

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>  
Copyright © 2017 by the author(s). Published here under license by the Resilience Alliance.

*Methodology*, part of a Special Feature on Advancing bird population monitoring with acoustic recording technologies

## Paired sampling standardizes point count data from humans and acoustic recorders

Steven L. Van Wilgenburg<sup>1</sup>, Péter Sólymos<sup>2</sup>, Kevin J. Kardynal<sup>3</sup> and Matthew D. Frey<sup>1</sup>

<sup>1</sup>Environment and Climate Change Canada, Canadian Wildlife Service, <sup>2</sup>Department of Biological Sciences, University of Alberta, Edmonton, Alberta, Canada, <sup>3</sup>Science & Technology Branch, Environment and Climate Change Canada

**ABSTRACT.** Acoustic recordings are increasingly used to quantify occupancy and abundance in avian monitoring and research. The recent development of relatively inexpensive programmable autonomous recording units (ARUs) has further increased the utility of acoustic recording technologies. Despite their potential advantages, persistent questions remain as to how comparable data are between ARUs and traditional (human observer) point counts. We suggest that differences in counts obtained from ARUs versus human observers primarily stem from differences in the effective detection radius of humans ( $EDR_H$ ) versus ARUs ( $EDR_A$ ). We describe how paired sampling can be used in conjunction with generalized linear (GLM) or generalized linear mixed models (GLMM) to estimate correction factors ( $\delta$ ) to remove biases between ARUs and traditional point counts. Furthermore, if human observers conduct distance estimation, we show that density estimates can be derived from single ARUs by estimating  $EDR_A$  as a function of  $EDR_H$  and  $\delta$ , thus providing alternatives to more complicated and expensive approaches. We demonstrate our approach using data from 363 point count stations in 105 unique boreal study sites at which field staff conducted point count surveys that were simultaneously recorded by an ARU and later transcribed in the lab. Finally, we used repeated random subsampling of the data to split the data into model creation (70%) and validation (30%) subsets to iteratively estimate  $\delta$  and validate density estimates from ARUs against densities calculated from human observers at the same independent validation locations. We modeled density of 35 species of boreal forest birds and show that incorporating  $\delta$  in statistical offsets successfully removes systematic biases in estimated avian counts and/or density between human and ARU derived surveys. Our method is therefore easily implemented and will facilitate the integration of ARU and human observer point count data, facilitating expanded monitoring efforts and meta-analyses with historic point count data.

## Standardisation de données de points d'écoute issus d'échantillons appariés d'observateurs et d'enregistrements acoustiques

**RÉSUMÉ.** Les enregistrements acoustiques sont de plus en plus utilisés pour quantifier l'occurrence et l'abondance dans le cadre de suivis et de recherches aviaires. Le développement récent d'unités d'enregistrement autonomes (ARU, pour *autonomous recording units*), programmables et plutôt bon marché a contribué à hausser l'utilité des technologies d'enregistrement acoustique. En dépit de leurs avantages possibles, des questions demeurent quant à la comparabilité des données issues de points d'écoute traditionnels (observateur humain) et d'ARU. Nous supposons que les différences obtenues entre les dénombrements issus d'ARU et ceux issus d'observateurs proviennent de différences dans le rayon de détection effectif (EDR, pour *effective detection radius*) des humains ( $EDRH$ ) comparativement à celui des ARU ( $EDRA$ ). Nous décrivons de quelle façon un échantillonnage apparié peut être utilisé de pair avec des modèles linéaires généralisés ou linéaires généralisés à effets mixtes pour estimer les facteurs de correction ( $\delta$ ) permettant de supprimer les biais entre les points d'écoute traditionnels et ceux provenant d'ARU. De plus, si les observateurs doivent estimer les distances, nous montrons comment des estimations de densité peuvent être dérivées à partir de simples ARU, en estimant  $EDRA$  en fonction de  $EDRH$  et de  $\delta$ , fournissant ainsi une alternative aux approches plus complexes et coûteuses. Nous démontrons notre approche à partir de données provenant de 360 stations de point d'écoute localisées dans 105 sites d'étude boréaux, où du personnel de terrain a réalisé des dénombrements par point d'écoute qui ont été enregistrés simultanément par ARU et ensuite transcrits en laboratoire. Enfin, nous avons procédé par sous-échantillonnage aléatoire répété des données pour les séparer en sous-ensembles destinés à la création de modèles (70 %) ou à la validation (30 %) afin d'estimer  $\delta$  de façon itérative et de valider les estimations de densité obtenues au moyen d'ARU comparativement à celles calculées à partir des données récoltées par des observateurs aux mêmes sites indépendants de validation. Nous avons modélisé la densité de 35 espèces d'oiseaux de forêt boréale et montrons que l'incorporation de  $\delta$  dans les compensations statistiques enlève correctement les biais systématiques des estimations de dénombrements aviaires et/ou de densités trouvés entre les relevés faits par des observateurs et ceux obtenus au moyen d'ARU. Notre méthode est simple à utiliser, facilite l'intégration de données de points d'écoute issus d'observateurs et d'ARU, et permettra des efforts de suivi plus grands et des méta-analyses avec des données de point d'écoute historiques.

**Key Words:** *ARU; autonomous recording unit; avian monitoring; bioacoustics; detection probability; detection radius; point count; survey bias*

## INTRODUCTION

Long-term monitoring programs such as the North American Breeding Bird Survey (hereafter BBS; Sauer et al. 2014) and the Christmas Bird Count (Link et al. 2006) provide crucial data that help to guide species status assessments and conservation efforts (Downes et al. 2016). Despite the success of those monitoring programs, many species are insufficiently monitored because of large geographic gaps in avian monitoring programs (Sauer et al. 2003, Francis et al. 2009, Machtans et al. 2014), or the species life history traits, e.g., crepuscular/nocturnal behavior, make them hard to monitor using traditional census methods (Goyette et al. 2011, Zwart et al. 2014). Many of the gaps in monitoring efforts exist in regions, e.g., boreal forest, or habitats where poor access and logistical constraints, high cost, and a lack of skilled observers have hindered monitoring efforts (Sauer et al. 2003, Francis et al. 2009, Machtans et al. 2014). Thus, it would be beneficial to augment monitoring efforts with alternative methods for poorly sampled species, habitats, and regions to better guide species assessments, conservation prioritization, and test hypotheses for causes of population change.

One potential solution to monitor bird populations in regions and habitats in which it is difficult to send skilled observers is to augment human observers with stereo recordings (Hobson et al. 2002, Francis et al. 2009, Klingbeil and Willig 2015). This approach is a plausible solution because most avian monitoring and research programs collect count data using point count surveys that primarily rely upon detection of acoustic cues (Hobson et al. 2002, Blumstein et al. 2011, Matsuoka et al. 2014). Indeed, several comparisons of data from stereo recordings against human observers in the field suggest that recordings provide relatively similar estimates of species abundance and community composition (Hobson et al. 2002, Blumstein et al. 2011, Venier et al. 2012, Klingbeil and Willig 2015).

Although many comparisons of acoustic recordings with point counts conducted by human observers suggest the data are generally comparable, subtle differences nonetheless do exist (Hutto and Stutzman 2009, Venier et al. 2012). For example, Hutto and Stutzman (2009) found that significantly more species were detected using point counts than acoustic recordings and speculated that this may in part be due to differences in the radius over which species were audible between methods. In addition to potential differences between a given recording system and human observers, previous research has also shown variable species detection between different recording systems (Rempel et al. 2013). Thus, broader incorporation of ARU-based data into monitoring programs may require correcting for differential detectability between ARUs and human observers in the field (Sidie-Slettedahl et al. 2015). We describe a sampling design and analysis framework to relate counts from ARUs to traditional point count data and derive estimated bird densities for both by correcting data for biases in species availability and perceptibility following earlier developments by Bart and Earnst (2002) and Sólymos et al. (2013). Development of this method will allow for efficient and cost-effective augmentation of acoustic monitoring programs with ARU technology for species and regions with poor survey coverage.

We develop a framework to use simultaneously conducted human point count and ARU surveys to estimate statistical offsets to adjust for systematic differences between counts conducted by humans versus ARUs (following Sólymos et al. 2013). We use field data to test whether our approach removes bias in ARU-based counts relative to densities estimated from field observers conducting point counts using both distance estimation and time removal sampling. Our proposed approach to correcting counts between ARUs and humans would be most easily applied if the statistical offsets do not vary with other factors affecting detection, and we therefore tested whether the offsets differed by habitat or with environmental noise. Based on the literature and field experience, we hypothesized that ratio of counts from ARUs relative to counts from human observers conducted at the same time and location would be  $< 1$  on the assumption that the detection radius of the ARU would be less than that of a human observer. We also hypothesized that the same ratio would be smaller when recordings were made in deciduous vs. other forest types and/or windy conditions because microphones tend to amplify leaf rustle (B. Turnbull, *personal communication*).

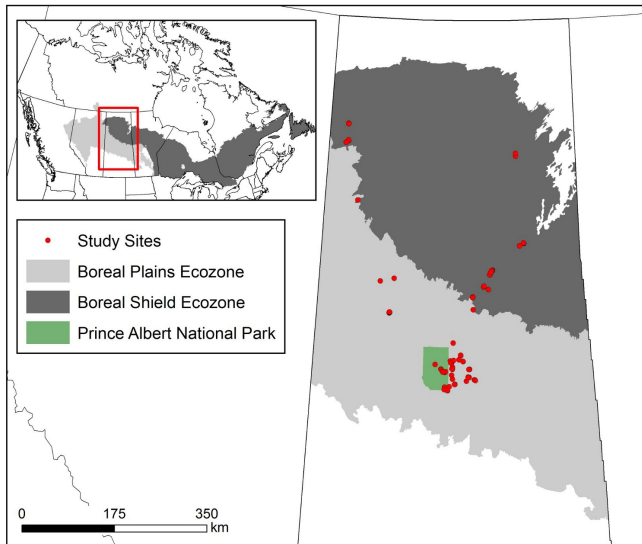
## METHODS

### Avian sampling

We conducted our study in the boreal forest of Saskatchewan, Canada. Study sites were located in Bird Conservation Region 6 (Boreal Plains Ecozone) and Bird Conservation Region 8 (Boreal Shield Ecozone) between 53° 34'N, 103° 43' W and 58° 08'N, 109° 28' W (Fig. 1) in the summers of 2014 and 2015. Surveys were conducted at 363 unique point count stations distributed among 105 study sites, each of which constituted unique forest stands. We sampled between 1 and 12 point count locations per study site (median = 3). Effort was approximately equally distributed between years ( $n = 205$  in 2014 and  $n = 200$  in 2015), with 42 point count stations in 9 separate study sites surveyed in both years. Based on the 250 m resolution Land Cover Map of Canada 2005 (LCC05; Latifovic et al. 2008), the most frequently sampled land cover classes were open coniferous (29.2%), closed mature mixed (16.3%), and open mature deciduous (11.6%) forest, with the remainder of the point count samples distributed among 13 other land cover classes (Table 1).

All surveys were conducted by one of five observers between 15 minutes prior to sunrise and 4.5 hours after sunrise between 1–29 June. Upon arriving at the point count station, observers attached a Song Meter SM2+ ARU with a pair of SMX-II microphones (Wildlife Acoustics Inc. ©, Maynard, MA) to the nearest tree at approximately head height and began a manual recording. ARUs were programmed to record in stereo in wav file format, using a sampling rate of 44,100 samples per second and using factory default acoustic gain settings for the microphone preamplifier. The observer stood approximately 3 m from the ARU to avoid introducing extraneous noise into the recordings. Observers then announced when they began and ended a simultaneous 10-minute point count to ensure data from the subsequent transcription of ARU recordings were collected over the identical time frame to point counts conducted by the human observers.

**Fig. 1.** Distribution of 105 study sites (forest stands) surveyed using paired human observer and autonomous recording unit point counts ( $n = 363$ ) in the Boreal Plains ( $n = 66$  stands) and Boreal Shield ( $n = 39$  stands) ecozones.



**Table 1.** Distribution of point count samples amongst land cover classes derived from the 250 m resolution Land Cover Map of Canada 2005 (LCC05; Latifovic et al. 2008).

Land cover class	Number of Stations	Percent
Open coniferous	106	29.2
Closed Mature Mixed	59	16.3
Open Mature Deciduous	42	11.6
Open Young Mixed	28	7.7
Sparse Coniferous	22	6.1
Poorly Drained	20	5.5
Closed Coniferous	19	5.2
Closed Young Mixed	16	4.4
Open Young Deciduous	15	4.1
Sparse Coniferous Shield	12	3.3
Closed Deciduous Mixed	9	2.5
Burns	4	1.1
Open Herb/Grass	4	1.1
Closed Deciduous	3	0.8
Open Northern	3	0.8
Open Mixed	1	0.3
<b>Total</b>	<b>363</b>	<b>100.0</b>

Our point count protocol followed those recommended in Matsuoka et al. (2014). In brief, field observers placed observed or acoustically detected individuals into one of three distance bins (0–50, 50–100, > 100 m) while conducting point counts. Observers were trained in distance estimation prior to field work in addition to opportunistically ground-truthing distance estimates using GPS units to estimate distances between point count centroids and birds heard or observed while walking between point count

locations. To account for differences in availability, observers additionally coded observations to the time interval (0–3, 3–5, and 5–10 minutes) of initial detection; thus treating any subsequent detections of the same individual as though it were “removed” from the population (Farnsworth et al. 2002).

To avoid introducing additional observer biases, ARU recordings were transcribed by the same observer that conducted the field count. During transcriptions, each acoustically identified individual was coded into one of 10 subset 1-minute long time periods (0–1, 1–2, through 9–10 minutes, respectively) to facilitate estimation of availability using a count-removal approach. ARU based intervals were subsequently collapsed to the 0–3, 3–5, and 5–10 minute intervals to match the human observer-based design. Transcribers were not privy to field notes during transcription and transcription was conducted after field season. Unlike counts conducted under field conditions, transcribers were allowed to pause and/or rewind the recording, e.g., to confirm identification, as is frequently done in data transcription. Finally, transcribers categorized environmental noise recorded from the ARU data for each point count using a five point scale (1 = none, 2 = light, 3 = moderate, 4 = heavy, and 5 = excessive).

## Framework

Sólymos et al. (2013) previously demonstrated that point count data can be adjusted for differences in field methodologies if the data include information on the time (Farnsworth et al. 2002, Sólymos et al. 2013) and distance (Matsuoka et al. 2012, Sólymos et al. 2013) intervals in which the individuals were first heard. These extra data allow the application of removal (Farnsworth et al. 2002, Sólymos et al. 2013) and distance modeling (Buckland et al. 2001, Matsuoka et al. 2012) to estimate components of detection probability. Specifically, removal or time-of-detection methods allow the estimation of the probability that an individual bird present at the time of survey gave a visual or auditory cue and was therefore available for detection, i.e. availability ( $p$ ), while distance sampling allows estimation of the probability that the available birds were detected (perceptibility [ $q$ ]) given that they were available (Alldredge et al. 2007a, Nichols et al. 2009). The two components of the observation process can be estimated independent of each other using conditional maximum likelihood estimation (see Appendix in Sólymos et al. 2013). Sólymos et al. (2013) established that incorporating the components of detection probability as statistical offsets in generalized linear (GLM) or generalized linear mixed effects (GLMM) models effectively adjusts count data for differences in point count methodology. The offset based method of Sólymos et al. (2013) forms the basis of our approach to placing ARUs and human observers on a similar footing, but assumes we can approximate both components of detection for both humans and ARUs. Obtaining  $p$  for both ARUs and humans is simply a matter of removal sampling (Farnsworth et al. 2002, Sólymos et al. 2013) or employing time-of-detection methods (Alldredge et al. 2007a, b). The use of Global Positioning System (GPS) synchronized ARU arrays allows distance to sound source to be directly estimated via differences in timing of sound arrival to linked ARUs (Dawson and Efford 2009, Mennill et al. 2012). The use of synchronized ARU arrays can provide accurate and precise estimates of density (Dawson and Efford 2009, Mennill et al. 2012), but is expensive owing to the need for many ARUs spaced

over small distances, e.g., ~30m (Mennill et al. 2012). To reduce costs, it would therefore be advantageous to devise methods of estimating distance-related detection error for single ARUs sampled with an unknown effective detection radius (hereafter EDR). Thus, we require a method to indirectly estimate detectability for single ARUs to apply the methods of Sólymos et al. (2013).

For a count conducted by the human observer ( $H$ ), the expected value of a count for a single species from a point count survey observer can be expressed as:

$$E[Y_H] = Np_{(t_j)}q_{(r_k)} = DA_Hp_H \quad (1)$$

where  $Y_H$  is the count,  $N$  is the species' abundance,  $D$  is the point level density (per unit area),  $A_H$  is the area sampled,  $p_{(j)}$  is the probability of an individual singing (and being detected) at least once during the cumulative duration of the count ( $j$ ) given that it is present to be detected ( $j=1, \dots, J$ ; the number of time intervals), and  $q_{(r_k)}$  is the probability that an individual bird within point count radius ( $r_k$ ) is detected given that it provides a cue, e.g., song, ( $k=1, \dots, K$ ; the number of distance intervals). Although the area sampled ( $A_H$ ) is typically unknown, it can be estimated via distance sampling, for example using binomial or multinomial distance estimators to estimate the effective detection radius (EDR, denoted here as  $\tau$ ) assuming perfect detectability ( $q = 1$ ) within this effective distance:

$$\hat{A}_H = \pi \hat{\tau}_H^2 \quad (2)$$

The simplest approach to determine the relationship between counts from human observers and those from ARUs is to conduct paired sampling (or "double sampling" sensu Bart and Earnst 2002). If we simultaneously use an ARU ( $A$ ) to record the same acoustic environment in which a human observer ( $H$ ) is conducting a point count, the population density to which both are exposed is identical by design; i.e.,  $D = D_H = D_A$ . As a result, if all else is equal then differences in the observed counts from ARUs and human observers should be primarily due to differences in the area sampled by each method. We note, however, that minor differences in estimated abundances could also be due to differences in the probability of detecting cues from individuals birds ( $p_H$  versus  $p_A$ ) related to differences in how detections are made in the field versus in laboratory, e.g., possibility of double checking recordings or lack of external distractions in the laboratory. This assumption ( $p_H = p_A$ ) can be explicitly tested by estimating  $p_H$  and  $p_A$  from the data by recording time intervals in which individuals were first detected and using removal models (Farnsworth et al. 2002, Sólymos et al. 2013) or using time-of-detection methods (Allredge et al. 2007a,b). For the sake of simplicity, we start by assuming that  $p_H$  and  $p_A$  are equal. If we divide the expected values of the counts, we can observe the expected relationship between the areas sampled by humans versus ARUs:

$$\frac{E[Y_A]}{E[Y_H]} = \frac{D\hat{A}_A p}{D\hat{A}_H p} = \frac{\pi \hat{\tau}_A^2}{\pi \hat{\tau}_H^2} = \frac{\hat{\tau}_A^2}{\hat{\tau}_H^2} \quad (3)$$

So, if we let  $\delta$  be such that:

$$\delta = \frac{\hat{\tau}_A}{\hat{\tau}_H} \quad (4)$$

then

$$\hat{\tau}_A = \delta \hat{\tau}_H \quad (5)$$

and

$$\hat{\tau}_A^2 = \delta^2 \hat{\tau}_H^2 \quad (6)$$

Therefore, Equation 3 could be written as:

$$\frac{E[Y_A]}{E[Y_H]} = \frac{\delta^2 \hat{\tau}_H^2}{\hat{\tau}_H^2} = \delta^2 \quad (7)$$

As a result, the ratio of mean counts derived from the ARU to mean counts by the human observer provides an estimate of a squared scaling constant ( $\delta^2$ ) that mathematically relates  $\tau_H$  to the unknown EDR of an ARU ( $\tau_A$ ).

Counts are often modeled in log-linear Poisson general linear (GLM) or generalized linear mixed (GLMM) models. If we estimate  $\tau_H$  and  $p_H$  using distance and removal sampling, respectively, following Sólymos et al. (2013) we can calculate a correction factor ( $C$ ) and the mean for a count made at point count location  $i$  ( $i=1, \dots, n$ ; number of locations) by the human observer that can be expressed as:

$$\lambda_{i,H} = D_i \hat{\tau}_H^2 \pi \hat{p} = D_i \hat{C}_{H,i} \quad (8)$$

Poisson or negative binomial GLM or GLMMs can be fit in this fashion combining human observer and ARU based counts using an indicator function ( $I_A$ ) taking 0 value for human observers and 1 for ARU based counts:

$$\log(\lambda_i) = \log(D_i) + \log(C_i) + I_A \log(\delta^2) \quad (9)$$

Log density is estimated as a linear combination of predictor variables and corresponding coefficients.

## Data analysis

Prior to analysis, we removed species, e.g., gulls, ducks, that are poorly monitored using point count methods because they are frequently detected as flyovers and thus violate the closure assumption. We then began by estimating EDR<sub>H</sub> based on our model calibration data. We limited analyses to species with at least 15 detections, and fit half-normal binomial distance models to estimate EDRs (Matsuoka et al. 2012). We then fit count removal models to both the human observer and ARU data using a model in which we included survey type as a factor to test for a difference in species availability. We considered  $p$  unequal if the 95% confidence interval (hereafter 95% CI) for ARU survey parameter estimate did not overlap zero. Distance and removal models were fit using the "detect" package based on conditional maximum likelihood estimating procedure (Sólymos et al. 2016).

Although  $\delta^2$  can be approximated based on the square root of the ratio between arithmetic mean ARU and human observer counts (see above), we are interested in deriving maximum likelihood estimates and associated confidence intervals of  $\delta$  that account for sampling design. These can be derived from coefficients ( $\delta^2 = \exp[\beta]$ ) from Poisson or negative binomial regression, which can be interpreted as the ratio of the count between levels of a treatment. We used Poisson GLMMs to estimate  $\delta$  by including

a fixed effect factor for survey type (ARU vs. human as the reference category), and included random intercepts for station and visit to account for paired observations between human observers and ARUs. Following Sólymos et al. (2013), we used our human observer data to derive species specific estimates of  $\log(\text{EDR}_H^2 \cdot \pi \cdot p)$  and included these as statistical offsets in our GLMMs. We estimated  $\delta$  for each species in which the comparison of availability between ARU and humans (above) showed  $p_H$  is approximately equal to  $p_A$  as per the assumptions of our approach.

We validated the predictive performance of our models and examined bias in density estimates (relative to those estimated by the human observers in the field) by using repeated random subsampling of the data. In each repeated subsample, data were partitioned by randomly selecting 70% of the study sites ( $n = 74$ ) for developing GLMMs from which we estimated  $\delta$ , and 30% of the study sites ( $n = 31$ ) were withheld as independent validation samples. We repeated this random sampling 50 times. In each repeated sample, we estimated  $\delta$  using the aforementioned GLMM structure and calculated the 95% CI across the 50 replicated analyses. We also calculated empirical estimates of  $\delta$  by dividing the mean ARU count by the mean human observer count in the withheld validation data for each of the 50 repeated subsampling events. We then assessed whether the 95% CIs for the GLMM-based estimates of  $\delta$  overlapped with 95% CIs from the empirical estimates of  $\delta$  and examined the (Pearson's) correlation between both estimates of  $\delta$ . In addition, we also examined whether the inclusion of  $\delta$  in statistical offsets reduced bias in predicted densities from ARU surveys within each random subsampling. We began by estimating density for human observations by fitting a GLMM to the subset of each validation subsample in which we included a random intercept for study site and a statistical offset, i.e.,  $\log(\text{EDR}_H^2 \cdot \pi \cdot p)$  to generate mean study site level density estimates. We then fit two competing models to the ARU data from the same sites with the same random effects structure as used for the human observer data, but fit one (our "null" model) in which we included the statistical offset used for the human observer data, and a competing model in which we used the  $\delta$  estimate from the model calibration data within the same iteration to estimate the offset as  $\log([\delta \cdot \text{EDR}_H^2 \cdot \pi \cdot p])$ . Based on these models, we calculated bias as the difference between the mean density predicted from the models fit to the ARU data minus the predicted mean density estimated from the human observer data. We were interested in testing whether  $\delta$  values estimated by different approaches were statistically different. Because  $\delta$  represents a relative difference between two methods and we used the same  $\text{EDR}_H$  estimates, incorporating the uncertainty around  $\text{EDR}_H$  would not have changed our results. It should be noted however that propagating the error through modeling (e.g. as described in Sólymos et al. 2013) might be required when estimating bird densities.

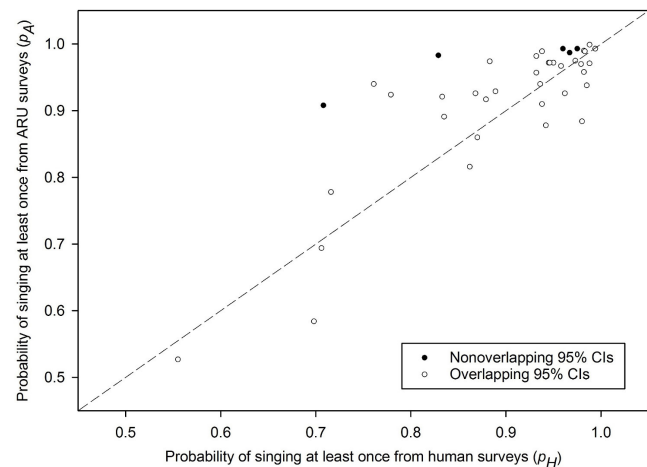
Finally, we used our full data set to assess whether  $\delta$  varied between habitat types (deciduous/mixed wood habitat types versus all other categories) and environmental noise conditions. We constructed six a priori GLMM models that all included random intercepts for station and visit as per above and offsets based on those calculated from human observer data. We included a null (intercept only) model, a model with fixed effects for survey type (two-level factor: ARU vs. human), habitat type (two-level

factor: deciduous/mixed vs. other) and models with both survey type plus habitat type or survey type plus environmental noise as main effects. Although noise was an ordinal variable, previous analyses suggest a reasonably linear response of counts to our noise variable and thus we treated noise as a linear covariate. Finally, in addition to the main effects models, we included two models that incorporated interactions between survey type and habitat type versus one with survey type and noise interaction. We did not consider models including all three main effects and interactions. We selected among competing models based on Akaike's information criterion (AIC, Burnham and Anderson 2002), and we only considered models with a  $\Delta\text{AIC}$  of  $< 2$  as potentially competitive. If habitat or environmental noise differentially impacted the detection radius of an ARU relative to a human observer, models including the interaction terms should receive the greatest support.

## RESULTS

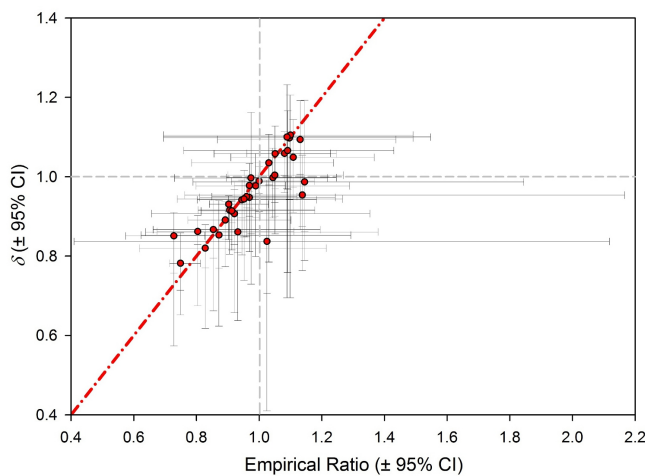
Forty-one species met our minimum sample size criteria (Table A1.1). Across species, effective detection radii ( $\text{EDR}_H$ ) ranged from ~34 to 167 m (median = 68 m; Table A1.1). Of the 41 species for which we estimated availability based on count-removal models, models did not successfully converge for one species (Northern Flicker, *Colaptes auratus*), and five species did not meet the assumption of equal availability based on parameter estimates for the ARU factor in removal models. Estimates of availability ( $p$ ) were strongly correlated (Pearson's  $r = 0.81$ ,  $p < 0.001$ ) between count removal models fit to human observer versus ARU count data (Fig. 2; Table A1.1). Species not meeting the assumption of equal availability include Ruffed Grouse, *Bonasa umbellus* (survey effect for ARU;  $\beta = 0.84$ ,  $\text{SE} = 0.37$ ), Red-eyed Vireo, *Vireo olivaceus* ( $\beta = 0.30$ ,  $\text{SE} = 0.09$ ), Tennessee Warbler, *Oreothlypis peregrina* ( $\beta = 0.25$ ,  $\text{SE} = 0.10$ ), Chestnut-sided Warbler, *Setophaga pensylvanica* ( $\beta = 0.43$ ,  $\text{SE} = 0.20$ ), and Chipping Sparrow, *Spizella passerina* ( $\beta = 0.66$ ,  $\text{SE} = 0.24$ ).

**Fig. 2.** Relationship between probability of a bird singing at least once (and being detected) during a 10-minute point count survey estimated from count-removal models fit to data from 41 species from simultaneous point counts conducted by human observers in the field (x-axis) versus from autonomous recording units (ARUs, y-axis). Dashed diagonal line indicates 1:1 correspondence.



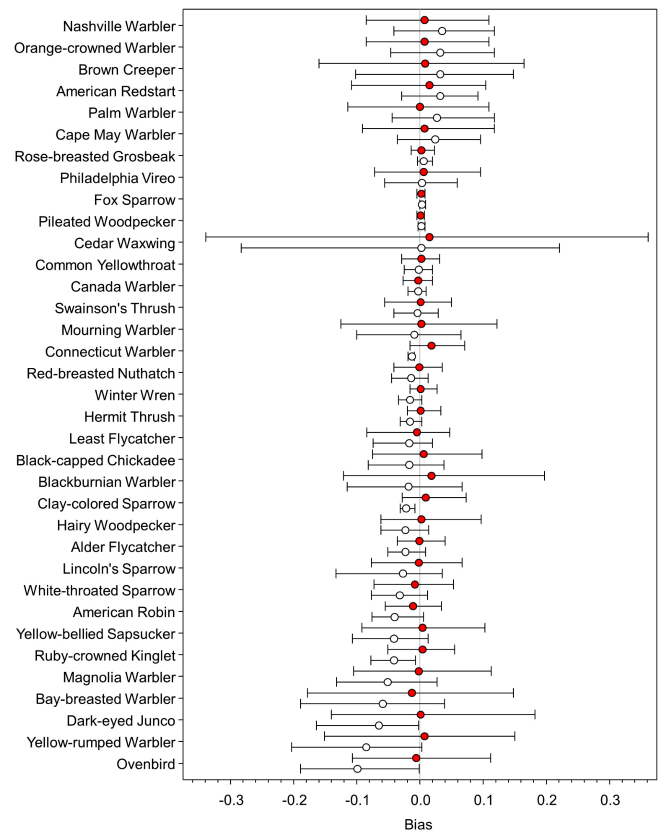
Models examining variation in paired counts between human observers and ARUs suggested that the majority of species were slightly less detectable on ARU recordings than in the field (Fig. 3; Table A1.2). Across species, the median estimate of  $\delta$  was 0.95 (minimum = 0.78, maximum = 1.11); however, 95% CIs overlapped one for 18 out of 35 species (Fig. 3; Table A1.2). Comparison of 95% CIs around estimates of  $\delta$  against those for empirical ratios of ARU to human observer counts showed overlap for all 35 species (Table A1.2). Across species, estimates of  $\delta$  from our calibration models were positively correlated with empirical ratios derived from the withheld validation samples (Fig. 3; Pearson's  $r = 0.84, p < 0.001$ ).

**Fig. 3.** Comparison of Poisson generalized linear mixed model based (GLMM) estimates of  $\delta$  ( $\pm$  95% confidence intervals) against empirical estimates of  $\delta$  ( $\pm$  95% confidence intervals) calculated by dividing the mean ARU count by the mean human observer count in the withheld validation data. GLMM estimates were derived from iteratively (50 repeated random subsamples) fitting GLMMs to 70% of the study sites and compared against the empirical estimates of  $\delta$  which were estimated from the 30% of withheld study sites. Dot-dash diagonal line (red) indicates 1:1 correspondence.



Applying  $\delta$  estimates within the statistical offsets resulted in reduced bias for 33 out of 35 species compared to modeling the data using offsets taken solely from human observer data (Fig. 4). Failing to incorporate  $\delta$  estimates within the statistical offsets resulted in 24 species with negative biases in their density estimates (Fig. 4). Of the 24 species with negatively biased density estimates derived using uncorrected offsets (taken from human observer data), five species (Ovenbird [*Seiurus aurocapilla*], Dark-eyed Junco [*Junco hyemalis*], Ruby-crowned Kinglet [*Regulus calendula*], Clay-colored Sparrow [*Spizella pallida*], and Connecticut Warbler [*Oporornis agilis*]) had 95% CIs that did not overlap zero (Fig. 4), whereas density estimates for these same species were unbiased when  $\delta$  was incorporated in the offsets (Fig. 4). For Ovenbird, failing to incorporate  $\delta$  within the offset resulted in density being underestimated by 0.10 birds  $\text{ha}^{-1}$  on average. Similarly, densities of Dark-eyed Junco, Ruby-crowned Kinglet, Clay-colored Sparrow, and Connecticut Warbler were underestimated by 0.07 birds  $\text{ha}^{-1}$ , 0.04 birds  $\text{ha}^{-1}$ , 0.02 birds  $\text{ha}^{-1}$ , and

**Fig. 4.** Bias in estimated densities (birds  $\text{ha}^{-1}$ ) from point count data derived from autonomous recording units (ARUs) compared to densities derived from human point counts conducted at the same time and location. Densities from human point counts were derived by adjusting counts for biases in availability ( $p$ ) and perceptibility ( $q$ ) using QPAD approach (Sólymos 2013) by inclusion in a statistical offset (i.e.,  $\log[\text{EDR}_H^2 \cdot \pi \cdot p]$ , see Methods). Densities from ARU surveys were derived by applying QPAD offsets from human observer data (open circles -  $\log[\text{EDR}_H^2 \cdot \pi \cdot p]$ ), versus adjusting the offsets to account for the scaling constant  $\delta$  (i.e.,  $\log[(\delta * \text{EDR}_H^2) \cdot \pi \cdot p]$ , closed red circles). Bias was estimated from fitting models to 70% of the study sites and validated against the withheld external validation sites (30%) over 50 repeated random subsamples of the data. See Table 2 for scientific names of bird species.



0.01 birds  $\text{ha}^{-1}$ , respectively. Conversely, estimates of Philadelphia Vireo (*Vireo philadelphicus*) and Cedar Waxwing (*Bombycilla cedrorum*) densities were less biased on average when the statistical offsets did not incorporate  $\delta$ ; however, 95% CIs overlapped zero for both offset approaches (Fig. 4), suggesting both methods produced unbiased estimates. Although the remainder of the species had 95% CIs that overlapped zero for both of the statistical offset approaches, incorporating  $\delta$  also (on average) reduced overestimation of densities (Fig. 4). For example, using uncorrected versus  $\delta$  corrected statistical offsets resulted in greater overestimation of density on average for Cape May Warbler (*Setophaga tigrina*; 0.02 vs. 0.01 birds  $\text{ha}^{-1}$ ), Palm Warbler

**Table 2.** Model selection based on Akaike's Information Criterion (AIC) for Poisson generalized linear mixed effects models examining variation in counts of 35 species of boreal forest birds. Presented are  $\Delta$ AIC values that rank models relative to the model with the lowest AIC value, with the lowest value (i.e., 0) representing the most parsimonious model.

Species	Null	Survey Type	Survey Type + Habitat	Survey Type * Habitat	Noise + Habitat	Noise * Habitat
Yellow-bellied Sapsucker ( <i>Sphyrapicus varius</i> )	2.41	0.30	<b>0.00</b>	3.46	1.30	4.33
Hairy Woodpecker ( <i>Picooides villosus</i> )	0.16	<b>0.00</b>	1.98	5.84	1.41	5.28
Pileated Woodpecker ( <i>Dryocopus pileatus</i> )	2.54	4.04	<b>0.00</b>	3.75	4.37	7.91
Alder Flycatcher ( <i>Empidonax alnorum</i> )	<b>0.00</b>	1.71	3.55	5.53	3.66	7.29
Least Flycatcher ( <i>Empidonax minimus</i> )	<b>0.00</b>	1.51	3.47	6.93	3.38	7.38
Philadelphia Vireo ( <i>Vireo philadelphicus</i> )	<b>0.00</b>	0.66	2.33	6.29	2.22	5.44
Black-capped Chickadee ( <i>Poecile atricapillus</i> )	<b>0.00</b>	1.45	3.24	7.07	2.70	6.62
Red-breasted Nuthatch ( <i>Sitta canadensis</i> )	0.93	2.70	4.11	8.09	<b>0.00</b>	3.98
Brown Creeper ( <i>Certhia americana</i> )	<b>0.00</b>	1.53	3.48	7.26	1.90	5.76
Winter Wren ( <i>Troglodytes hiemalis</i> )	1.52	1.44	<b>0.00</b>	3.98	1.97	5.94
Ruby-crowned Kinglet ( <i>Regulus calendula</i> )	12.52	8.77	<b>0.00</b>	3.93	9.65	13.58
Swainson's Thrush ( <i>Catharus ustulatus</i> )	10.13	11.93	12.41	16.25	<b>0.00</b>	3.76
Hermit Thrush ( <i>Catharus guttatus</i> )	2.16	0.89	2.89	6.72	<b>0.00</b>	3.99
American Robin ( <i>Turdus migratorius</i> )	8.66	3.39	2.37	6.20	<b>0.00</b>	3.21
Cedar Waxwing ( <i>Bombycilla cedrorum</i> )	<b>0.00</b>	1.91	3.41	7.38	2.78	5.84
Ovenbird ( <i>Seiurus aurocapilla</i> )	20.86	18.64	<b>0.00</b>	3.68	17.41	21.40
Orange-crowned Warbler ( <i>Oreothlypis celata</i> )	<b>0.00</b>	1.25	0.20	4.18	3.25	5.29
Nashville Warbler ( <i>Oreothlypis ruficapilla</i> )	<b>0.00</b>	1.75	1.27	5.27	3.71	6.36
Connecticut Warbler ( <i>Oporornis agilis</i> )	<b>0.00</b>	1.19	2.49	6.29	3.18	6.88
Mourning Warbler ( <i>Geothlypis philadelphia</i> )	4.76	6.13	<b>0.00</b>	3.91	7.89	10.56
Common Yellowthroat ( <i>Geothlypis trichas</i> )	<b>0.00</b>	1.88	3.58	7.51	3.84	7.07
American Redstart ( <i>Setophaga ruticilla</i> )	<b>0.00</b>	1.99	0.26	3.58	3.49	7.18
Cape May Warbler ( <i>Setophaga tigrina</i> )	<b>0.00</b>	1.80	3.77	7.71	3.44	7.23
Magnolia Warbler ( <i>Setophaga magnolia</i> )	5.21	6.75	8.33	12.18	<b>0.00</b>	3.85
Bay-breasted Warbler ( <i>Setophaga castanea</i> )	<b>0.00</b>	1.28	2.98	6.93	2.98	6.69
Blackburnian Warbler ( <i>Setophaga fusca</i> )	<b>0.00</b>	1.95	3.65	6.37	3.94	6.96
Palm Warbler ( <i>Setophaga palmarum</i> )	<b>0.00</b>	1.98	2.62	4.60	3.98	7.77
Yellow-rumped Warbler ( <i>Setophaga coronata</i> )	9.79	4.40	<b>0.00</b>	4.00	5.73	9.64
Canada Warbler ( <i>Cardellina canadensis</i> )	1.27	3.24	<b>0.00</b>	4.00	5.23	9.15
Clay-colored Sparrow ( <i>Spizella pallida</i> )	1.40	<b>0.00</b>	1.56	4.98	1.93	3.39
Fox Sparrow ( <i>Passerella iliaca</i> )	<b>0.00</b>	1.96	2.67	5.33	3.80	7.31
Lincoln's Sparrow ( <i>Melospiza lincolni</i> )	<b>0.00</b>	1.90	2.86	4.39	3.87	7.68
White-throated Sparrow ( <i>Zonotrichia albicollis</i> )	<b>0.00</b>	1.61	3.19	7.11	3.60	7.39
Dark-eyed Junco ( <i>Junco hyemalis</i> )	25.81	26.44	0.00	3.43	27.51	30.13
Rose-breasted Grosbeak ( <i>Pheucticus ludovicianus</i> )	5.23	7.17	0.00	3.99	9.12	12.96

(*Setophaga palmarum*; 0.03 vs. 0.00 birds ha<sup>-1</sup>), American Redstart (*Setophaga ruticilla*; 0.03 vs. 0.02 birds ha<sup>-1</sup>), Brown Creeper (*Certhia americana*; 0.03 vs. 0.01 birds ha<sup>-1</sup>), Orange-crowned Warbler (*Oreothlypis celata*; 0.03 vs. 0.01 birds ha<sup>-1</sup>), and Nashville Warbler (*Oreothlypis ruficapilla*; 0.04 vs. 0.01 birds ha<sup>-1</sup>).

Based upon AIC model selection, the null model was the most parsimonious for 18 species, the model only including the survey type factor was the most parsimonious for two species, the model including factors for both survey and habitat type was the most parsimonious for 10 species, and the model including survey type and noise was the most parsimonious model for five species (Table 2). There was substantial model uncertainty for virtually all species; however, neither of the interaction models received substantial support (Table 2). Across species, the minimum  $\Delta AIC$  (3.43) for the survey type by habitat type interaction was observed for Dark-eyed Junco, and parameter estimates from the interaction in that model show little evidence for an effect ( $\beta = -0.38$ , SE = 0.51). Similarly, the most substantial support for the survey type by noise interaction was observed for American Robin (*Turdus migratorius*;  $\Delta AIC = 3.21$ ), which similarly showed little evidence for an effect ( $\beta = 0.17$ , SE = 0.32).

## DISCUSSION

Our results provide further evidence supporting the conclusions of previous researchers (Haselmayer and Quinn 2000, Hobson et al. 2002, Celis-Murillo et al. 2009, Blumstein et al. 2011) that the raw counts derived from both acoustic recordings and human observers are relatively comparable. In paired comparisons between ARU and human observers, we found that the null models were favored over models incorporating a survey type effect for 18 out of 35 species. This was further supported by parameter estimates for the survey effect ( $\delta$ ) that overlapped one for the majority of the species that we modeled. Together, these lines of evidence suggest occasional minor biases exist in count data from ARUs relative to human observations.

Despite the relative similarity of many of the raw counts, systematic biases were apparent, and for five species the 95% CIs for estimated bias in densities did not overlap zero if the statistical offsets did not incorporate  $\delta$ ; thus analytically dealing with these biases will be important for data integration. This may be especially important since the biases may differ between acoustic recorder types and brands (Rempel et al. 2013, Yip et al. 2017) and/or may change with equipment wear (Turgeon et al. 2017). We demonstrated that correcting ARU data for differential detectability can effectively remove the majority of these biases. Because our experimental design and statistical analysis resulted in similar species availability, we suggest that the key source of bias in the counts derives from differences in detection radius between human observers and ARUs.

Our results suggest that the relatively simple approach of pairing human observers with ARUs allows data to be successfully corrected for systematic biases between counts. Repeated random subsampling of our data suggested that  $\delta$  estimates were relatively robust to sampling variation and were correlated with empirical estimates calculated from withheld data from independent study sites. In addition, applying offsets incorporating  $\delta$  reduced bias for almost every species examined. Furthermore, we found little

support for interactions between survey type and habitat type and survey type and environmental noise effects on  $\delta$ . The lack of support for the interaction models suggests that biases between ARUs and human observers are apparently relatively consistent (but see below). Therefore, paired sampling can generally be used to derive corrections that can be readily obtained using relatively common Poisson (GLM or GLMM) regression models simply by including survey type as a factor available in most modern statistical software.

In addition to our approach facilitating integration of ARU data with human point counts, it has the added benefit that avian density estimates can be derived from single ARUs as long as the human observers include distance estimation in their survey protocol. Alternative methods exist to derive densities from ARUs, typically employing acoustic localization from arrays of synchronized ARUs (Dawson and Efford 2009, Campbell and Francis 2012, Mennill et al. 2012). Acoustic arrays require a greater financial investment in ARUs because multiple ARUs are required for an array and each is more expensive owing to additional hardware (GPS). For example, as of the time of writing, Wildlife Acoustics Inc. (<http://www.wildlifeacoustics.com/store#song-meter-sm3>) charges US\$1049 for a single SM3 ARU plus an additional US\$299 for the accompanying GPS module. Although acoustic localization is rapidly evolving, it can be logistically and computationally difficult. Thus, our approach provides a logistically feasible and affordable alternative to more complicated designs. Future paired comparisons between ARU and/or human point counts placed within acoustic arrays or traditional spot mapping grids (Bart and Earnst 2002) would further improve certainty in point count density estimation and would naturally fit with the analytical approach we describe here.

An alternative method to correcting biases between ARUs and human point counts would be through playback experiments. Yip et al. (2017) conducted playback experiments in which species calls were played along transects at various distances away from human observers and ARUs. An experimental approach has the advantage of the calls coming from known distances, however it also requires the experimenter to make assumptions about the amplitude at which birds sing/call because how loud wild birds sing is variable and generally unknown (Brackenbury 1979). In addition, the effects of directionality on song amplitude in wild birds are not well described but are known to impact detection probability from experimental playbacks (Allredge et al. 2007a, c). Therefore experimentally replicating the impact of bird orientation relative to point count location complicates the approach. In contrast, with respect to density estimation our approach assumes that observers accurately estimated distance, which can be inaccurate (Allredge et al. 2007c). We suggest that both paired comparisons and experimental playbacks could be used in a complementary fashion to estimate correction factors between human point counts and ARUs. Further, we suggest that paired sampling is a pragmatic approach to obtain statistical offsets for the majority of species and does not require assumptions about amplitude and directionality of songs. Where sample sizes become limiting because of species rarity, the experimental approach of Yip et al. (2017) would allow estimates to be obtained for species for which  $\delta$  cannot be estimated because of a lack of detections.



Similar to our results, Yip et al. (2017) generally found estimates of  $\delta$  that were less than 1. Unlike our results however, Yip et al. (2017) found evidence for habitat related variation in  $\delta$ . Presumably detection for both human observers and ARUs were similarly affected by habitat and environmental noise in our experiment and thus our density estimates may be biased low; however, systematic differences between survey types were apparently corrected. Given the apparent difference between our results and those of Yip et al. (2017), future analyses under a broader set of habitat conditions and with a broader range of species may provide evidence suggesting the need for stratification to improve the corrections we have employed here. For example, we did not fit observer or habitat specific distance models that would presumably improve precision because there can be substantial interobserver variation in distance estimation (Nadeau and Conway 2012). Despite not having estimated observer or habitat specific offsets, our validation still suggests a substantial reduction in bias despite randomly sampling among habitats and observers. Greater effort should be put into replicating our design with more combinations of species, habitats, and environmental conditions to facilitate estimating how much annual effort should be placed on paired sampling because the added time for transcription is an added cost of our method.

Our results and external validation provide evidence that data from both human observers and ARUs can be placed on a similar footing. We therefore recommend monitoring and research programs begin further integration of ARUs and human observed point counts to take advantage of the relative merits of both methods. Not only would this improve sample sizes, but would also allow researchers to gain a better understanding of factors influencing detection probability owing to the ease of obtaining repeated samples with programmable ARUs. Although further sampling could provide refinements to our estimates, we have shown that our approach reduces bias related to survey type. Our method therefore provides an easily implemented method that facilitates the integration of ARU data with human observer point counts to allow expanded monitoring efforts and will facilitate meta-analyses with historic point count data to examine factors influencing avian populations (Cumming et al. 2010, Sóllymos et al. 2013).

Responses to this article can be read online at:  
<http://www.ace-eco.org/issues/responses.php/975>

---

#### Acknowledgments:

This work was supported by operating grants to SVW from the Wildlife and Habitat Assessment Section (PR) of the Canadian Wildlife Service - Environment and Climate Change Canada. We thank E. Cumming, B. Obermayer, and C. Chutter for field assistance on this project. This publication is a contribution of the Boreal Avian Modelling (BAM) Project, an international research collaboration on the ecology, management, and conservation of boreal birds. We acknowledge BAM's members, avian and biophysical Data Partners, and funding agencies (including Environment and Climate Change Canada and the U.S. Fish &

Wildlife Service), listed in full at <http://www.borealbirds.ca/index.php/placknowledgements>. We would like to thank K. Hobson, B. Klingbeil, and two anonymous reviewers for comments that helped us to improve the manuscript.

---

#### LITERATURE CITED

- Allredge, M. W., K. H. Pollock, T. R. Simons, J. A. Collazo, and S. A. Shriner. 2007b. Time-of-detection method for estimating abundance from point-count surveys. *Auk* 124:653-664. [http://dx.doi.org/10.1642/0004-8038\(2007\)124\[653:TMFEAF\]2.0.CO;2](http://dx.doi.org/10.1642/0004-8038(2007)124[653:TMFEAF]2.0.CO;2)
- Allredge, M. W., T. R. Simons, and K. H. Pollock. 2007c. A field evaluation of distance measurement error in auditory avian point count surveys. *Journal of Wildlife Management* 71(8):2759-2766. <http://dx.doi.org/10.2193/2006-161>
- Allredge, M. W., T. R. Simons, K. H. Pollock, and K. Pacifici. 2007a. A field evaluation of the time-of-detection method to estimate population size and density for aural avian point counts. *Avian Conservation and Ecology* 2(2):13. <https://doi.org/10.5751/ACE-00205-020213>
- Bart, J., and S. Earnst. 2002. Double sampling to estimate density and population trends in birds. *Auk* 119:36-45. [http://dx.doi.org/10.1642/0004-8038\(2002\)119\[0036:DSTEDA\]2.0.CO;2](http://dx.doi.org/10.1642/0004-8038(2002)119[0036:DSTEDA]2.0.CO;2)
- Blumstein, D. T., D. J. Mennill, P. Clemins, L. Girod, K. Yao, G. Patricelli, J. L. Deppe, A. H. Krakauer, C. Clark, K. A. Cortopassi, S. F. Hanser, B. McCowan, A. M. Ali, and A. N. Kirschel. 2011. Acoustic monitoring in terrestrial environments using microphone arrays: applications, technological considerations and prospectus. *Journal of Applied Ecology* 48:758-767. <http://dx.doi.org/10.1111/j.1365-2664.2011.01993.x>
- Brackenbury, J. H. 1979. Power capabilities of the avian sound-producing system. *Journal of Experimental Biology* 78(1):163-166.
- Buckland, S. T., D. R. Anderson, K. P. Burnham, J. L. Laake, D. L. Borchers, and L. Thomas. 2001. *Introduction of distance sampling: estimating abundance of biological populations*. Oxford University Press, Oxford, UK.
- Burnham, K. P., and D. R. Anderson. 2002. *Model selection and inference: a practical information-theoretic approach*. Springer-Verlag, New York, New York, USA. <http://dx.doi.org/10.1007/978-1-4757-2917-7>
- Campbell, M., and C. M. Francis. 2012. Using microphone arrays to examine effects of observers on birds during point count surveys. *Journal of Field Ornithology* 83:391-402. <http://dx.doi.org/10.1111/j.1557-9263.2012.00389.x>
- Celis-Murillo, A., J. L. Deppe, and M. F. Allen. 2009. Using soundscape recordings to estimate bird species abundance, richness, and composition. *Journal of Field Ornithology* 80:64-78. <http://dx.doi.org/10.1111/j.1557-9263.2009.00206.x>
- Cumming, S. G., K. L. Lefevre, E. Bayne, T. Fontaine, F. K. A. Schmiegelow, and S. J. Song. 2010. Toward conservation of Canada's boreal forest avifauna: design and application of ecological models at continental extents. *Avian Conservation and Ecology* 5(2):8. <http://dx.doi.org/10.5751/ACE-00406-050208>

- Dawson, D. K., and M. G. Efford. 2009. Bird population density estimated from acoustic signals. *Journal of Applied Ecology* 46:1201-1209. <http://dx.doi.org/10.1111/j.1365-2664.2009.01731.x>
- Downes, C. M., M.-A. R. Hudson, A. C. Smith, and C. M. Francis. 2016. The Breeding Bird Survey at 50: scientists and birders working together for bird conservation. *Avian Conservation and Ecology* 11(1):8. <http://dx.doi.org/10.5751/ACE-00855-110108>
- Farnsworth, G. L., K. H. Pollock, J. D. Nichols, T. R. Simons, J. E. Hines, and J. R. Sauer. 2002. A removal model for estimating detection probabilities from point-counts surveys. *Auk* 119:414-425. [http://dx.doi.org/10.1642/0004-8038\(2002\)119\[0414:ARMFED\]2.0.CO;2](http://dx.doi.org/10.1642/0004-8038(2002)119[0414:ARMFED]2.0.CO;2)
- Francis, C. M., P. J. Blancher, and R. D. Phoenix. 2009. Bird monitoring programs in Ontario: What have we got and what do we need? *Forestry Chronicle* 85:202-217. <http://dx.doi.org/10.5558/tfc85202-2>
- Goyette, J. L., R. W. Howe, A. T. Wolf, and W. D. Robinson. 2011. Detecting tropical nocturnal birds using automated audio recordings. *Journal of Field Ornithology* 82:279-287. <http://dx.doi.org/10.1111/j.1557-9263.2011.00331.x>
- Haselmayer, J., and J. S. Quinn. 2000. A comparison of point counts and sound recording as bird survey methods in Amazonian southeast Peru. *Condor* 102:887-893. [http://dx.doi.org/10.1650/0010-5422\(2000\)102\[0887:ACOPCA\]2.0.CO;2](http://dx.doi.org/10.1650/0010-5422(2000)102[0887:ACOPCA]2.0.CO;2)
- Hobson, K. A., R. S. Rempel, H. Greenwood, B. Turnbull, and S. L. Van Wilgenburg. 2002. Acoustic surveys of birds using electronic recordings: new potential from an omnidirectional microphone system. *Wildlife Society Bulletin* 30:709-720.
- Hutto, R. L., and R. J. Stutzman. 2009. Humans versus autonomous recording units: a comparison of point-count results. *Journal of Field Ornithology* 80:387-398. <http://dx.doi.org/10.1111/j.1557-9263.2009.00245.x>
- Klingbeil, B. T., and M. R. Willig. 2015. Bird biodiversity assessments in temperate forest: the value of point count versus acoustic monitoring protocols. *PeerJ* 3:e973. <https://doi.org/10.7717/peerj.973>
- Latifovic, R., I. Olthof, D. Pouliot, and J. Beaubien. 2008. *Land cover map of Canada 2005 at 250m spatial resolution*. Natural Resources Canada/ESS/Canada Centre for Remote Sensing, Ottawa, Ontario, Canada. [online] URL: [ftp://ftp.ccrs.nrcan.gc.ca/ad/NLCC/LandCover/LandcoverCanada2005\\_250m/](ftp://ftp.ccrs.nrcan.gc.ca/ad/NLCC/LandCover/LandcoverCanada2005_250m/)
- Link, W. A., J. R. Sauer, and D. K. Niven. 2006. A hierarchical model for regional analysis of population change using Christmas Bird Count data, with application to the American Black Duck. *Condor* 108:13-24. [http://dx.doi.org/10.1650/0010-5422\(2006\)108\[0013:AHMFRA\]2.0.CO;2](http://dx.doi.org/10.1650/0010-5422(2006)108[0013:AHMFRA]2.0.CO;2)
- Machtans, C. S., K. J. Kardynal, and P. A. Smith. 2014. How well do regional or national Breeding Bird Survey data predict songbird population trends at an intact boreal site? *Avian Conservation and Ecology* 9(1):5. <http://dx.doi.org/10.5751/ACE-00649-090105>
- Matsuoka, S. M., E. M. Bayne, P. Sólymos, P. C. Fontaine, S. G. Cumming, F. K. A. Schmiegelow, and S. J. Song. 2012. Using binomial distance-sampling models to estimate the effective detection radius of point-count surveys across boreal Canada. *Auk* 129:268-282. <http://dx.doi.org/10.1525/auk.2012.11190>
- Matsuoka, S. M., C. L. Mahon, C. M. Handel, P. Sólymos, 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. <http://dx.doi.org/10.1650/CONDOR-14-108.1>
- Mennill, D. J., M. Battiston, D. R. Wilson, J. R. Foote, and S. M. Doucet. 2012. Field test of an affordable, portable, wireless microphone array for spatial monitoring of animal ecology and behaviour. *Methods in Ecology and Evolution* 3:704-712. <http://dx.doi.org/10.1111/j.2041-210X.2012.00209.x>
- Nadeau, C. P., and C. J. Conway. 2012. Field evaluation of distance-estimation error during wetland-dependent bird surveys. *Wildlife Research* 39:311-320. <http://dx.doi.org/10.1071/WR11161>
- Nichols, J. D., L. Thomas, and P. B. Conn. 2009. Inferences about landbird abundance from count data: recent advances and future directions. Pages 201-236 in D. L. Thomson, E. G. Cooch, M. J. Conroy, editors. *Modeling demographic processes in marked populations*. Environmental and Ecological Statistics series. Springer, New York, New York, USA. [http://dx.doi.org/10.1007/978-0-387-78151-8\\_9](http://dx.doi.org/10.1007/978-0-387-78151-8_9)
- Rempel, R. S., C. M. Francis, J. N. Robinson, and M. Campbell. 2013. Comparison of audio recording system performance for detecting and monitoring songbirds. *Journal of Field Ornithology* 84:86-97. <http://dx.doi.org/10.1111/jfo.12008>
- Sauer, J. R., J. E. Fallon, and R. Johnson. 2003. Use of North American Breeding Bird Survey data to estimate population change for bird conservation regions. *Journal of Wildlife Management* 67:372-389. <http://dx.doi.org/10.2307/3802778>
- Sauer, J. R., J. E. Hines, J. E. Fallon, K. L. Pardieck, D. J. Ziolkowski Jr., and W. A. Link. 2014. *The North American Breeding Bird Survey, results and analysis 1966 - 2013*. Version 01.30.2015. USGS Patuxent Wildlife Research Center, Laurel, Maryland, USA.
- Sidie-Slettedahl, A. M., K. C. Jensen, R. R. Johnson, T. W. Arnold, J. E. Austin, and J. D. Stafford. 2015. Evaluation of autonomous recording units for detecting 3 species of secretive marsh birds. *Wildlife Society Bulletin* 39:626-634. <http://dx.doi.org/10.1002/wsb.569>
- Sólymos, P., S. M. Matsuoka, E. M. Bayne, S. R. Lele, P. Fontiane, S. G. Cumming, D. Stralberg, F. K. A. Schmiegelow, and S. J. Song. 2013. Calibrating indices of avian density from non-standardized survey data: making the most of a messy situation. *Methods in Ecology and Evolution* 4:1047-1058. <http://dx.doi.org/10.1111/2041-210X.12106>
- Sólymos, P., M. Moreno, and S. R. Lele. 2016. *detect: Analyzing wildlife data with detection error*. R package version 0.4-0. The R Project for Statistical Computing, Vienna, Austria. [online] URL: <https://CRAN.R-project.org/package=detect>

Turgeon, P. J., S. L. Van Wilgenburg, and K. L. Drake. 2017. Microphone variability and degradation: implications for monitoring programs employing autonomous recording units. *Avian Conservation and Ecology* 12(1):9. <http://dx.doi.org/10.5751/ace-00958-120109>

Venier, L. A., S. B. Holmes, G. W. Holborn, K. A. Mcilwrick, and G. Brown. 2012. Evaluation of an automated recording device for monitoring forest birds. *Wildlife Society Bulletin* 36:30-39. <http://dx.doi.org/10.1002/wsb.88>

Yip, D. A., 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. <http://dx.doi.org/10.5751/ace-00997-120111>

Zwart, M. C., A. Baker, P. J. K. McGowan, and M. J. Whittingham. 2014. The use of automated bioacoustic recorders to replace human wildlife surveys: an example using nightjars. *PLoS ONE* 9(7):e102770. <http://dx.doi.org/10.1371/journal.pone.0102770>



## Appendix 1.

Table A1.1 Estimated effective detection radius (EDR) and parameter estimates from count removal models including both the human observer and autonomous recording unit (ARU) data with survey type included as a fixed effect factor (ARU (1) versus human as the reference category (0)) to test for a difference in species availability for 41 species (see Methods). Bold denotes parameter estimates for which 95% confidence intervals suggest the probability of an individual singing (and being detected) at least once during the cumulative duration of the count differed between human observer ( $p_H$ ) and transcription from ARU ( $p_A$ ).

Species	Latin Name	Sample size			Intercept		Survey Effect	
		Human	ARU	EDR (m)	$\beta$	SE	$\beta$	SE
Ruffed Grouse	<i>Bonasa umbellus</i>	43	87	95	-1.73	0.34	<b>0.84</b>	<b>0.37</b>
Yellow-bellied Sapsucker	<i>Sphyrapicus varius</i>	65	54	68	-2.10	0.36	-0.03	0.56
Hairy Woodpecker	<i>Picoides villosus</i>	18	16	53	-1.02	0.27	0.02	0.43
Northern Flicker	<i>Colaptes auratus</i>	39	24	128	-15.50	NA	12.95	NA
Pileated Woodpecker	<i>Dryocopus pileatus</i>	47	55	167	-1.71	0.31	0.21	0.39
Alder Flycatcher	<i>Empidonax alnorum</i>	69	64	102	-1.12	0.13	-0.22	0.22
Least Flycatcher	<i>Empidonax minimus</i>	18	19	46	-0.95	0.22	-0.10	0.32
Philadelphia Vireo	<i>Vireo philadelphicus</i>	24	30	53	-1.54	0.35	0.53	0.43
Red-eyed Vireo	<i>Vireo olivaceus</i>	229	217	81	-1.00	0.07	<b>0.30</b>	<b>0.09</b>
Black-capped Chickadee	<i>Poecile atricapilla</i>	33	31	56	-1.62	0.30	-0.16	0.48
Red-breasted Nuthatch	<i>Sitta canadensis</i>	65	56	88	-2.12	0.38	-0.31	0.67
Brown Creeper	<i>Certhia americana</i>	47	63	42	-1.28	0.21	-0.14	0.30
Winter Wren	<i>Troglodytes hiemalis</i>	111	100	101	-0.89	0.11	0.09	0.16
Ruby-crowned Kinglet	<i>Regulus calendula</i>	98	77	86	-1.21	0.12	0.18	0.18
Swainson's Thrush	<i>Catharus ustulatus</i>	135	136	91	-1.23	0.11	0.20	0.14
Hermit Thrush	<i>Catharus guttatus</i>	163	145	139	-1.29	0.10	0.03	0.15
American Robin	<i>Turdus migratorius</i>	64	52	92	-1.94	0.30	0.68	0.36
Cedar Waxwing	<i>Bombycilla cedrorum</i>	42	49	34	-2.51	0.57	-0.08	0.83
Ovenbird	<i>Seiurus aurocapilla</i>	203	201	77	-0.91	0.06	0.14	0.08
Tennessee Warbler <sup>†</sup>	<i>Oreothlypis peregrina</i>	200	192	63	-1.08	0.07	<b>0.25</b>	<b>0.10</b>
Orange-crowned Warbler	<i>Oreothlypis celata</i>	43	35	59	-2.07	0.43	0.18	0.58
Nashville Warbler	<i>Oreothlypis ruficapilla</i>	33	41	46	-1.55	0.30	0.16	0.39
Connecticut Warbler	<i>Oporornis agilis</i>	24	25	74	-1.28	0.25	0.48	0.34
Mourning Warbler	<i>Geothlypis philadelphia</i>	73	76	64	-1.31	0.16	0.15	0.22
Common Yellowthroat	<i>Geothlypis trichas</i>	19	19	62	-0.91	0.27	-0.24	0.41
American Redstart	<i>Setophaga ruticilla</i>	40	42	51	-0.66	0.14	-0.05	0.19
Cape May Warbler	<i>Setophaga tigrina</i>	22	27	37	-1.31	0.32	0.40	0.40
Magnolia Warbler	<i>Setophaga magnolia</i>	122	122	54	-1.24	0.11	0.21	0.16
Bay-breasted Warbler	<i>Setophaga castanea</i>	62	59	39	-1.60	0.22	0.25	0.30
Blackburnian Warbler	<i>Setophaga fusca</i>	40	39	41	-0.93	0.19	-0.60	0.34

Chestnut-sided Warbler	<i>Setophaga pensylvanica</i>	51	47	55	-1.13	0.15	<b>0.43</b>	<b>0.20</b>
Palm Warbler	<i>Setophaga palmarum</i>	34	37	65	-1.51	0.26	0.19	0.34
Yellow-rumped Warbler	<i>Setophaga coronata</i>	134	115	64	-1.72	0.16	0.35	0.21
Canada Warbler	<i>Cardellina canadensis</i>	20	22	61	-0.82	0.24	0.49	0.34
Chipping Sparrow	<i>Spizella passerina</i>	160	151	73	-2.10	0.21	<b>0.66</b>	<b>0.24</b>
Clay-colored Sparrow	<i>Spizella pallida</i>	20	15	70	-1.25	0.29	-0.31	0.55
Fox Sparrow	<i>Passerella iliaca</i>	17	19	131	-1.59	0.44	-0.04	0.60
Lincoln's Sparrow	<i>Melospiza lincolnii</i>	26	28	69	-0.82	0.23	-0.22	0.34
White-throated Sparrow	<i>Zonotrichia albicollis</i>	260	259	99	-1.15	0.06	0.08	0.08
Dark-eyed Junco	<i>Junco hyemalis</i>	76	65	70	-1.89	0.23	0.54	0.29
Rose-breasted Grosbeak	<i>Pheucticus ludovicianus</i>	33	39	85	-0.87	0.21	-0.40	0.32

---

Table A1.2 Estimated scaling constants ( $\delta$  and both lower (LCL) and upper (UCL) 95% confidence intervals) relating counts from autonomous recording units (ARUs) to those by human observers for 35 species and calculated as  $\sqrt{\exp(\beta)}$  from generalized linear mixed effects regression models (see Methods). Presented are maximum likelihood (MLE) estimates of  $\delta$  and 95% CIs estimated as the median and the 2.5 and 97.5% quantiles from 10,000 Monte Carlo simulations drawn from a multivariate normal distribution treating the model coefficients (including the survey effect  $\beta$ ) as the mean and the estimated variance-covariance matrix. Also presented are calibration  $\delta$  and empirical ratios (dividing the mean ARU count by the mean human observer count) derived from 50 repeated random (70% versus 30%) sub-samples of the data (see Methods).

Species	MLE estimates			Repeated Random Sampling					
	LCL	$\delta$	UCL	Calibration			Empirical Ratio		
	LCL	$\delta$	UCL	LCL	$\delta$	UCL	LCL	$\delta$	UCL
Yellow-bellied Sapsucker	0.71	0.86	1.03	0.80	0.87	0.94	0.66	0.85	1.20
Hairy Woodpecker	0.62	0.84	1.14	0.71	0.84	0.99	0.41	1.03	2.12
Pileated Woodpecker	0.84	1.03	1.26	1.00	1.06	1.11	0.96	1.08	1.23
Alder Flycatcher	0.76	0.90	1.06	0.81	0.86	0.92	0.64	0.93	1.38
Least Flycatcher	0.73	0.93	1.18	0.90	0.95	1.03	0.81	0.97	1.24
Philadelphia Vireo	0.77	1.00	1.31	0.86	1.00	1.16	0.73	0.97	1.22
Black-capped Chickadee	0.70	0.88	1.11	0.83	0.91	0.97	0.66	0.92	1.35
Red-breasted Nuthatch	0.74	0.90	1.08	0.89	0.93	0.99	0.82	0.91	1.16
Brown Creeper	0.83	1.01	1.23	0.97	1.05	1.14	0.91	1.11	1.37
Winter Wren	0.78	0.91	1.05	0.88	0.92	0.94	0.81	0.91	1.18
Ruby-crowned Kinglet	0.76	0.87	1.00	0.83	0.86	0.90	0.68	0.80	1.11
Swainson's Thrush	0.91	1.01	1.12	0.96	0.99	1.03	0.91	1.00	1.18
Hermit Thrush	0.84	0.93	1.02	0.88	0.91	0.94	0.82	0.91	1.05
American Robin	0.68	0.82	0.98	0.77	0.82	0.88	0.62	0.83	1.21
Cedar Waxwing	0.85	1.02	1.23	0.94	1.00	1.07	0.90	1.04	1.25
Ovenbird	0.86	0.92	0.99	0.91	0.93	0.95	0.84	0.90	1.03
Orange-crowned Warbler	0.75	0.95	1.20	0.94	1.10	1.21	0.70	1.10	1.49
Nashville Warbler	0.87	1.10	1.39	1.00	1.11	1.21	0.70	1.10	1.49
Connecticut Warbler	0.69	0.88	1.13	0.76	0.85	0.91	0.57	0.73	0.83
Mourning Warbler	0.86	1.01	1.18	0.90	0.98	1.06	0.80	0.99	1.29
Common Yellowthroat	0.74	1.00	1.34	0.87	0.95	1.05	0.76	1.14	2.17
American Redstart	0.92	1.09	1.30	0.98	1.04	1.11	0.79	1.03	1.24
Cape May Warbler	0.80	1.06	1.42	0.97	1.07	1.17	0.76	1.09	1.43
Magnolia Warbler	0.82	0.93	1.05	0.92	0.95	0.98	0.85	0.96	1.12
Bay-breasted Warbler	0.77	0.92	1.09	0.90	0.94	0.99	0.80	0.95	1.18
Blackburnian Warbler	0.80	1.00	1.25	0.85	0.94	1.02	0.74	0.95	1.27
Palm Warbler	0.87	1.11	1.42	0.97	1.10	1.23	0.70	1.09	1.55
Yellow-rumped Warbler	0.79	0.89	1.00	0.84	0.89	0.94	0.77	0.89	1.10
Canada Warbler	0.73	1.00	1.36	0.94	1.00	1.07	0.90	1.05	1.27
Clay-colored Sparrow	0.55	0.77	1.09	0.65	0.78	0.86	0.71	0.75	0.81
Fox Sparrow	0.84	1.12	1.49	1.01	1.09	1.19	0.87	1.13	1.44
Lincoln's Sparrow	0.81	1.12	1.55	0.86	0.99	1.19	0.79	1.15	1.85
White-throated Sparrow	0.91	0.97	1.03	0.96	0.98	1.00	0.92	0.97	1.08

Dark-eyed Junco	0.71	0.84	0.98	0.77	0.85	0.94	0.62	0.87	1.29
Rose-breasted Grosbeak	0.86	1.10	1.39	0.97	1.06	1.13	0.86	1.05	1.25

---