2018

Quantifying False Positives in Avian Survey Data

Kaitlyn M. Strickfaden
kaitlyn.strickfaden@umontana.edu

Quantifying false positives in avian survey data
10 May 2018

Kaitlyn Strickfaden, Wildlife Biology Program (Senior), Avian Science Center, W.A. Franke College of Forestry and Conservation, University of Montana, 32 Campus Drive, Missoula, MT 59812, USA.
E-Mail: kaitlyn.strickfaden@umontana.edu

Abstract

Imperfect detection is a known issue when conducting count-based surveys in wildlife studies. False positive detections, observed occurrences of individuals that truly are not present, are often assumed to not occur. This assumption can bias detection rates and create misleading results when calculating population estimates. Survey methods such as the dependent double-observer method are suggested to reduce the occurrence of false positives (Nichols et al. 2000). My study quantified and compared rates of false positives in a single-observer method and a dependent double-observer method using computer-generated auditory surveys. I categorized volunteer observers as either inexperienced or experienced and asked them to identify vocalizations of ten grassland songbird species native to central Montana. False positive rates of experienced observers declined from 0.095 in single-observer surveys to 0.032 in dependent double-observer surveys. False positive rates of inexperienced observers declined from 0.511 in single-observer surveys to 0.391 in dependent double-observer surveys. Further evaluation will provide information on the effectiveness of the dependent double-observer method in providing more precise and less biased population estimates.

Introduction

Imperfect detection is a prevalent issue in count-based data used in wildlife studies. Imperfect detection takes two forms: false negatives, when an individual is not counted as present when it truly is present; and false positives, when an individual is counted as present when it truly is not present (Royle and Link 2006, Fitzpatrick et al 2009, Miller at al. 2011, Miller et al. 2012, Connors et al. 2014, Miller et al. 2015). False negatives result from observers' inability to see or hear every individual within the sampled area. False positives result from either misidentifying or double-counting individuals. Mechanisms are currently used to include false negative errors in estimates of imperfect detection when calculating abundance. However, false positives are often left unaddressed because calculating a false positive rate relies on the assumption that true presence at field sites is confidently known. (Royle 2004). Ignoring false positive errors in detection estimates significantly biases population estimates (Miller et al. 2015). Because many population estimates use count-based data to extrapolate total occupancy or abundance, even small inaccuracies in detection estimates can seriously inflate or deflate population estimates. The magnitude of biases also often varies by species in multispecies surveys. Given the prevalence of both abundance and occupancy estimates in wildlife management, it is important to understand how often false positives occur and how to account for their occurrence in deriving population estimates.

The independent single-observer (ISO) method is the most common wildlife survey method. ISO surveys are flexible in that the duration of surveys and recording method (simple counts, distance sampling, etc.) can be adjusted to fit the needs and goals of the study. False negative rates can be calculated after several return trips to the same plot. However, this survey method is limited in its ability to determine the occurrence of false positive errors. Researchers recognize that false positives occur but almost never account for them in calculating detection rates. Instead, researchers rely on experienced observers to avoid false positives in ISO surveys (Aldredge et al. 2008). However, even experienced observers report false positives (Aldredge et al. 2008; Miller et al. 2012). The few studies that have calculated false positive rates (Aldredge 2008, Fitzpatrick et al. 2009, Miller et al. 2012) have used ISO or independent double-observer survey methods.

It has been suggested that the dependent double-observer (DDO) method reduces the occurrence of false positives (Nichols et al. 2000; Golding and Dreitz 2017), but this idea has not yet been tested. The DDO method is based on capture-recapture removal methodology that treats the 'captures' of counted individuals as being removed by one of the two observers in a two-person observer team (Nichols et al. 2000, Golding and Dreitz 2016). During the count survey, the primary observer reports all observations to the secondary observer, who records the observations. In addition, the secondary observer records any observations that the primary observer failed to detect. The secondary observer must avoid cueing the primary observer to any observations that the primary observer missed. Observers can collaborate with each other in identifying an individual as long as detection is attributed to the correct observer. This method has been successfully applied to surveys of songbirds (Tipton et al. 2009; Golding and Dreitz 2016; Golding et al. 2017; Leston et al. 2015), butterflies (Henry and Anderson 2016), crocodiles (Shirley et al. 2011), and gull nests (Barbraud et al. 2005).

Avian surveys are an optimal means of studying imperfect detection in complex survey situations. Avian surveys often sample several different species. Many bird species are difficult to confidently distinguish from other species with only a quick visual or aural stimulus. The DDO method addresses this problem in avian surveys by providing an observer with a resource (i.e. another observer) to detect additional birds within the sampling plot and keep track of the birds during surveys. This survey method has already proven to be successful in measuring grassland bird abundance on Colorado shortgrass prairie (Tipton et al. 2009), Montana sagebrush grasslands (Golding and Dreitz 2017), and Alberta mixed-grass prairie (Leston et al. 2015).

Though the importance of avian monitoring programs such as EBird and the Integrated Monitoring in Bird Conservation Regions (IMBCR) program to conservation efforts cannot be denied, the likelihood of false positive errors calls into question the credibility of the biological inferences drawn from these monitoring programs. EBird entries are reported by observers with a wide variety of bird-identifying experience, so identification in these surveys could range from flawless to spurious. IMBCR employs a training process for observers, but Miller et al. (2012) still reported a notable false positive rate despite using experienced observers. In addition, Aldredge et al. (2008) found that experienced observers performing distance sampling surveys in forests often could not reliably determine the direction of a call. EBird and IMBCR data are invaluable, and they should not be entirely discredited, but false positives must be addressed if they are to continue to be common sources of data to inform conservation.

I answer whether there are significant differences in false positive rates between the ISO and DDO survey methods with auditory data in which truth is known. My ‘truth’ data consist of randomly-generated vocalizations of grassland bird species native to Montana. Unlike in field surveys, the true identity of each individual is known from a generated list to determine when a false positive error occurs within a survey. I considered two different groups of observers: 1) experienced, having prior experience identifying the selected study species, or 2) inexperienced, having no prior experience identifying the study species. Both groups performed 3-minute auditory ISO and DDO surveys. My objectives were to compare false positive rates between 1) the ISO and DDO survey methodologies and 2) experienced and inexperienced observers. I predicted that false positive rates would be lower in DDO surveys than in ISO surveys. I also predicted that false positive rates would vary between species. Lastly, I predicted that false positive rates would be lower in experienced observers than in inexperienced observers and that both observer groups would report false positives.

Methods

I chose vocalizations from ten avian species that commonly occur in eastern Montana: Brewer's Sparrow (*Spizella breweri*), Brown-Headed Cowbird (*Molothrus ater*), Horned Lark (*Eremophila alpestris*), Killdeer (*Charadrius vociferous*), Lark Bunting (*Calamospiza melanocorys*), Long-Billed Curlew (*Numenius americanus*), McCown's Longspur (*Rhynchophanes mccownii*), Savannah Sparrow (*Passerculus sandwichensis*), Vesper Sparrow (*Pooectes graminius*), and Western Meadowlark (*Sturnella neglecta*). Some species have vocalizations that are similar to other species in this study whereas other have vocalizations that are very unique. I retrieved the vocalizations from the Macauley Library of the Cornell Lab of Ornithology (https://www.macaulaylibrary.org). I filtered each vocalization to remove background noise that might cue the observers to the identity of the call (Audacity Team; Bioacoustics Research Program) and then clipped each vocalization to four seconds of audio space (hereafter, 'audio clip').

Survey generation – Three-minute auditory surveys were generated which contained both bird vocalizations and white noise. For each species' audio clip, I first randomly selected a number between one and four to be the number of times that species' audio clip was played during a particular survey. After repeating this for all ten study species, I randomly generated a list of where each audio clip would occur within the 3-minute survey. White noise was placed where no audio clip occurred to complete the 3-minute survey. I generated 1000 unique surveys. These surveys had a mean of 18.3 vocalizations per survey, a standard deviation of 2.8 vocalizations, and a range of 11-27 vocalizations per survey. The surveys were generated in R (R CoreTeam).

Data collection - Observers consisted of 13 volunteers with a wide range of bird identification experience. Seven observers were categorized as inexperienced and six were categorized as experienced based on self-assessments of ability conducted by each observer. For both the ISO and DDO surveys, observers were seated in a room with little noise or distraction. The 3-minute auditory surveys were played through the speaker of a Dell laptop computer placed in front of the observer or the two observers. Observers recorded vocalizations they heard within three-second time intervals (Fig. 1). The 3-second interval was determined to be the average amount of time needed for an observer to report an observation. A timer on the computer screen kept track of time throughout the survey and helped observers to properly record observations. No visual cues were provided to aid in identification. Example vocalizations for each study species were provided before surveys, but no training was required to take part in surveys.

For ISO surveys, the observer listened to a randomly-assigned survey and recorded the species heard by circling or highlighting the name of the species in the proper time interval. For DDO surveys, the two observers were seated in view of the computer screen and timer. The two observers were randomly assigned the role of primary or secondary observer. The secondary observer recorded all observations voiced by the primary observer as well as any observations they detected that the primary observer failed to detect. An additional column on the DDO datasheet provided space to indicate observations that were only detected by the secondary observer. Observers switched roles after each DDO survey.

Interval	Observer	Brewer's Sparrow	Brown-Headed Cowbird	Horned Lark	Killdeer	Lark Bunting	Long-Billed Curlew	McCown's Longspur	Savannah Sparrow	Vesper Sparrow	Western Meadowlark
0:00-0:03	S	Brewer's Sp.	Cowbird	H. Lark	Killdeer	L. Bunting	L.B. Curlew	M. Longspur	Savannah Sp.	Vesper Sp.	W. Meadowlark
0:03-0:06	S	Brewer's Sp.	Cowbird	H. Lark	Killdeer	L. Bunting	L.B. Curlew	M. Longspur	Savannah Sp.	Vesper Sp.	W. Meadowlark
0:06-0:09	S	Brewer's Sp.	Cowbird	H. Lark	Killdeer	L. Bunting	L.B. Curlew	M. Longspur	Savannah Sp.	Vesper Sp.	W. Meadowlark
0:09-0:12	S	Brewer's Sp.	Cowbird	H. Lark	Killdeer	L. Bunting	L.B. Curlew	M. Longspur	Savannah Sp.	Vesper Sp.	W. Meadowlark
0:12-0:15	S	Brewer's Sp.	Cowbird	H. Lark	Killdeer	L. Bunting	L.B. Curlew	M. Longspur	Savannah Sp.	Vesper Sp.	W. Meadowlark

Figure 1: Illustration of a DDO datasheet for recording avian survey data. Full names of species are provided at the top and then shortened within-survey for convenience. The "Observer" column is marked when an observation is made by the secondary observer.

Data analysis - Following the completion of surveys, observations recorded by observers were compared to truth. Because each vocalization was a 4-second audio clip while the data were collected at 3-second intervals, I used the midpoint of the vocalization to determine which 3-second interval was considered the 'true' 3-second interval in which the vocalization occurred. A detection was correct when it was recorded either in the 'true' 3-second interval or ± one 3-second interval. For example, if a vocalization was recorded in the 1:12-1:15 time interval when it was truly played in the 1:15-1:18 time interval, this was considered correct. A false positive was an incorrectly-identified vocalization or an observation reported outside of the allowed time interval (more than one time interval away from truth).

False positive rates by survey method and by species were calculated as:

$$\hat{p}_{\mathrm{FP}} = N_{FP} / (N_{FP} + N_{TP})$$

, where $\hat{p}_{\mathrm{FP}}$ is the false positive rate, N_{FP} is the number of false positive observations, and N_{TP} is the number of true positive observations. Therefore, $\hat{p}_{\mathrm{FP}}$ is a proportion of how many detections were false positive detections. I used analysis of variance (ANOVA) and Tukey's Honestly Significant Differences (Tukey HSD) to determine if there were statistically-significant ($p \leq$

0.05) differences in false positive rates by survey method, observer group, species, and their combinations. All data were analyzed in R (R Core Team).

Results

The total number of unique surveys conducted was 183. The number of observations reported in all surveys was 3163, and the true number of vocalizations played was 3306. Inexperienced observers reported 797 observations in ISO surveys and 545 observations in DDO surveys, while experienced observers reported 862 observations in ISO surveys and 959 observations in DDO surveys. The overall false positive rate was 0.232 (SD = 0.008). False positive rates per observer ranged from 0.000 – 0.788 in ISO surveys and from 0.007 – 0.581 in DDO surveys.

Survey method affected false positive rates in both groups (Fig. 2). False positive rates in DDO surveys were significantly lower than false positive rates in ISO surveys ($p \approx 0$) (Fig. 3). Experienced observers had a false positive rate of 0.095 (SD = 0.001) in ISO surveys and 0.032 (SD = 0.006) in DDO surveys. This 6.3 percentage point decrease in false positive rates is statistically significant ($p \approx 0$). Inexperienced observers had a false positive rate of 0.511 (SD = 0.018) in ISO surveys and 0.391 (SD = 0.021) in DDO surveys. This 12 percentage point decrease in false positive rates is also statistically significant ($p = 0.0017$).

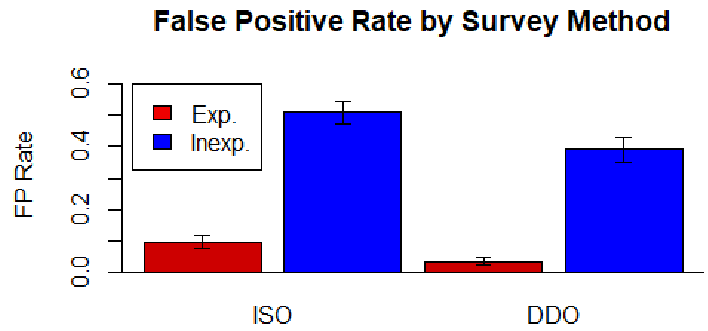

Figure 2: False positive rates by survey method given experience level, calculated as the proportion of detections that were misidentifications. Error bars are 95% confidence intervals.

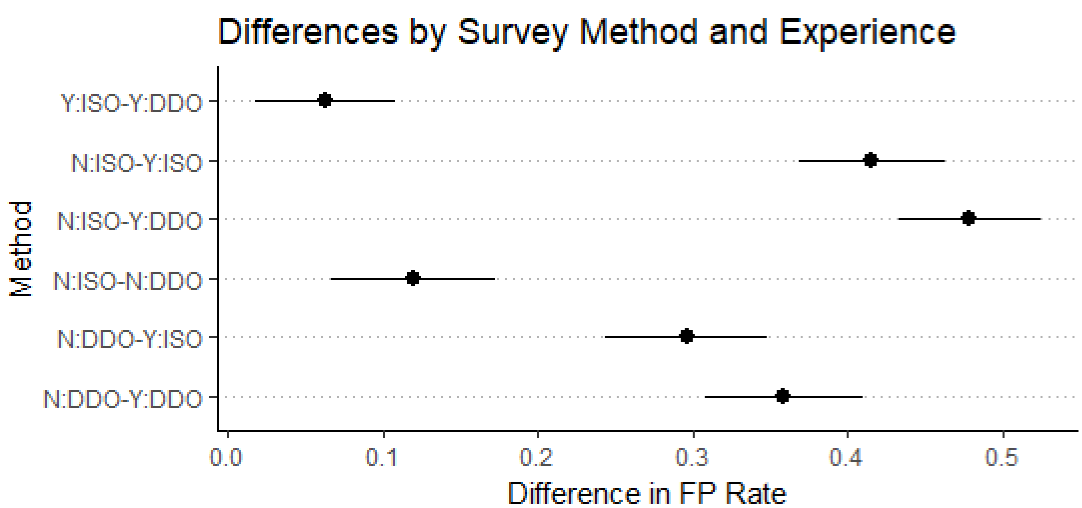

Figure 3: Tukey HSD of false positive rates by survey method and experience level in avian survey data. N is inexperienced observers, while Y is experienced observers; ISO is the independent single-observer method, and DDO is the dependent double-observer method. In a given pair, the former survey method/experience combination reported the higher false positive rate, while the latter reported the lower rate. All differences were significant ($p < 0.05$).

False positive rates were significantly different across the two experience levels ($p \approx 0$). False positive rates by species (Fig. 4) differed between the two experience levels (Fig. 5). Within experienced observers, the false positive rate of Horned Lark (HOLA) was significantly higher than rates of all other study species with a false positive rate of 0.168 (SD = 0.026). The next closest species, Western Meadowlark (WEME), had a false positive rate of 0.085 (SD = 0.021). Within the inexperienced group, false positive rates were high. The Brown-Headed Cowbird (BHCO) had the lowest false positive rate of 0.181 (SD = 0.032) within the inexperienced observers, and Killdeer (KILL) had a false positive rate of 0.314 (SD = 0.034).

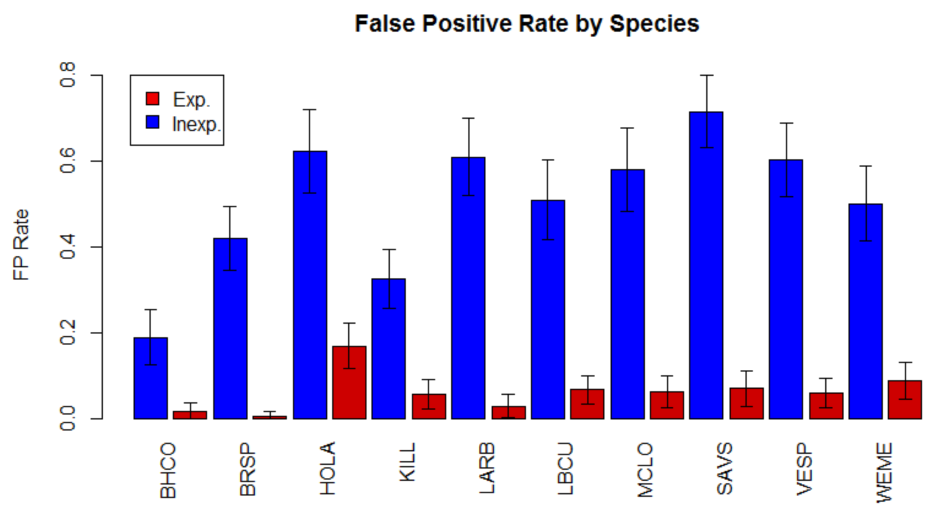

Figure 4: False positive rates by species given experience level, calculated as the proportion of detections that were misidentifications. Species are Brown-Headed Cowbird (BHCO), Brewer's Sparrow (BRSP), Horned Lark (HOLA), Killdeer (KILL), Lark Bunting (LARB), Long-Billed Curlew (LBCU), McCown's Longspur (MCLO), Savannah Sparrow (SAVS), Vesper Sparrow (VESP), and Western Meadowlark (WEME). Error bars are 95% confidence intervals.

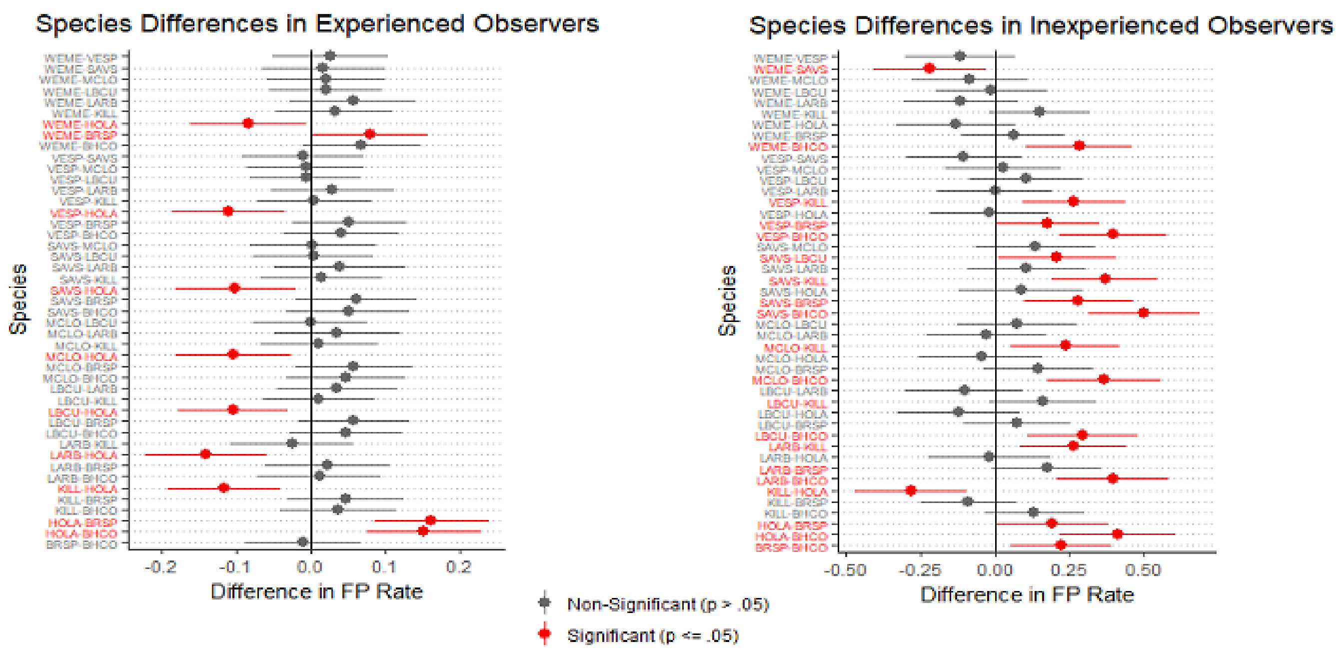

Figure 5. Tukey HSD of differences in false positive rates between species in experienced (left) and inexperienced (right) observers. For each species pair, the data point is the false positive rate of the first species minus the false positive rate of the second species. Points to the left of zero indicate that the second species in the pair had a higher false positive rate, while points of the right of zero indicate that the first species had a higher false positive rate. Species are Brown-Headed Cowbird (BHCO), Brewer's Sparrow (BRSP), Horned Lark (HOLA), Killdeer (KILL), Lark Bunting (LARB), Long-Billed Curlew (LBCU), McCown's Longspur (MCLO), Savannah Sparrow (SAVS), Vesper Sparrow (VESP), and Western Meadowlark (WEME). Bars and text in red are significant differences ($p < .05$). The false positive rate of HOLA was significantly higher than false positive rates of all other species in experienced observer surveys, while BHCO, BRSP, and KILL false positive rates tended to be significantly lower than those of other species in inexperienced observers.

Discussion

In many studies, false positives are assumed not to occur. This assumption is clearly violated given the results of this experiment. False positives occurred in both survey types and with observers of both experience levels. The decline in false positive rates using the DDO method was 6.3 percentage points in experienced observers and 12 percentage points in inexperienced observers.

Some species in this study were commonly mistaken for similar species. Killdeer and Long-Billed Curlew were regularly interchanged in both experience groups. McCown's Longspur was often misidentified as Horned Lark in experienced observer surveys because of the similarity of their vocalizations. Brewer's Sparrow had the lowest false positive rate in the experienced group and the third-lowest false positive rate in the inexperienced group.

This research is significant for studies on wildlife species that use count-based methods. This study uses 10 species, but a field study would be subject to dozens of potential (and not preemptively known) species as well as effects of environment and distance, so false positive rates are likely higher in field studies. The assumption that false positives do not occur biases estimates and renders them inaccurate or misleading. Inaccuracies in detection hamper timely and effective management practices. A species that should be a target for conservation efforts may be overlooked if its abundance is overestimated due to false positives. For avian surveys that are often short-term, accurate counts are important for detecting small changes before it is too late to take necessary action. Studies using ISO methods will estimate populations more accurately by incorporating this study's ISO false positive rate (with adjustments for their specific contexts) into detection rates.

This experiment is limited by a few factors. First, and most obviously, there is no visual component to these surveys. Logistical restraints have prevented the addition of visual components to these surveys. Second, observers were only exposed to a 4-second long stimulus for identifying an individual. In a field setting, some species may vocalize for a longer time period, giving observers more time to identify the individual. Alternatively, a 4-second vocalization may be more stimulus than is typical for species that are quiet or cryptic. Third, applications of this study to non-bird taxa may be limited. Songbird surveys are unusual in their complexity relative to other wildlife surveys where the DDO method has been applied (Barbraud et al. 2005; Shirley et al. 2011; Henry and Anderson 2016). These other studies do not mention misidentification as a potential problem, while in multispecies avian surveys, this is often the primary problem. Thus, this study may be less informative for wildlife surveys in which there is little to no potential for misidentification.

This research on overall and species-specific rates of false positives can be used to guide future research decisions towards methods that are most suitable. Because DDO surveys do have significantly lower overall false positive rates than ISO surveys, then it may be advantageous for multispecies surveys to use the DDO method to most accurately assess occupancy and abundance. Proper training on the implementation of the DDO method would be crucial to ensuring that detection rates are calculated as accurately as possible so that those detection rates can later be incorporated into population models. However, for single-species surveys with a focal species that is easily recognizable to even naïve observers, the cost of employing a second observer to aid in surveys may outweigh biases in detection rates. A false positive rate should be incorporated into detection rates in these studies to more accurately assess detection. No matter the survey method, technicians should be trained in identification of local species prior to the field season to ensure that misidentifications are limited as much as possible.

These data serve as a baseline for further research into false positives. Future research should apply these false positive rates to population estimates to determine the amount of bias incurred by false positives. Because truth would be known, the magnitude of bias for each species could be very accurately determined. Future research should also incorporate visual detections into surveys of known truth to calculate false positive rates with visual detections.

Acknowledgements

I would first like to thank Dr. Victoria Dreitz for giving me the opportunity to work on this project. Vicky, I will never be able to thank you enough for all of the resources, time, guidance, and support that you gave me as I worked through this project. You believed that I could get this research done, and done well, when I sometimes didn't believe it myself. Thank you for both encouraging me to look into more avenues and for curbing my enthusiasm when I got ahead of myself. I hope I was able to make this project into what you envisioned it could be years ago.

Special thanks go to the Avian Science Center and particularly to Jessie Golding, Alan Harrington, Jason Tack, Danielle Fagre, and Kaitlyn Reinstma for all of the preliminary work that they did to get this project off the ground. Without their particular contributions, whether they be vision, methods, or code, I would never have been able to get this research done in the brief amount of time I had. Thank you also to Drs. Chad Bishop and Mike Mitchell for taking time out of your busy schedules to serve on my committee and provide much-needed guidance. Your thoughts and ideas helped me immensely in my data collection and revision processes and truly made this project into something of widespread scientific value.

Finally, I would like to thank my volunteer observers for their time and patience. If I've discovered anything through this project, it's that coordinating lots of volunteers is no easy task. You brave folks dealt with my many last-minute scheduling changes with grace. Without all of you, I would have no data to present. Thank you for your commitment. I hope you found this project to be as fun and valuable of an experience as I did.

Literature Cited

Aldredge, M.W., K. Pacifici, T.R. Simons, and K.H. Pollock. 2008. A novel field evaluation of the effectiveness of distance and independent observer sampling to estimate aural avian detection probabilities. Journal of Applied Ecology 45: 1349 –1356.

Audacity Team (2014). Audacity(R): Free Audio Editor and Recorder [Computer program]. Version 2.0.0 retrieved July, 1st 2015 from http://audacity.sourceforge.net/

Bioacoustics Research Program. (2014). Raven Pro: Interactive Sound Analysis Software (Version 1.4) [Computer software]. Ithaca, NY: The Cornell Lab of Ornithology. Available from http://www.birds.cornell.edu/raven.

Barbraud, C. and G. Gélinaud. 2005. Estimating the sizes of large gull colonies taking into account nest detection probability. Waterbirds 28: 53–60.

Connors, B. M., A. B. Cooper, R. M. Peterman, and N. K. Dulvy. 2014. The false classification of extinction risk in noisy environments. Proceedings of the Royal Society B 281: 20132935.

Diefenbach, D.R., D. W. Brauning, and J. A. Mattice. 2003. Variability in grassland bird counts related to observer differences and species detection rates. The Auk, 120, 1168–1179.

Fitzpatrick, M.C., E.L. Preisser, A.M. Ellison, and J.S. Elkinton. 2009. Observer bias and the detection of low-density populations. Ecological Applications 19: 1673–1679.

Forcey, E.M., J.T. Anderson, F.K. Ammer, and R.C. Whitmore. 2006. Comparison of two double-observer point-count approaches for estimating breeding bird abundance. Journal of Wildlife Management 70: 1674–1681.

Golding, J. D. 2015. Assessing changes in avian communities. Master's Thesis. University of Montana: Missoula, Montana.

Golding, J. D. and V. J. Dreitz. 2016. Comparison of removal-based methods for estimating abundance of prairie songbirds. Journal of Field Ornithology. In Review.

Golding JD, Nowak JJ, Dreitz VJ. A multispecies dependent double-observer model: A new method for estimating multispecies abundance Ecol Evol. 2017;7:3425–3435.

Kissling, M.L. and E.O. Garton. 2006. Estimating detection probability and density from point-count surveys: a combination of distance and double –observer sampling. The Auk 123: 735–752.

Miller, D.A., J.D. Nichols, B.T. McClintock, E.H. Campbell 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: 1422–1428.

Miller, D.A., L.A. Weir, B.T. McClintock, E.H. Campbell Grant, L.L. Bailey, and T.R. Simons. 2012. Experimental investigation of false positive errors in auditory species occurrence surveys. Ecological Applications 22: 1665–1674.

Miller, D.A., L.A. Bailey, E.H. Campbell Grant, B.T. McClintock, L.A. Weir, and T.R. Simons. 2015. Performance of species occurrence estimators when basic assumptions are not met: a test using field data where true occupancy status is known. Methods in Ecology and Evolution 6: 557–565.

Nichols, J. D., J. E. Hines, J. R. Sauer, F. W. Fallon, J. E. Fallon, and P. J. Heglund. 2000. A double-observer approach for estimating detection probability and abundance from point counts. The Auk 117: 393–408.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

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

Royle, J.A. 2004. N-mixture models for estimating population size from spatially replicated counts. Biometrics 60: 108–115.

Shirley, M. H., R. M. Dorazio, E. Abassery, A. A. Elhady, M. S. Mekki, H. H. Asran. 2011. A sampling design and model for estimating abundance of Nile crocodiles while accounting for heterogeneity of detectability of multiple observers. Journal of Wildlife Management 76: 966–975.

Tipton, H.C., P.F. Doherty Jr., and V.J. Dreitz. 2009. Abundance and density of mountain plover (Charadruis montanus) and burrowing owl (Athene cunicularia) in eastern Colorado. The Auk 126: 493–499.

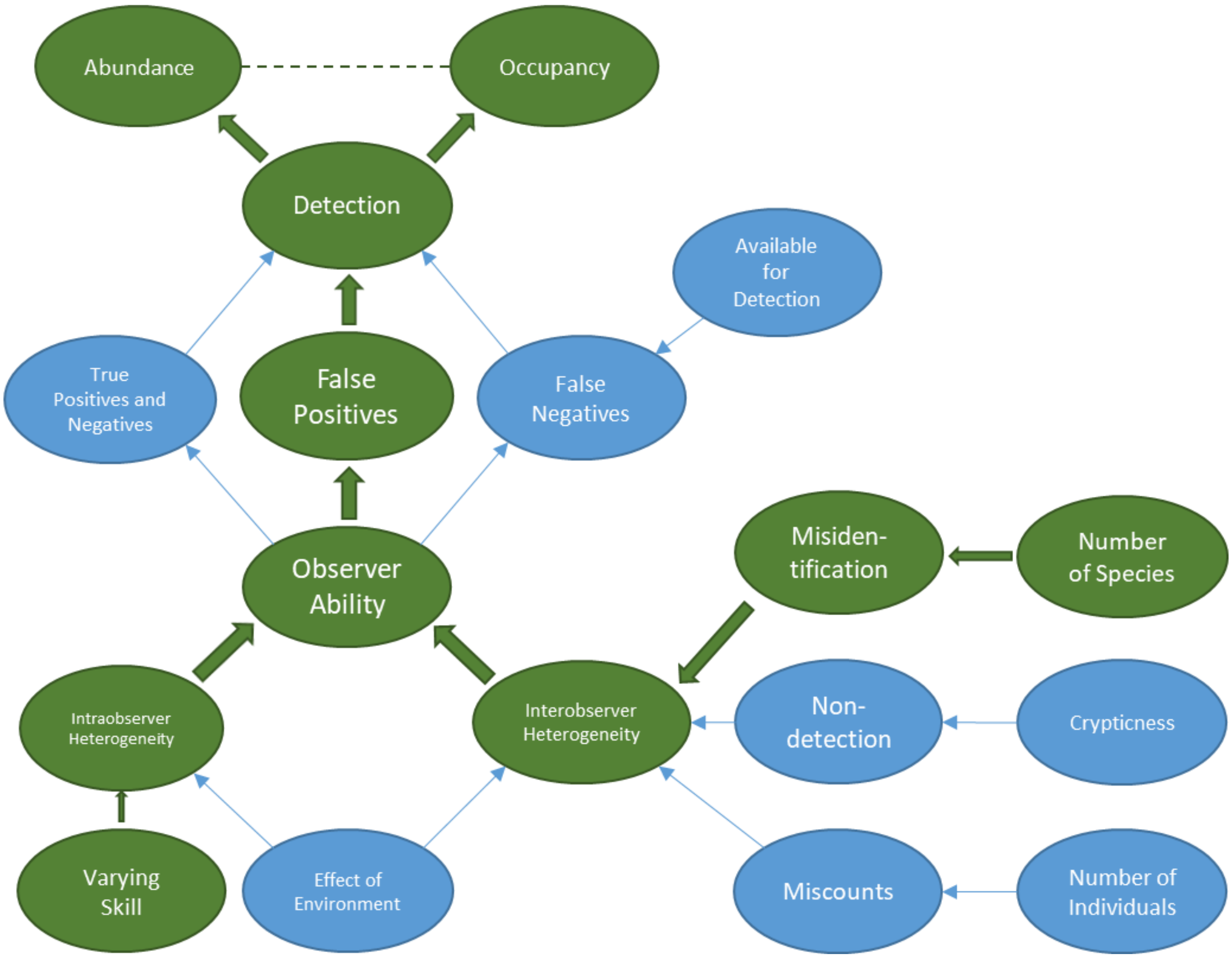

Figure 6: Conceptual diagram of factors affecting population estimates. Bubbles in green are the factors pertinent to this study.