human observers enable integrated analysis of point count data. *Avian Conservation and Ecology* 12(1):11. <https://doi.org/10.5751/ACE-00997-120111>

Appendix 1. Annotated Excel spreadsheet for generating simulated observation data with detection and misidentification error. This version is for simulating data where 6 repeat observations (recordings) occur at a site.

Please [click here](#) to download file 'appendix1.xlsm'.

Appendix 2. Annotated Excel spreadsheet for generating simulated observation data with detection and misidentification error. This version is for simulating data where 3 repeat observations (recordings) occur at a site.

Please [click here](#) to download file 'appendix2.xlsm'.

APPENDIX 3.

Table A3. Species names and codes, and detection covariates associated with AIC supported models[†].

Common Name	Scientific Name	AOU Code	Detection covariates [†] Study Area 1	Detection covariates Study Area 2	Detection covariates Study Area 3
Alder Flycatcher	<i>Empidonax alnorum</i>	ALFL	null	null	null
American Bittern	<i>Botaurus lentiginosus</i>	AMBI	PHW	-	-
American Redstart	<i>Setophaga ruticilla</i>	AMRE	TSSR_QUALITY	TSSR_QUALITY_DSS_PHW	PHW
American Robin	<i>Turdus migratorius</i>	AMRO	null	TSSR_QUALITY_DSS	PHW
Black-and-white Warbler	<i>Mniotilta varia</i>	BAWW	null	DSS	TSSR_PHW
Bay-breasted Warbler	<i>Setophaga castanea</i>	BBWA	DSS	TSSR	QUALITY*
Black-capped Chickadee	<i>Poecile atricapillus</i>	BCCH	TSSR_DSS_PHW	PHW	null
Blue-headed Vireo	<i>Vireo solitarius</i>	BHVI	TSSR_DSS_PHW	TSSR_QUALITY	DSS
Blackburnian Warbler	<i>Setophaga fusca</i>	BLBW	DSS_PHW	QUALITY_DSS	null
Blue Jay	<i>Cyanocitta cristata</i>	BLJA	TSSR_DSS	TSSR_DSS_PHW	TSSR_PHW
Boreal Chickadee	<i>Poecile hudsonicus</i>	BOCH	DSS_PHW	-	-
Brown Creeper	<i>Certhia americana</i>	BRCR	null	null	TSSR_QUALITY
Black-throated Blue Warbler	<i>Setophaga caerulescens</i>	BTBW	QUALITY_DSS_PHW	-	TSSR_PHW
Black-throated Green Warbler	<i>Setophaga virens</i>	BTNW	TSSR_QUALITY	-	null
Canada Warbler	<i>Cardellina canadensis</i>	CAWA	null	QUALITY_PHW	DSS
Cedar Waxwing	<i>Bombycilla cedrorum</i>	CEDW	TSSR	TSSR_QUALITY	TSSR
Chipping Sparrow	<i>Spizella passerina</i>	CHSP	TSSR_DSS_PHW	PHW	TSSR_PHW
Cape May Warbler	<i>Setophaga tigrina</i>	CMWA	-	TSSR_QUALITY	-
Common Nighthawk	<i>Chordeiles minor</i>	CONI	TSSR_DSS_PHW	TSSR_DSS_PHW	TSSR_QUALITY_DSS_PHW*
Connecticut Warbler	<i>Oporornis agilis</i>	CONW	-	PHW	-
Common Yellowthroat	<i>Geothlypis trichas</i>	COYE	QUALITY_DSS_PHW	PHW	PHW
Chestnut-sided Warbler	<i>Setophaga pensylvanica</i>	CSWA	TSSR	null	TSSR_DSS_PHW
Dark-eyed Junco	<i>Junco hyemalis</i>	DEJU	TSSR_PHW	null	TSSR_QUALITY_DSS
Eastern Wood-Pewee	<i>Contopus virens</i>	EAWP	-	-	null
Eastern Whip-poor-will	<i>Caprimulgus vociferus</i>	EWPW	-	-	QUALITY_PHW*
Golden-crowned Kinglet	<i>Regulus satrapa</i>	GCKI	TSSR_QUALITY	TSSR_QUALITY_PHW	PHW

Gray Jay	<i>Perisoreus canadensis</i>	GRAJ	DSS	TSSR_QUALITY	TSSR_PHW
Hairy Woodpecker	<i>Picoides villosus</i>	HAWO	TSSR_DSS	PHW	QUALITY_PHW
Hermit Thrush	<i>Catharus guttatus</i>	HETH	TSSR_QUALITY_DSS_PHW	TSSR_QUALITY_DSS	TSSR_PHW
Least Flycatcher	<i>Empidonax minimus</i>	LEFL	TSSR_QUALITY_PHW	DSS	PHW
Lincoln's Sparrow	<i>Melospiza lincolnii</i>	LISP	null	QUALITY_DSS_PHW	PHW
Magnolia Warbler	<i>Setophaga magnolia</i>	MAWA	QUALITY_DSS_PHW	null	QUALITY_PHW
Mourning Warbler	<i>Geothlypis philadelphia</i>	MOWA	QUALITY_PHW	TSSR_PHW	QUALITY_DSS_P HW
Nashville Warbler	<i>Oreothlypis ruficapilla</i>	NAWA	TSSR_PHW	TSSR_DSS	TSSR_QUALITY_ PHW
Northern Flicker	<i>Colaptes auratus</i>	NOFL	TSSR_QUALITY_PHW	TSSR	TSSR_QUALITY
Northern Parula	<i>Setophaga americana</i>	NOPA	TSSR_DSS_PHW	QUALITY_PHW	PHW
Northern Waterthrush	<i>Parkesia noveboracensis</i>	NOWA	TSSR_QUALITY	QUALITY_PHW	null
Olive-sided Flycatcher	<i>Contopus cooperi</i>	OSFL	TSSR_PHW	null	null
Ovenbird	<i>Seiurus aurocapilla</i>	OVEN	TSSR_DSS_PHW	PHW	TSSR_DSS_PHW
Palm Warbler	<i>Setophaga palmarum</i>	PAWA	TSSR_QUALITY_DSS	TSSR_PHW	TSSR_PHW
Philadelphia Vireo	<i>Vireo philadelphicus</i>	PHVI	PHW	-	QUALITY_PHW
Pine Siskin	<i>Spinus pinus</i>	PISI	-	TSSR_QUALITY_DSS	QUALITY*
Pileated Woodpecker	<i>Dryocopus pileatus</i>	PIWO	TSSR	TSSR	QUALITY_DSS
Purple Finch	<i>Carpodacus purpureus</i>	PUFI	-	-	-
Rose-breasted Grosbeak	<i>Pheucticus ludovicianus</i>	RBGR	-	TSSR_QUALITY	null
Red-breasted Nuthatch	<i>Sitta canadensis</i>	RBNU	TSSR_QUALITY	TSSR_PHW	TSSR_DSS
Ruby-crowned Kinglet	<i>Regulus calendula</i>	RCKI	TSSR_PHW	TSSR_DSS_PHW	null
Red-eyed Vireo	<i>Vireo olivaceus</i>	REVI	TSSR_QUALITY_DSS_PHW	TSSR_PHW	PHW
Ruffed Grouse	<i>Bonasa umbellus</i>	RUGR	DSS_PHW	TSSR_QUALITY	TSSR_QUALITY_ PHW*
Red-winged Blackbird	<i>Agelaius phoeniceus</i>	RWBL	TSSR	TSSR	TSSR_PHW
Song Sparrow	<i>Melospiza melodia</i>	SOSP	-	TSSR_DSS_PHW	QUALITY
Swamp Sparrow	<i>Melospiza georgiana</i>	SWSP	TSSR	PHW	PHW
Swainson's Thrush	<i>Catharus ustulatus</i>	SWTH	TSSR_QUALITY_PHW	TSSR	TSSR_QUALITY_ DSS_PHW
Tennessee Warbler	<i>Oreothlypis peregrina</i>	TEWA	PHW	DSS	-
Veery	<i>Catharus fuscescens</i>	VEER	null	TSSR_QUALITY_DSS	TSSR_QUALITY_ DSS
Wilson's Warbler	<i>Cardellina pusilla</i>	WIWA	DSS_PHW	QUALITY_PHW	PHW
Winter Wren	<i>Troglodytes hiemalis</i>	WIWR	DSS	DSS	null
White-throated Sparrow	<i>Zonotrichia albicollis</i>	WTSP	TSSR_QUALITY_PHW	TSSR	TSSR_PHW

Yellow-bellied Flycatcher	<i>Empidonax flaviventris</i>	YBFL	TSSR_DSS_PHW	TSSR_QUALITY_PHW	TSSR_DSS
Yellow-bellied Sapsucker	<i>Sphyrapicus varius</i>	YBSA	TSSR_QUALITY_DSS_PHW	TSSR_PHW	null
Yellow-rumped Warbler	<i>Setophaga coronata</i>	YRWA	TSSR_QUALITY_DSS	null	null
Yellow Warbler	<i>Setophaga petechia</i>	YWAR	DSS_PHW	TSSR	TSSR_QUALITY_PHW

†Models were selected with lowest AIC values, and in some cases the model with no detection covariates was most strongly supported. In a few cases number of observations was too low to model.

‡Detection covariates (dc): TSSR - time since sunrise/sunset; DSS - days since spring – May15th; RQ - recording quality; PHW - percent hardwood.

APPENDIX 4. Simulation comparison of known (true) occupancy model parameters with parameters estimated from the multiple detection state (MDS) model.

Figure A4.1. Comparison of parameter values for the known simulation versus estimated values, where true occupancy rate (OR) was 0.2 - 0.8, and false positive (FP) rate from about 0.001 to 0.07. Horizontal lines are mean values, and boxes represent the interquartile range; true OR is based on the known simulation data, ψ is the occupancy estimate from the MDS model, d is detection rate ($p11$), FP is false positive rate ($p10$) and b is probability a detection was a confirmed detection. The ^ indicates a model estimate.

Figure A4.2. Comparison of parameter values for the known simulation versus estimated values, where true occupancy rate (OR) varied from 0.2 - 0.8, and detection rate (d) varied from about 0.01 to 0.2. Horizontal lines are mean values, and boxes represent the interquartile range; true OR is based on the known simulation data, ψ is the occupancy estimate from the MDS model, d is detection rate ($p11$), FP is false positive rate ($p10$) and b is probability a detection was a confirmed detection. The $\hat{}$ indicates a model estimate.