2023

Determining species-specific false-positive rates using visual and auditory cues: a case study with sagebrush steppe songbirds

Amelia K. Evavold
University of Montana, Missoula, amelia.evavold34@gmail.com

Determining species-specific false-positive rates using visual and auditory cues: a case study with sagebrush steppe songbirds

Amelia Evavold, Wildlife Biology Program, Avian Science Center, W.A. Franke College of Forestry and Conservation, University of Montana, 32 Campus Drive, Missoula MT 59812, USA.
Email: amelia.evavold34@gmail.com

Abstract

Errors in wildlife field data threaten to bias resulting abundance and occupancy estimates if not properly accounted for or minimized. Methods to account for false-positive errors in wildlife data have not been as thoroughly developed as those for false-negative errors despite false-positives being present across diverse wildlife taxa and study systems. The calibration method to account for false-positives involves assessing the field detection method to determine how often false-positive errors occur in the field data. Rates can then be incorporated into estimations based on the field data to improve estimation accuracy. This study presents an application of the calibration approach for multispecies avian abundance surveys of seven songbird species (Brewer's Sparrow, Chestnut-collared Longspur, Horned Lark, Long-billed Curlew, Thick-billed Longspur, Vesper Sparrow, Western Meadowlark) in the sagebrush steppe and grassland ecosystem of eastern Montana. Completion of simulated avian surveys resulted in estimation of species-specific false-positive rates as well as examination of how rates may change with the availability of different identification cues. Visual identification cues (video of birds) were always available but auditory identification cues (bird vocalizations) were not always present. Approximately 15% of focal species identifications were false-positives (SD= 0.36). False-positive rates varied significantly between 15 out of 21 focal species pairs, ranging from rates of 0.003 to 0.402 (SD=0.054, SD=0.49). The availability of bird vocalizations in tandem with bird visuals did not differ significantly from false-positive rates based on visuals only (p=2e-16). These results suggest that among these species and study system false-positive rates are primarily a product of similarities in species morphology rather than vocalizations.

Introduction

Accurate data is critical to produce reasonable wildlife population estimates to inform appropriate management action. However, detection is rarely perfect in survey data and imperfect detection can result in biased occupancy or abundance estimates (Kéry and Schmidt 2008). Two principal types of bias-inducing errors result in imperfect detection: false-negatives and false-positives. A false-negative error occurs when a species or individual is present but undetected. A false-positive error occurs when a species or individual is absent but counted as present. False-positives result from species misidentification or double counting (Royle and Link 2006, Miller et al. 2011, Strickfaden et al. 2019). Since it is often reasonable to assume that detecting every individual within a sample area is nearly impossible, there are various well-developed approaches to account for false-negatives in survey data (Gu and Swihart 2004, Guillera-Arroita 2017).

False-positives, however, have not received the same amount of attention. This is in part due to a combination of the idea that detecting an animal that is not present seems impossible or uncommon, and that it is challenging to verify truth in the field to determine when false-positives

are occurring. However, false-positives are present across diverse target species and ecosystems, including during auditory avian surveys in sagebrush-grassland (Strickfaden et al. 2019), among lynx in the Alps Mountains (Molinari-Jobin 2011), murrelets along the Alaskan coast (Schaefer et al. 2015), wolves in northern Montana (Miller et al. 2013), and cheetahs, hyaenas, and leopards in south-west Kenya (Madsen 2020). Not only are false-positives present in wildlife survey data, but they can also significantly bias population estimates even at relatively low rates (Royle and Link 2006).

While often present in wildlife count-based data, false-positive rates can be extremely variable depending on study design, observer experience, observer expectations, field conditions, and similarity between species. Several studies have shown that observer experience is an important factor influencing false-positive rates, and results consistently demonstrate that as observer experience increases, false-positive rates decrease (Farmer et al. 2012, Schaefer et al. 2015, Strickfaden et al. 2019). Further, Farmer et al. (2012) found that observer experience interacted with species rarity. Experienced observers misidentified common species as rare, and less experienced observers misidentified rare species as common (Farmer et al. 2012). Environmental factors such as field conditions can also impact false-positive rates. For example, in a study conducted on murrelets where observers conducted surveys from boats, false-positive rates were higher during surveys when the sea state was choppy rather than calm (Schaefer et al. 2015). Because false-positives are primarily a result of misidentification, similarities amongst species morphologies or vocalizations can greatly impact false-positive rates, especially in multispecies surveys (Royle and Link 2006, Strickfaden et al. 2019). These findings demonstrate that false-positive frequency and distribution within datasets can change significantly with subtle shifts in who is collecting the data, field conditions, and methods. It is therefore important to streamline approaches to assess false-positives across a range of study designs and systems.

We can address false-positives both by limiting false-positive occurrence with intentional survey design or statistically accounting for them after data collection. There are several methods proposed to address false-positives. Chambert et al. (2015) outlined three primary approaches to statistically account for false-positives in count-based data: site-confirmation, observation-confirmation, and calibration. Site-confirmation and observation confirmation generally require that truth is confirmed at roughly the same location and time that the primary field data is collected. To do this, survey methods are compared to another method that is assumed to reflect truth such as images, recordings, or analyzed fur/scat samples. If this additional truthful dataset is not available, then application of these approaches to account for false-positives is limited. Rather than determining site or observation-level truth, the calibration approach involves assessing the accuracy of the field survey method by employing the same method in a situation where truth is known so that detection error rates can be calculated and factored into the field data and subsequent models (Chambert et al. 2015). Approaches for assessing survey methods can be tailored to different study designs, species, and habitats. Survey methods can also be assessed at any point before, during, or after primary data collection which creates opportunities to decrease false-positives by adjusting methods pre data collection and/or accounting for the expected false-positive rates post data collection.

Surveys of avian species serve as a prime case study for false-positive calibration since they are often multispecies surveys, quick-moving, cryptic, and abundant. For example, Strickfaden et al. (2019) created simulated avian surveys by randomizing bird song recordings to gauge identification accuracy and false-positive rates in multispecies avian auditory abundance surveys. However, many avian count surveys employ both auditory and visual cues to identify

species. Incorporating visual and auditory cues in tandem introduces the potential for both higher false-positive rates for some species, if there are morphological similarities, and decrease false-positive rates among other species, if vocally similar species have very distinct appearances. The addition of visual cues to the survey method likely impacts detection rates and identification accuracy; therefore, a new assessment of the survey method is required to determine false-positive rates and subsequently apply the calibration method.

This study contributes to the limited available examples of calibration method implementation to account for false-positives in wildlife count-based studies, and further attempts to simulate the variable conditions present in field settings. Specifically, I determine species-specific false-positive rates from multispecies avian abundance surveys employing a combination of purely visual observations and visual observations in tandem with aural observations. Simulated surveys were created with video and corresponding audio of seven songbird species collected in their native sagebrush steppe and grassland habitats. Volunteer identifications from these surveys were used to calculate false-positive rates. Since my study involves multiple species, several with similar morphology and vocalizations, false-positives are likely to occur. I expected that false-positive rates would vary significantly by species. I also predicted that false-positive rates would be lower when there are two identification cues available to the observer (visual and vocalization of a bird) versus one identification cue (visual of a bird) since the addition of a vocalization provides more opportunity to differentiate species. However, species morphology was expected to be the dominant factor affecting when false-positives occur rather than vocalizations since visual cues are consistently present for birds in the surveys. Overall, by further developing this calibration framework and streamlining methods for determining survey method- and species-specific false-positive rates, existing and future wildlife count-based data can better account for false-positive errors, improving management actions and conservation agendas based on this count-based data.

Methods

Study System – Video footage was collected in the field from June 11, 2020 to June 15, 2020 on lands managed by the Bureau of Land Management west of Roundup, Montana. The area is on the eastern edge of sagebrush steppe and grassland ecosystems in the western US. Vegetation is composed of sparse, short shrubs of Wyoming big sagebrush (*Artemisia tridentata* spp. *wyomingensis*) and silver sagebrush (*Artemisia cana*) intermixed with grasses, primarily needle-and-thread grass (*Hesperostipa comata*) and western wheatgrass (*Pascopyrum smithii*). Sufficient video footage was available for seven avian species in the surveys. The focal species include a sagebrush obligate species, Brewer's Sparrow (*Spizella breweri*); four grassland obligate species, Chestnut-collared Longspur (*Calcarius ornatus*), Horned Lark (*Eremophila alpestris*), Long-billed Curlew (*Numenius americanus*), Thick-billed Longspur (*Rhynchophanes mccownii*), and two sagebrush and grassland associated species Vesper Sparrow (*Pooecetes gramineus*), and Western Meadowlark (*Sturnella neglecta;* Baker et al. 1976, Cochran & Anderson 1987, Dubois 1935, Johnson et al. 2019, Miller et al. 2017).

Simulated survey creation – Original video footage was first stabilized to reduce shakiness and then cut into eight-second clips (hereafter referred to as clips). The clips contain either one visible bird or empty habitat with zero visible birds. Clips did not contain multiple birds to avoid confounding factors that could arise with multiple species or individuals. In both clips with a bird and empty habitat, audio could include other birds vocalizing that were not visible. Audio associated with each clip is the original audio collected at the time of video

footage collection and is unedited. Stabilization and clipping were completed in Adobe® Premiere® Pro software. Multiple observers with experience surveying these specific avian species reviewed each clip to confirm species identifications in each bird clip. In addition, images from these clips were uploaded to iNaturalist, where identifications were further verified by iNaturalist members (iNaturalist).

Clips were separated by species and whether the visible bird was singing or not. Hereafter "nonvocal" will refer to clips that show a bird that is not singing, and "vocal" will refer to clips that show a bird which can be heard singing. The number of unique clips for each species was unequal. To set the number of clips equal for each species a set of both random nonvocal bird clips and vocal bird clips from each species were duplicated so that each species had 168 total clips (126 of nonvocal individuals, 42 of vocal individuals; Figure 1). This allowed for an equal likelihood that each species appeared when developing the surveys, a 75% likelihood that a bird was nonvocal, and a 25% likelihood that a bird was vocal. This same random duplication was conducted on empty habitat clips so that 50% of all clips available for random selection when creating surveys are empty habitat clips (Table 1).

Once random duplication was complete the final clip set contained 2,352 clips, made up of 1,176 focal species clips and 1,176 empty habitat clips. Forty-five clips were randomly selected from this final clip set to create a six-minute survey. Each list of clips in a survey includes the species present and serves as a "truth list" to determine correct and incorrect observer identifications. Clip duplication and random clip selection for survey creation was completed using the software KNIME (Berthold et al. 2007). In total, 118 unique six-minute surveys were created. Each survey had a different proportion of bird clips to habitat clips, and different proportions of each focal species. However, across all surveys the set parameters resulted in approximately half of all survey clips containing birds, and approximately equal occurrences of each species.

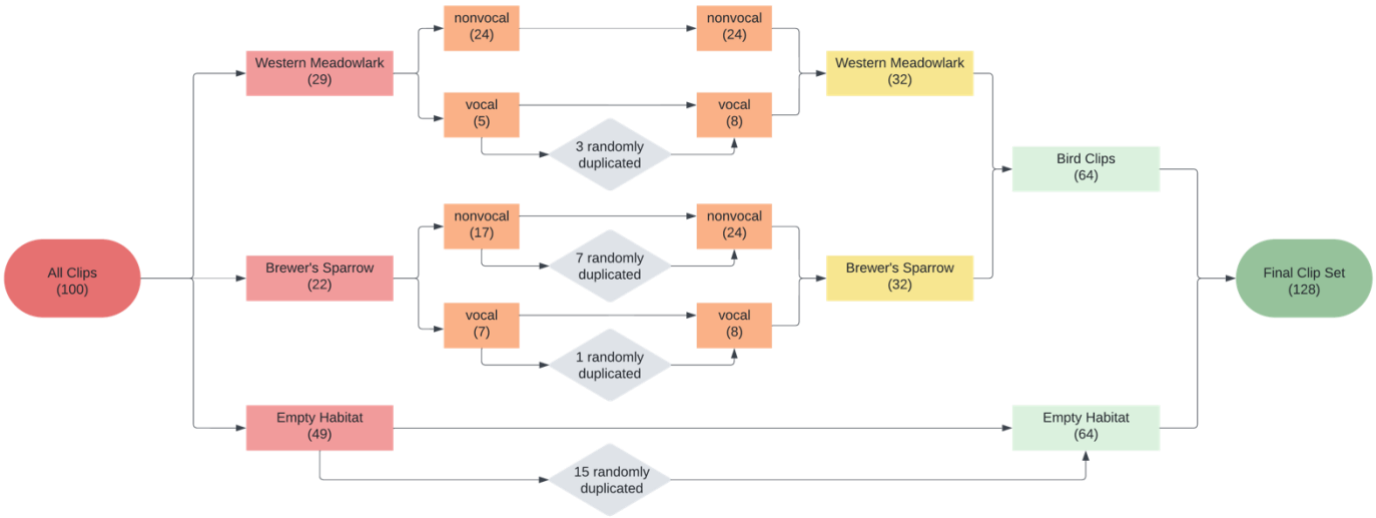

Figure 1. Example workflow of randomized clip duplication to ensure equal species likelihood and average empty habitat clip frequency. Shown here for two species and empty habitat with example clip numbers (not real numbers). Random clips are duplicated from each group until all species have equal clip numbers, and equal proportions of vocal and nonvocal clips (25% and 75% of bird clips respectively). Empty habitat clips are duplicated until they equal total bird clips, so that approximately 50% of all survey clips are empty habitat. The "final clip set" is what is sampled from to create the survey sequences.

In addition to the six-minute surveys, a shorter "practice survey" was created for volunteers to become familiar with the survey format before completing a full survey. The practice survey was two-minutes long and contained fifteen clips. It was created using the same process and parameters for the full surveys outlined above. An overview of the different video terms and their respective processing is included below (Table 1).

Table 1. Video terms, and respective length and properties. For reference when discussing video processing and survey creation parameters.

Video Terminology	Length	Processing and Properties
Clip	8 sec	- stabilized - video and audio - finalized material for survey creation 3 types: - bird clips: one bird that remains in frame, exhibiting one behavior, may hear nonvisible birds vocalizing - nonvocal: visible bird is not singing - vocal: visible bird is seen and heard singing - empty habitat clips: no birds visible, can still potentially hear birds vocalizing
Survey	6 min	- partially randomized sequences of 45 clips - each survey is unique Parameters: - 50% likelihood of an empty habitat clip - 50% likelihood of a bird clip - 75% likelihood that a bird clip is nonvocal - 25% likelihood that a bird clip is vocal - equal likelihood that each species appears in survey
Practice Survey	2 min	- same processsing and parameters outlined for Surveys above, but only 15 clips - only one created, same practice survey used for all volunteers

Administering surveys – Nine undergraduate student volunteers from the University of Montana completed the simulated surveys. Before data collection, volunteers were provided with basic information about the study area where the video footage was collected, including all bird species observed in the study area and an identification guide for the twenty most common bird species (which included the seven focal species). This was to emulate the resources they would have been provided before surveying in a field setting. Observers were not otherwise informed of the specific species present in the surveys. As with survey design, survey administration was structured as close as possible to field surveys in order to produce false-positive rates applicable to field data. Volunteers were assigned a unique identification number to avoid associating volunteer names with the data. Self-assessments have been suggested to represent ability more accurately than years of experience (Miller et al. 2012). Therefore, volunteers were asked to self-classify themselves as naïve, beginner, proficient, or expert skill level for bird identification in sagebrush-grassland ecosystem. Reported skill levels ranged from beginner to proficient, with most volunteers self-identifying as beginners.

Surveys were administered in a quiet room free of distractions. Volunteers sat facing a desktop monitor and attached speaker. Surveys were played on the monitor with consistent volume levels across all surveys. Volunteers were instructed to identify visible birds to the species-level, with the option to use vocalizations to aid in identifications (e.g., if a visible bird is

heard vocalizing). Bird identifications were stated aloud so I could record the identifications next to the appropriate survey clip. Volunteers completed a practice survey at the first data collection session to become familiar with the survey format and instructions. After completing the practice, volunteers estimated the percentage of identifications they felt were correct (as another method to explore volunteer ability) and then began completing full six-minute surveys. Visual or written identification aids were not allowed during the surveys. There was a two-minute break between each survey. Volunteers completed no more than three surveys within one data collection session to avoid identification errors introduced by observer fatigue (Norton-Griffiths 1976). A similar limit is often enforced during field surveys to limit errors from technician fatigue. Volunteers completed between ten and twenty data collection sessions each over several weeks.

Data analysis – Volunteer observations were compared to the true species identifications in the surveys. A detection refers to an instance when an observer identified that a bird was present, regardless of correct or incorrect species identification. The data also included detections in which an individual stated a 'bird' was present but were not confident in identifying to the species-level. A false-positive is when a species is identified as present when it is not truly present (either due to misidentification, or identification when there is no bird present). Detection and false-positive rates were calculated with the counts of detections and false-positive occurrences from volunteer identifications. Overall detection probability, p_D, was calculated using the following equation:

$$p_D = B_D / B_T ,$$

where B_D is the number of birds detected by observers, and B_T is the true number of birds that occurred in the surveys. False-positive rates, p_{FP}, were calculated as the following:

$$p_{FP} = B_{FP} / (B_{FP} + B_{TP}),$$

where B_{FP} is the number of false-positives recorded, and B_{TP} is the number of true positives recorded. Both total and species-specific false-positive rates were determined using this equation. Detections that did not include a specific species were not included in false-positive calculations. Two-sample *t*-tests were used to compare false-positive rates of vocal and nonvocal clips for a) the combination of all seven focal species and b) individually within each of the seven focal species. Analysis of variance (ANOVA) and Tukey's Honestly Significant Difference (Tukey's HSD) were used to determine significant differences in false-positive rates between a) focal species and b) nonvocal versus vocal bird clips for each focal species. Focal species false-positives due to misidentifications were further explored graphically to identify patterns. For instance, to determine if a specific species frequently results in false-positives for another species or if a pattern changed if a species was vocalizing. All statistical analysis was conducted using the program R (R Core Team 2023).

Results

Nine volunteers completed 118 unique six-minute surveys for a total of 708 survey minutes. There were 2,589 clips of birds included in the surveys (between 353 and 386 per species) and volunteers detected 2,590 birds for an overall detection rate of 1.0004. There

were 21 false-negatives when present birds were not detected and 22 false-positives when birds were identified during empty habitat clips resulting in the detection rate being greater than one.

Out of 2,512 identifications made to the species level, there were 597 false-positives resulting from both misidentifications and detection of birds in empty habitat. This resulted in an overall false-positive rate of 0.238 (SD=0.43). This rate includes false-positives from when non-focal species were identified from clips of focal species. These overall false-positive rates were also significantly lower when birds were vocal in clips as opposed to nonvocal (p=0.0223). When visible birds were nonvocal, false-positives occurred at a rate of 0.248 (SD=0.43). When visible birds were vocal, the false-positive rate decreased to 0.205 (SD=0.40).

Once results were filtered to remove identifications of non-focal species, 338 false-positives occurred of the 2,253 total focal species identifications made. This resulted in a cumulative focal species false-positive rate of 0.150 (SD= 0.36, Table 2). Among only identifications of focal species, there was no significant difference between false-positive rates for clips of nonvocal birds (0.157, SD=0.34) versus vocal birds (0.129, SD=0.36, p=0.0957).

Table 2. False-positive rates and standard deviations (SD) of each focal species when birds are nonvocal versus vocal. These numbers are only for detections when focal species were identified (disregarding false-positives for non-focal species). Totals include false-positives resulting from both misidentifications and identifications from empty habitat clips. Nonvocal and vocal numbers include only false-positives from misidentifications. Species are Western Meadowlark (WEME), Vesper Sparrow (VESP), Thick-billed Longspur (TBLO), Long-billed Curlew (LBCU), Horned Lark (HOLA), Chestnut-collared Longspur (CCLO), and Brewer's Sparrow (BRSP).

Species		Detections	False-positives	False-positive rate	SD
BRSP	total:	332	117	0.352	0.48
	nonvocal:	272	92	0.338	0.47
	vocal:	52	17	0.327	0.47
CCLO	total:	289	23	0.080	0.27
	nonvocal:	225	16	0.071	0.26
	vocal:	64	7	0.109	0.31
HOLA	total:	329	15	0.046	0.21
	nonvocal:	236	8	0.034	0.18
	vocal:	88	2	0.023	0.15
LBCU	total:	347	1	0.003	0.05
	nonvocal:	253	1	0.004	0.06
	vocal:	94	0	0	0
TBLO	total:	342	72	0.211	0.41
	nonvocal:	248	49	0.198	0.40
	vocal:	93	22	0.237	0.43
VESP	total:	266	107	0.402	0.49
	nonvocal:	205	84	0.410	0.49
	vocal:	61	23	0.377	0.49
WEME	total:	348	3	0.009	0.09
	nonvocal:	241	1	0.004	0.06
	vocal:	106	1	0.009	0.10
Total	total:	2253	338	0.150	0.36
	nonvocal:	1680	251	0.157	0.34
	vocal:	558	72	0.129	0.36

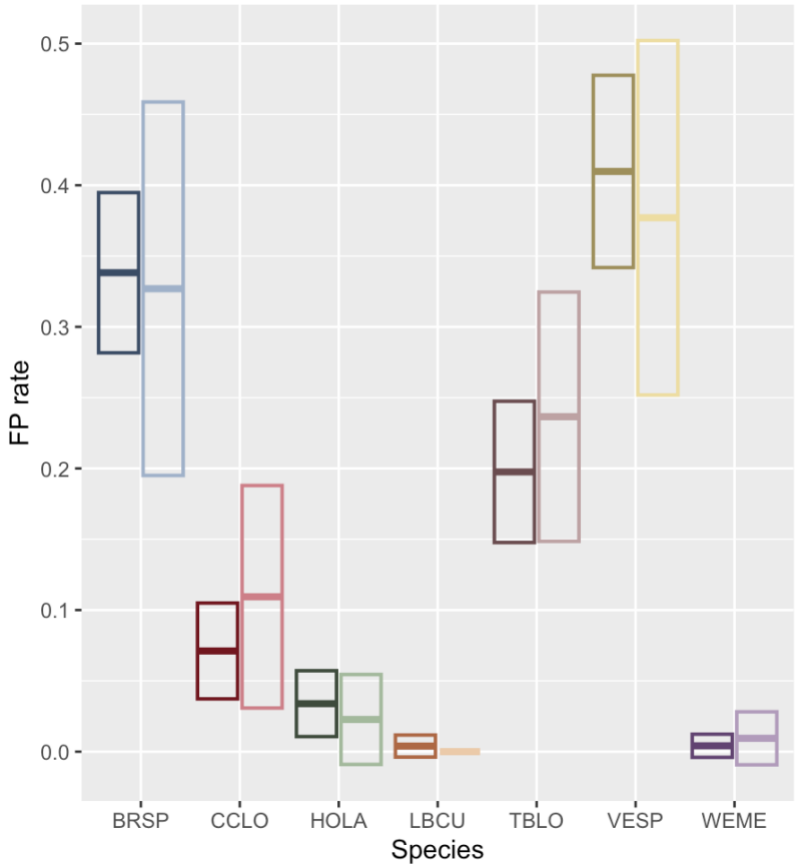

Figure 2. Focal species false-positive rates when birds were nonvocal (left bar in each pair, darker colors) versus when vocal (right bar in each pair, lighter colors). Central crossbars are false-positive rates. False-positives from empty habitat clips are excluded. Bars are 95% confidence intervals. Species are Western Meadowlark (WEME), Vesper Sparrow (VESP), Thick-billed Longspur (TBLO), Long-billed Curlew (LBCU), Horned Lark (HOLA), Chestnut-collared Longspur (CCLO), and Brewer's Sparrow (BRSP).

Focal species false-positive rates varied considerably between species but did not differ significantly between nonvocal and vocal bird clips within each species (Figure 2). Species-specific false-positive rates varied significantly between 15 of the 21 possible species pairs, demonstrating false-positive occurrence is not distributed evenly among species in this study system (Figure 3). Vesper Sparrow and Long-billed Curlew false-positive rates were the most different (p=5.12e-11), where Western Meadowlark and Long-billed Curlew were the most similar (p=1.000). Species false-positive rates ranged from a low of 0.003 (SD=0.054) for Long-billed Curlew to a high of 0.402 (SD=0.49) for Vesper Sparrow. Four focal species had slightly lower false-positive rates when detected birds were vocal, while three had slightly higher false-positive rates when detected birds were vocal. Within each focal species there were no significant differences between false-positive rates when nonvocal versus vocal (Figure 4).

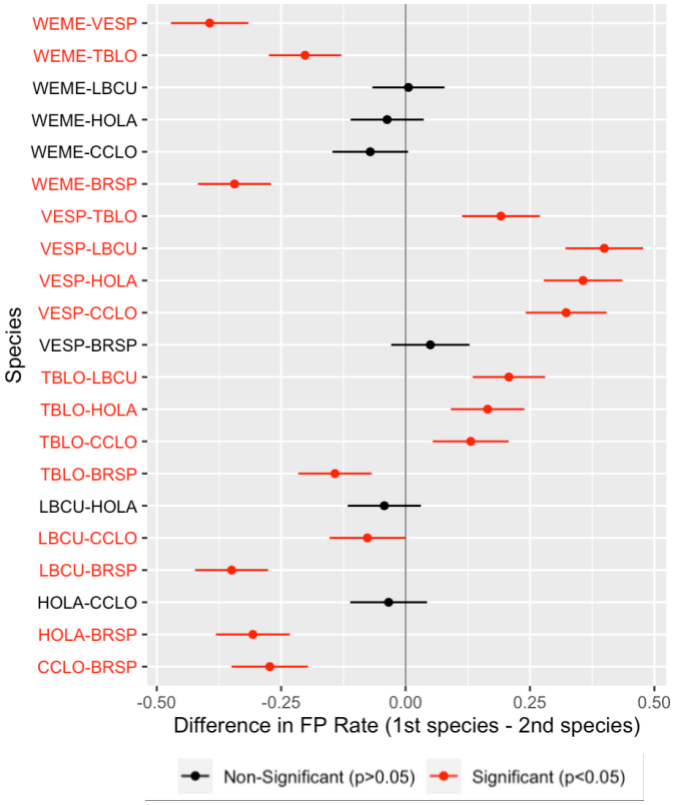

Figure 3. Tukey HSD of differences in focal species false-positive (FP) rates. Points depict the FP rate of the first species minus FP rate of the second species. Points to the left of zero signify higher FP rates for the second species, and points to the right of zero signify higher FP rates for the first species. Bars are 95% confidence intervals. Species are Western Meadowlark (WEME), Vesper Sparrow (VESP), Thick-billed Longspur (TBLO), Long-billed Curlew (LBCU), Horned Lark (HOLA), Chestnut-collared Longspur (CCLO), and Brewer's Sparrow (BRSP).

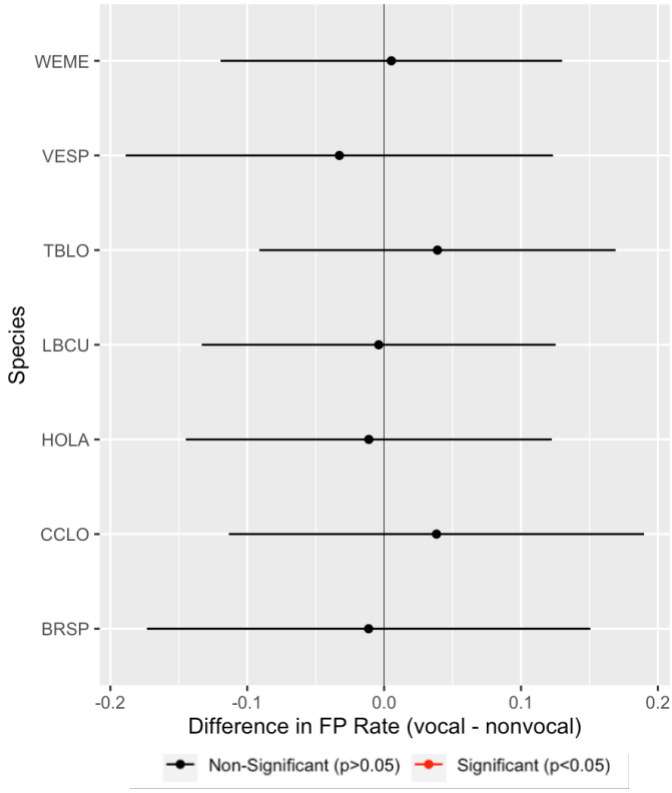

Figure 4. Tukey HSD of differences in focal species false-positive (FP) rates between nonvocal and vocal birds. Points depict the FP rate of the species when vocal minus the FP rate of the species when nonvocal. Points to the left of zero signify higher FP rates when birds are nonvocal, and points to the right of zero signify higher FP rates when birds are vocal. Plotted results are filtered to include only within-species comparisons. Bars are 95% confidence intervals. Species are Western Meadowlark (WEME), Vesper Sparrow (VESP), Thick-billed Longspur (TBLO), Long-billed Curlew (LBCU), Horned Lark (HOLA), Chestnut-collared Longspur (CCLO), and Brewer's Sparrow (BRSP).

Several patterns appear in how false-positives occur for each focal species (Figure 5). Distribution of false-positives for each focal species appear to be reciprocated. For example, Brewer's Sparrow false-positives occur primarily from Vesper Sparrows, and Vesper Sparrow false-positives occur primarily from Brewer's Sparrows. Distribution of false-positives across the true species present remain relatively consistent between nonvocal and vocal clips. However, for all focal species except Horned Lark, false-positives were spread across more true species when nonvocal and across fewer species when vocal (e.g., Thick-billed longspur nonvocal false-positives occurred from five different truly present species being misidentified, and Thick-billed longspur vocal false-positives occurred from three truly present species being misidentified). It is important to note however that vocal bird clip sample sizes were considerably smaller than nonvocal bird clip sample sizes.

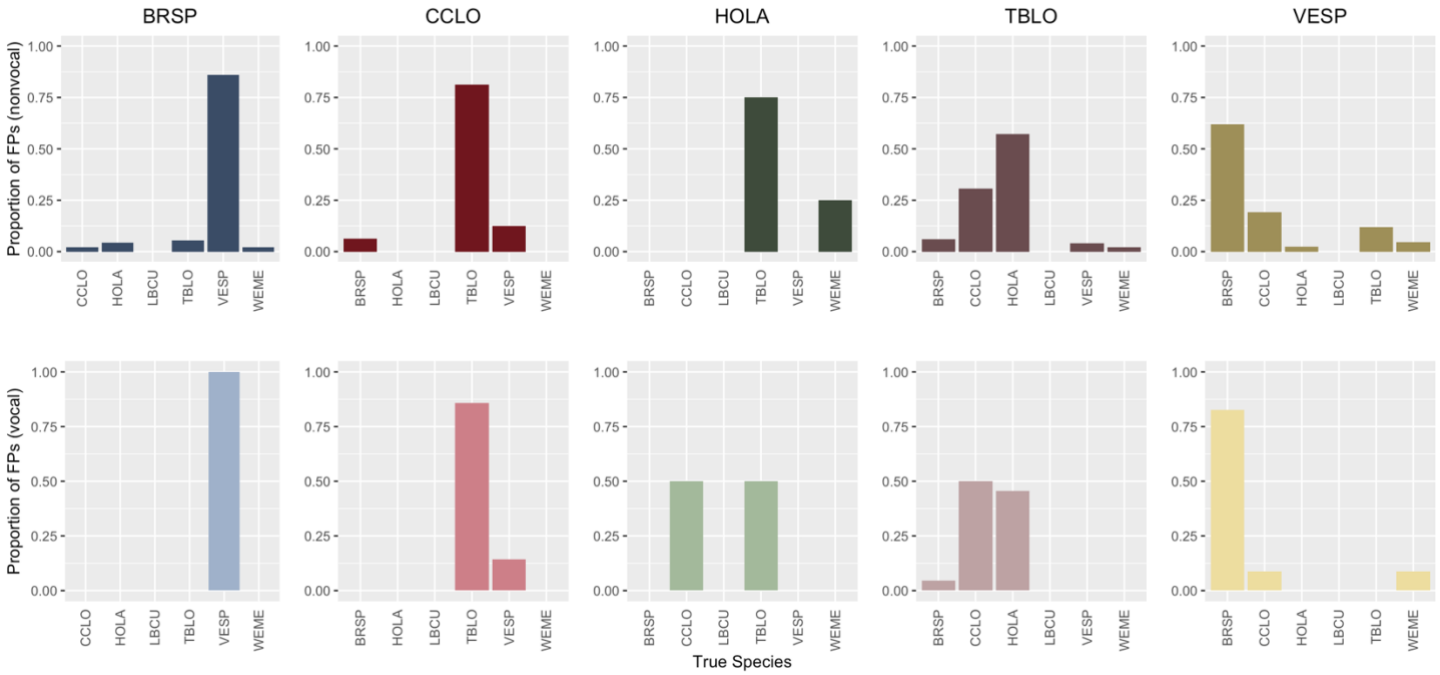

Figure 5. Focal species false-positive distributions across the true species present. Bars show the proportion of the focal species' false-positives that occurred when a clip of the true species was nonvocal (top row, darker colors) or vocal (bottom row, lighter colors). Species are Western Meadowlark (WEME), Vesper Sparrow (VESP), Thick-billed Longspur (TBLO), Long-billed Curlew (LBCU), Horned Lark (HOLA), Chestnut-collared Longspur (CCLO), and Brewer's Sparrow (BRSP). Plots for LBCU and WEME are not presented here since their false-positive occurrences were very low (1 and 3 total false-positives respectively).

Discussion

Species-specific false-positive rates were extremely variable across the seven focal species, ranging across two orders of magnitude. This is despite being observed in the same habitat types, by the same observers, and at similar overall frequencies. Results suggest that this level of variation in species false-positive rates is primarily due to morphological similarities rather than similarities in vocalizations since there was no significant change in false-positive errors for any focal species with the addition of the auditory cue of birdsong. Focal species morphologies and feather patterning compared across identified and true species reveals similar markings and forms between more frequently mistaken species (Figure 6). For example, Thick-billed Longspur were most often falsely identified from clips of Horned Lark and Chestnut-collared Longspur. These three species are similarly sized, and all have distinct black throat/chest bands as well as dark facial markings. Conversely, species less frequently mistaken as Thick-billed Longspur had notably different forms with less contrasting coloration (Brewers and Vesper Sparrow), or a conspicuous yellow chest (Western Meadowlark).

In this situation visual cues appear to dominate, and auditory cues may therefore be disregarded in favor of the "dominant" visual cue. It was expected that having another identification cue (visual + auditory) would aid in identification accuracy and therefore decrease false-positive rates compared to instances with one identification cue (visual). This is not the case. Instead, false-positive rates and patterns in which species were misidentified remained relatively stable between the single-cue and double-cue groups. Simulated surveys were designed

so that observers were only asked to identify visible birds but could use song to aid in identification if the visible bird was vocalizing. Visuals were therefore the consistent cue and within the hierarchy of human senses, sight lies at the top. When visual images are clear, vision dominates sound, and it is only when visual stimuli become unclear that sound may dominate human perception (Alais & Burr 2004). There was also only one bird visible at a time whereas audio could include multiple vocalizing birds (all of which could be non-visible birds). The need to sift through extraneous auditory information and stimuli, particularly when time-limited and as relatively inexperienced observers, likely also contributed to the auditory cue's lack of impact on false-positive rates.

It is important to note however that while vocalizations did not impact false-positive rates for focal species, when calculated for all identifications (including identification of non-focal species from focal species clips), false-positive rates did decline significantly when birds were vocal versus only nonvocal. Vocal bird clip false-positive rates were 4.3% lower than non-vocal bird clips. This may suggest that the addition of more species in surveys could increase importance and value of auditory cues, however this would need to be explored further to draw any definitive conclusions.

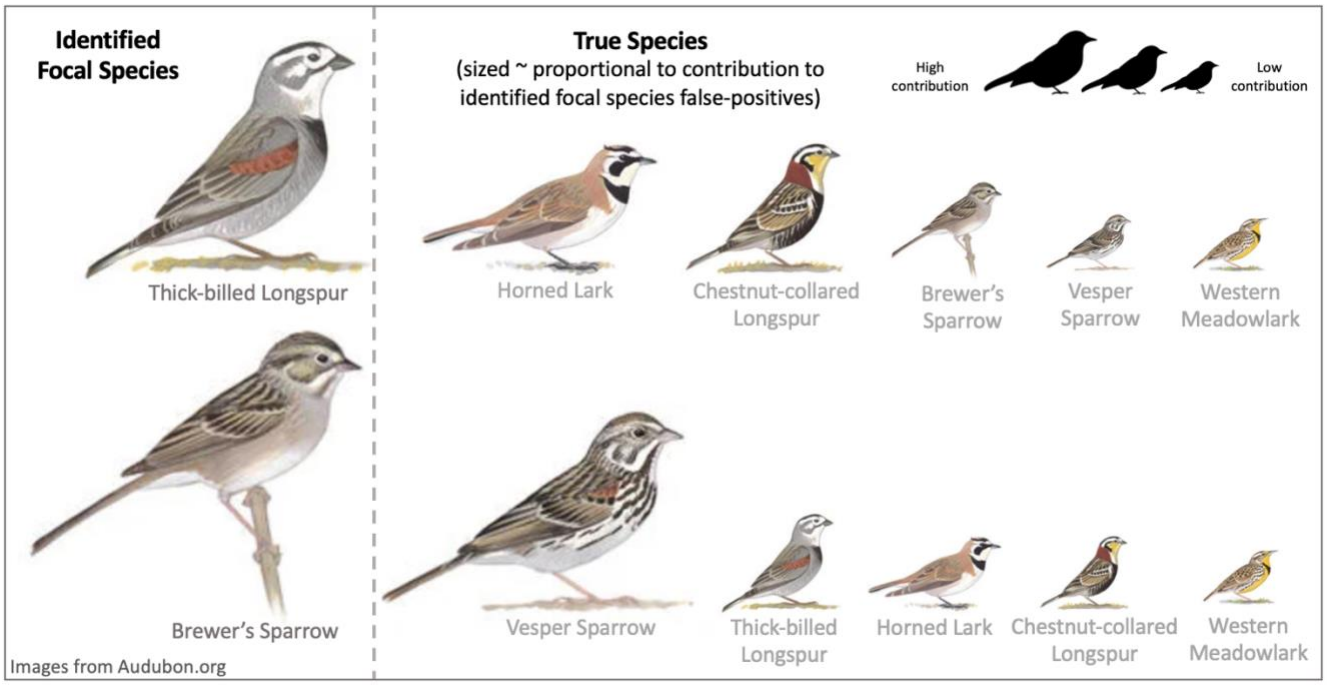

Figure 6. Proportional representation of Brewer's Sparrow and Thick-billed Longspur false-positive (FP) distributions across true species. (e.g., larger true species images represent more FPs for the focal species, smaller true species images represent fewer FPs for the focal species) Species are Brewer's Sparrow (BRSP), Chestnut-collared Longspur (CCLO), Horned Lark (HOLA), Thick-billed Longspur (TBLO), Vesper Sparrow (VESP), and Western Meadowlark (WEME).

Overall, the false-positive rate estimates presented here are expected to be conservative estimates for the represented observer experience level. The full surveying process and range of variability in field conditions could not be reasonably reflected in the simulated avian surveys created. There is ample opportunity for improved estimates by accounting for more complexity such as a larger set of species, multiple birds visible at once, bird movement into and out of the frame, panoramic survey viewing, and the many potential weather and lighting conditions. Each

of these variables has the potential to increase false-positive rates further, increase false-positive rate variability, and even introduce new sources of false-positive errors such as double counting. Following the calibration approach workflow these songbird false-positive rate estimates, however conservative, can be incorporated into existing abundance data of these focal species in the sagebrush-steppe ecosystem to improve resulting abundance estimates.

While the idea of even higher false-positive rates is disconcerting when considered in regard to past wildlife studies that have dismissed false-positive occurrence, there are many promising avenues for improving false-positive exploration methods. Rapidly advancing technologies such as virtual reality and artificial intelligence (AI) indicate that it may soon be possible to better simulate field variability and complexity while still establishing truth to account for false-positives. For example, one major limiting factor in this study was the challenge of collecting enough high-quality video footage of each species. The surveyed species therefore had to be limited to seven. With AI tools, rather than needing to collect footage for surveys, the visual material could be collected from existing media and manipulated to fit desired simulation formats. This would increase ease of survey creation and create more flexibility in modifying surveys to fit field conditions. More accurate simulations mean more accurate false-positive estimates. Just as this proposed study builds off an existing study design, it is expected that in the coming years it will become easier for other studies to tailor this calibration approach to fit their respective focal species, study environments, and detection methods to account for false-positive rates in their unique count-based data.

Understanding false-positive rates and the variables that impact them can improve wildlife abundance estimates by incorporating estimated error rates post hoc, but the knowledge can also target the very sources of false-positives during data collection. For example, determining which species are most frequently misidentified as other species (and what identification cues lead to these patterns) can inform where to focus efforts in training protocols for field technicians. Similarly, detection method comparisons can inform which detection method should be implemented to achieve lower false-positive rates. The simulated survey approach in particular creates exciting opportunities to use the surveys as training tools prior to field seasons to improve and gauge observer accuracy. As simulated surveys improve in realism they can serve as immersive virtual realities where observers can experience data collection prior to entering the field and get real-time feedback and tailored training focused on areas where simulations show their accuracy can improve.

Overall, both the results of this study and the methods employed are of interest. It is another example of the complexities within false-positive error occurrence, demonstrated through an application of the calibration approach using simulated surveys to assess false-positive error rates in multispecies songbird abundance surveys. False-positives occurred at high rates within the simulated sagebrush-steppe and grassland ecosystem but were not uniform across study species and did not vary based on available identification cues. The need for further development in the field of wildlife false-positive research is well-established and further confirmed by this study. The subject boasts many opportunities for exciting interdisciplinary collaboration and application of cutting-edge technologies. By improving understandings of false-positives this form of error can both be reduced at the source during data collection as well as incorporated into wildlife abundance estimates to increase accuracy. Improved wildlife abundance estimates mean that resulting wildlife policies and management actions can be better fit to the true status and needs of wildlife populations.

Acknowledgements

I would like to thank the University of Montana undergraduate students who volunteered to participate in this study, dedicating time and energy to support my data collection. I would also like to acknowledge the instrumental role of the Avian Science Center (ASC) at the University of Montana. Dr. Victoria Dreitz, director of the ASC, served as my research advisor and contributed a tremendous amount of guidance, encouragement, and expertise throughout the research process. Members of the ASC, particularly Kaitlyn Reintsma, Christian Dupree, and Miles Scheuring contributed extensively to my project formulation, species verification, survey design, analysis, writing, and more. I also thank the two additional members of my thesis committee, Dr. Zachary Cheviron and Dr. Michael Musick for their support and feedback. Finally, I would like to express my gratitude for the financial support to conduct this research received from the Irene Evers Competitive Undergraduate Research Scholarship and the Bill Gabriel Avian Science Center Scholarship. This research complies with the requirements of the Institutional Review Board at the University of Montana (IRB #149-21).

Literature Cited

Alais, D., & Burr, D. (2004). The Ventriloquist Effect Results from Near-Optimal Bimodal Integration. *Current Biology, 14*, 257-262.

Baker, M. F., Eng, R. L., Gashwiler, J. S., Schroeder, M. H., Braun, C. E. (1976). Conservation Committee Report on Effects of Alteration of Sagebrush Communities on Associated Avifauna. *The Wilson Bulletin, 88*(1), 165-171.

Berthold, M.R., Cebron, N., Dill, F., Gabriel, T.R., Kötter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K., Wiswedel, B. (2007). KNIME: the Konstanz information miner. Studies in Classification, Data Analysis, and Knowledge Organization (GfKl 2007). *Springer*.

Chambert, T., Miller, D. A., & Nichols, J. D. (2015). Modeling false positive detections in species occurrence data under different study designs. *Ecology, 96*(2), 332–339.

Cochran, J. F. & Anderson, S. H. (1987). Comparison of habitat attributes at sites of stable and declining long-billed curlew populations. *The Great Basin Naturalist, 47*(3), 459-466.

Dubois, D. (1935). Nests of Horned Larks and Longspurs on a Montana Prairie. *The Condor 32*(2), 56-72.

Farmer, R. G., Leonard, M. L., & Horn, A. G. (2012). Observer effects and avian-call-count survey quality: Rare-species biases and overconfidence. *Auk, 129*(1), 76–86.

Guillera-Arroita, G., Lahoz-Monfort, J. J., van Rooyen, A. R., Weeks, A. R., & Tingley, R. (2017). Dealing with false-positive and false-negative errors about species occurrence at multiple levels. *Methods in Ecology and Evolution, 8*(9), 1081–1091.

Gu, W., & Swihart, R. K. (2004). Absent or undetected? Effects of non-detection of species occurrence on wildlife-habitat models. *Biological Conservation, 116*(2), 195–203.

iNaturalist. Available from https://www.inaturalist.org. Accessed 5 January 2023.

Johnson, A. E. M., Sillett, T. S., Luther, D., Herrmann, V., Akre, T. A., McShea, W. J. (2019). Effect of Grassland Management on Overwintering Bird Communities. *The Journal of Wildlife Management, 83*(7), 1515-1526.

Kéry, M., & Schmidt, B. R. (2008). Imperfect detection and its consequences for monitoring for conservation. *Community Ecology, 9*(2), 207–216.

Madsen, E. K., & Broekhuis, F. (2020). Determining multi-species site use outside the protected areas of the Maasai Mara, Kenya, using false positive site-occupancy modelling. *Oryx*, *54*(3), 395–404.

Miller, D. A. W., Nichols, J. D., Gude, J. A., Rich, L. N., Podruzny, K. M., Hines, J. E., Mitchel, M. S. (2013). Determining Occurrence Dynamics when False Positives Occur: Estimating the Range Dynamics of Wolves from Public Survey Data. *PLoS ONE, 8*(6), e65808.

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

Miller, D. A. W., Weir, L. A., Mcclintock, B. T., Grant, E. H. C., Bailey, L. L., Simons, T. R., Miller, D. A. W., Weir, L. A., Mcclintock, B. T., Campbell, E. H., Bailey, L. L., & Simons, T. R. (2012). *Ecological Applications*,. *22*(5), 1665–1674.

Miller, R. A., Bond, L., Migas, P. N., Carlisle, J. D., Kaltenecker, G. S. (2017). Contrasting habitat associations of sagebrush-steppe songbirds in the intermountain west. *Western Birds, 48*, 35-55.

Molinari-Jobin, A., Kéry, M., Marboutin, E., Molinari, P., Koren, I., Fuxjäger, C., Breitenmoser-Würsten, C., Wölfl, S., Fasel, M., Kos, I., Wölfl, M., & Breitenmoser, U. (2011). Monitoring in the presence of species misidentification: The case of the Eurasian lynx in the Alps. *Animal Conservation*, *15*(3), 266–273.

Norton-Griffiths, M. (1976). Further Aspects of Bias in Aerial Census of Large Mammals. *The Journal of Wildlife Management, 40*(2), 368–371.

R Core Team (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Royle, J. A., & Link, W. A. (2006). Generalized Site Occupancy Models Allowing for False Positive and False Negative Errors. *Ecology*, *87*(4), 835–841.

Schaefer, A., Lukacs, P. M., & Kissling, M. L. (2015). Testing factors influencing identification rates of similar species during abundance surveys. *Condor, 117*(3), 460–472.

Strickfaden, K. M., Fagre, D. A., Golding, J. D., Harrington, A. H., Reintsma, K. M., Tack, J. D., & Dreitz, V. J. (2019). Dependent double-observer method reduces false-positive errors in auditory avian survey data. *Ecological Applications*, *30*(2), 1–9.