

2015

monitoR: Automation Tools For Landscape-scale Acoustic Monitoring

Jonathan Edward Katz
University of Vermont

Follow this and additional works at: <https://scholarworks.uvm.edu/graddis>



Part of the [Acoustics, Dynamics, and Controls Commons](#), and the [Natural Resources and Conservation Commons](#)

Recommended Citation

Katz, Jonathan Edward, "monitoR: Automation Tools For Landscape-scale Acoustic Monitoring" (2015). *Graduate College Dissertations and Theses*. 359.
<https://scholarworks.uvm.edu/graddis/359>

This Dissertation is brought to you for free and open access by the Dissertations and Theses at ScholarWorks @ UVM. It has been accepted for inclusion in Graduate College Dissertations and Theses by an authorized administrator of ScholarWorks @ UVM. For more information, please contact donna.omalley@uvm.edu.

MONITOR: AUTOMATION TOOLS FOR LANDSCAPE-SCALE ACOUSTIC MONITORING

A Dissertation Presented

by

Jonathan Katz

to

The Faculty of the Graduate College

of

The University of Vermont

In Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
Specializing in Natural Resources

May, 2015

Defense Date: August 29, 2014
Dissertation Examination Committee:

Therese Donovan, Ph.D., Advisor

Ruth Mickey, Ph.D., Chairperson

Brian Mitchell, Ph.D.

Allan Strong, Ph.D.

Cynthia J. Forehand, Ph.D., Dean of the Graduate College

ABSTRACT

Climate change coupled with land-use change will likely alter habitats and affect state parameters of the animal populations that dwell in them. Affected parameters are anticipated to include site occupancy and abundance, population range, and phenophase cycles (e.g., arrival dates on breeding grounds for migrant bird species). Detecting these changes will require monitoring many sites for many years, a process that is well suited for an automated system. We developed and tested *monitoR*, an R package that is designed for long-term, multi-taxa automated passive acoustic monitoring programs. *monitoR* correctly identified presence for black-throated green warbler and ovenbird in 64% and 72% of the 52 surveys using binary point matching, respectively, and 73% and 72% of the 52 surveys using spectrogram cross-correlation, respectively. Of individual black-throated green warbler song events, 73% of 166 black-throated green warbler songs and 69% of 502 ovenbird songs were identified by binary point matching. Spectrogram cross correlation identified 64% of 166 black-throated green warbler songs and 64% of 502 ovenbird songs. False positive rates were <1% for song event detection.

We describe a method to identify the probability of survey presence in a template-based automated detection system using known false positive rates for each template. True and false positive detection rates were observed in 146 training surveys. These probabilities were used in a Bayesian approach that discriminates between detections in occupied surveys and unoccupied surveys. We evaluated this approach in 146 test surveys. A total of 1142 Black-throated green warbler (*Setophaga virens*) songs were observed in the training surveys and test surveys, which we attempted to locate with 3 different binary point matching templates. When only posterior probabilities greater than 0.5 were considered detections, the average ratio of accurate identifications of survey presence to false positive identifications in 500 bootstrapped samples improved from 1.2:1 using a standard score cutoff approach to 2.8:1 using all 3 templates and a likelihood-based discriminator. With the selected score cutoffs the average true positive and false positive rates for the combined three templates were 0.18 and 0.002, respectively.

Automated detection methods are increasingly being used for identification and monitoring of landscape-scale responses to climate change and land-use change. Skepticism of automated acoustic monitoring software is largely due to higher false positive and negative error rates than those in traditional human surveys, but the false positive multiple method occupancy model is capable of estimating detection parameters and occupancy state when one method has occasional false positive detections. We test the accuracy of the model when automated detection of black-throated green warbler is mixed with human detection in 4 recorded surveys at 60 sites. Precision and accuracy are evaluated by simulation, and we use the results to optimize future sampling. In simulation, parameter estimates by the multiple method occupancy model are close to those we computed manually when two surveys are manually analyzed. Our results support the use of the multiple method false positive occupancy model to track detection rates in automated monitoring programs.

ACKNOWLEDGEMENTS

This work was supported by USDA National Needs Graduate Fellowship grant no. 2008-38420-19547, the Vermont Cooperative Fish and Wildlife Research Unit, the National Park Service cooperative agreement P10AC00288, and the National Phenology Network.

A big thank you to all who supported me as I sorted my thoughts, and another big thank you to those who stimulated new thoughts along the way. This includes, in no particular order, my advisor Terri Donovan - an amazing editor, ecologist, and idea factory; Allan Strong - ace ornithologist, enthusiastic birder, and critical reader; Brian Mitchell - monitoring expert, critical reader, and bioacoustic pioneer; Ruth Mickey - statistical expert and critical reader; Sasha Hafner - R wizard, informal mentor, and formal collaborator; Cori Brauer - lab mate and highly tolerant field collaborator; additional lab mates Michelle Brown, Jonathan Cummings, and Kurt Rinehart - all sources of original insights; my family Gary, Vicki, Sara, and Elliott - as caring and generous a family as a person can have; and the most thanks to my wife Meghan Gilbert - patient, rational, wise, and patient.

TABLE OF CONTENTS

Acknowledgements	ii
List of Figures	vii
1 Comprehensive Literature Review	1
Environmental Health: A Bird’s Eye View	2
Acoustic Monitoring Fields of Study	6
Automating Bioacoustics	10
Correlation In Automatic Detection Data	13
Estimating Occupancy From Automated Detection Data	21
Developing Validation Theory For Automated Acoustic Monitoring	26
References	27
2 Tools for automated acoustic monitoring within the R package <code>monitor</code>	41
Summary	42
Introduction	43
Package goals	44
Overview of Functions and Terminology	46
Acoustic Detection Workflow With Local Data Storage	48
Acoustic Detection Workflow	
With Database Storage	49
Managing Inputs	49
Surveys	49
Templates	51
Managing Outputs	52
Manipulating Templates	53
Manual Spectrogram Annotation	54
Manual Results Verification	55
The Acoustics Database	55
Toward An Automated Monitoring Framework	56
Acknowledgements	57
References	58
Table and Figure Legend	61
3 Assessment of Error Rates in Acoustic Monitoring With The R Package <code>monitor</code>	68
Abstract	70
Introduction	71
Methods	74
Database Development (Objective 1)	74
Field Methods (Objective 2)	75
Creating Gold and Silver Standards (Objective 3)	77

Templates (Objective 4)	78
Hits and Classification Rates of Hits (Objective 5)	80
Results	84
Database population (Objective 2)	84
Gold and Silver Standard (Objective 3)	84
Detections and Classification Rates of Detections (Objective 5)	85
Discussion	89
Acknowledgements	94
References	95
Table Legend	100
Figure Legend	100
4 Estimating Presence With Correlated Automatic Detection Results	108
Abstract	110
Introduction	111
Methods	115
Objective 1: Field recordings and manual identification of song events . . .	115
Objective 2: Automatic song detection with monitoR	116
Objective 3: Estimating detection rates in training surveys	118
Objective 4: Estimate posterior probability of presence using objective priors.	120
Objective 5. Compare inference about survey presence between the Bayesian method and a standard cutoff approach.	121
Results	122
Objective 1: Field recordings and manual identification of song events . . .	122
Objective 2: Automatic Song detection with monitoR	122
Objective 3: Estimating monitoR's detection rates in training surveys . . .	123
Objective 4: Objective priors to estimate the posterior probability of presence	124
Objective 5. Compare inference about survey presence between the Bayesian method and a standard cutoff approach	125
Discussion	127
Acknowledgements	132
References	133
Figure Legend	138
5 Occupancy modeling with automated acoustic detections	145
Abstract	147
Introduction	148
Methods	153
Field recordings	153
Study species	154
Manual detection	155
Automatic song detection	156

Objective 1. Estimating single-season multiple method parameters and parameter confidence	158
Objective 2. Comparison between automated and manual detection rates. .	159
Objective 3. Optimizing automated acoustic monitoring for the target species	161
Results	162
Objective 1. Estimating single-season multiple method parameters	162
Objective 2. Comparison of detection rates	163
Objective 3. Estimating monitoring power	164
Discussion	164
Acknowledgements	168
References	169
Figure Legend	173
A Appendix A. Binary point matching and spectrogram cross-correlation	179
Binary Point Matching	180
The template	180
Scoring	181
In monitoR	182
Spectrogram Cross-Correlation	184
The template	184
Scoring	185
In monitoR	186
Bibliography	189

LIST OF FIGURES

2.1	Workflow with inputs and results stored locally.	64
2.2	Workflow with input metadata and results stored in the supplied MySQL “acoustics” database.	65
2.3	A binary point template (a) and correlation template (b), plotted using the default <code>plot</code> method for <code>templateList</code> objects.	66
2.4	Three binary point matching detections of black-throated green warbler (tall rectangles) and two detections of ovenbird in a 30 second spectrogram of a recorded survey.	67
3.1	Black-throated green warbler song.	104
3.2	Spectrogram scores by both template matching methods for 30 seconds of a composite survey.	105
3.3	Sensitivity and specificity curves for binary point matching at varying score cutoffs.	106
3.4	Joint probability of black-throated green warbler and ovenbird detection.	107
4.1	Black-throated green warbler templates (blue and orange) used for presence assessment, as plotted over the grayscale spectrograms.	140
4.2	Detection results for a 10 minute survey from MABI02 on June 6, 2010.	141
4.3	Receiver Operating Characteristic (a) and total number of true hits (b) and false hits (c) for black-throated green warbler detection by all three templates in training surveys.	142
4.4	Distribution of posteriors estimated with an objective prior of 0.5 for three templates.	143
4.5	Distribution of TP:FP surveys in test group after 500 resampling iterations.	144
5.1	Binary point templates for black-throated green warbler detection.	174
5.2	Spectrogram of a survey with four detected black-throated green warbler songs (top, in boxes) and automated detection scores (bottom, solid black line) and “peaks” (points); dashed line indicates score cutoff, above which peaks become “hits”.	175
5.3	Count of negative and positive surveys per site as detected by monitoR (a) and human detection (b); $n = 60$	176
5.4	Automated detection parameter estimates using the multiple method occupancy model by Miller et al. (2011), 2 of 4 surveys manually evaluated.	177
5.5	Power of manual and automated detection by number of surveys.	178
A.1	A binary point matching template (a) and a short survey spectrogram (b). Note that the survey spectrogram must have at least as many time bins as the template, and in most cases will have many more.	183

A.2 A spectrogram cross-correlation template (a) and the same short survey spectrogram as in Figure 1 (b). Again, the survey spectrogram must have at least as many time bins as the template, and in most cases will have many more. 187

CHAPTER 1

COMPREHENSIVE LITERATURE REVIEW

ENVIRONMENTAL HEALTH: A BIRD'S EYE VIEW

The shared bird-human history is overshadowed by gloomy anecdotes of exploitation and extinction, but birds are a unique category of wildlife because they coexist with humans in nearly all environments. It is possible to see birds in urban green spaces, tropical rainforests, the open ocean, the frozen poles, and even the dry deserts. As the easiest wildlife to observe, it is logical that environmental issues will be first observed in bird populations. In this sense they could be considered sentinels of poor habitat quality, as suggested in Rachel Carson's iconic book "Silent Spring" (Carson 1987). Individual species have been explicitly used as sentinels of poor habitat quality; for example, during the industrial revolution canaries were caged and carried into mine tunnels to warn miners of carbon monoxide poisoning, and were immortalized in the phrase "canary in the coal mine". Despite legislation to protect populations of migratory species, songbird populations throughout the world have been in decline for several decades (Amar et al. 2006, Sanderson et al. 2006, Heldbjerg and Fox 2008). The loss of habitat in wintering grounds (Rappole and McDonald 1994) and the various effects of climate change are predicted to negatively affect populations across the full breeding range of many species. An observed effect of climate change is difficulty timing migration and breeding to coincide with sufficient food supplies to provision hatchlings (Butler 2003, Barbraud and Weimerskirch 2006, Both et al. 2006, Forero-Medina et al. 2011). Habitat loss due to climate change is anticipated to be especially difficult for high-elevation breeders—species who typically have little habitat to lose before population numbers suffer (Rodenhouse et al. 2008, Harris et al. 2014).

Inventory (assessing population size or another system state) is a necessary tool for wildlife managers. Repeated inventory of wildlife transitions to monitoring, and monitoring provides the presence data from which broader species occupancy can be extrapolated (MacKenzie

et al. 2002), and it may also serve to fulfill other more broad management objectives (Yoccoz et al. 2001, Legg and Nagy 2006, Nichols and Williams 2006). Targeted monitoring is typically used to assess the results of past management activities, or it can provide data to predictively model the effects of planned activities (Gibbs et al. 1999, Johnson et al. 2006, Runge et al. 2009, Martin et al. 2011*a*). Monitoring without an explicit end point is known as surveillance monitoring. There are circumstances for which surveillance monitoring is warranted, such as when a response to a stimulus is anticipated but it is unknown what the response will be and which species will be affected. In these cases, monitoring goals are quantified by the power of the monitoring itself, e.g. the goal may be to be able to detect changes with a specified statistical power. Many of the anticipated effects of climate change meet the criteria to warrant surveillance monitoring (Wintle et al. 2010), as responses are expected to be unevenly distributed among species. Both targeted and surveillance monitoring programs are difficult for wildlife managers to implement: they often consume many personell hours, and when projects are complete it is not always easy to draw conclusions from the results or to extrapolate local findings to a system-wide pattern. For example, Donovan and Flather (2002) found that local inventory did a poor job of predicting range-wide population trends. The high variation observed at the local scale suggests that the early effects of climate change on bird populations will be most accurately and most clearly visible to landscape scale monitoring programs. To implement landscape-scale monitoring, some recommend that a few species be singled out for occupancy state evaluation to improve efficiency (Noon et al. 2012).

The limitations of local monitoring present a different difficulty for wildlife managers. Ideally monitoring would be capable of identifying local trends and evaluating them in the context of larger landscape level patterns, as it is usually impractical to directly monitor at the landscape scale. With the exception of large volunteer projects such as the Breeding

ENVIRONMENTAL HEALTH: A BIRD'S EYE VIEW

Bird Survey and the Christmas Bird Count, many monitoring projects are limited by the number of people on the inventory crew and the number of sites each crew member can visit in a summer. Despite the high mobility and perceived low error rate of human inventory, there is increasing recognition that the non-zero error rate has a negative effect on overall data quality when humans collect data for species occurrence. Furthermore, data quality may vary not just between observers but also between surveys by the same observer (Kessler and Scott 1981, Bart and Schoultz 1984, Sauer et al. 1994, Kendall et al. 1996, Farmer et al. 2012). Observer bias is time consuming to measure, and as a result it is discounted in the majority of single-observer studies. In a single-observer study, observer bias leaves the study unable to support the assumption of population closure between visits, as unlike a double-observer study it cannot be determined whether presence was overlooked in a survey or whether occupancy state changed between surveys. Furthermore, the effect size of the bias in the estimate of occupancy increases as detection probability decreases (Moore et al. 2004). In extreme cases, observer bias may do more than violate analytical assumptions; at least one study of point count results suggested that a trend for increasing observer age, and therefore reduced detection capability, could explain a small part of trends of decreasing population estimates (Farmer et al. 2014).

Landscape or range-wide responses to climate change may include more than migration timing shifts. Other population responses may be growth or attrition, changes in foraging/hunting or mating behavior, and changes in range extent. These system states would also be most informatively monitored at the landscape level, but to allocate time to do so is often not possible, even for species with highly restricted ranges. It is unclear that knowing to what extent these system states are changing would allow us to reduce the threat to vulnerable species, but it is certain that we are leaving the fate of vulnerable species to chance if no management plan is established that includes repeated cycles with decision,

ENVIRONMENTAL HEALTH: A BIRD'S EYE VIEW

action, and monitoring phases.

An alternative monitoring strategy to minimizing observer bias in bird counts, frog counts, bat counts, and whale counts is to deploy weatherproof recording hardware to record surveys as digital audio files. In bird counts it has been suggested that most birds are heard but not seen; it is likely that most visual detections are also redundantly detected through song, and those that are not amount to fewer than 3% of observations (Brewster and Simons 2009). Recording audio of the survey is a viable option, and it allows the presence of reclusive species to be evaluated by extensive and repeated review, plus surveys can be archived for future reanalysis (Parker 1991, Haselmayer and Quinn 2000, Hobson et al. 2002, Acevedo and Villanueva-Rivera 2006). In contrast to human surveys, when recorded surveys are analyzed automatically a second set of advantages are realized: a) rapid data availability and b) elimination of observer-based errors. Both of these advantages are specific to automated detection methods, and both have positive consequences in wildlife management decision making. A potential disadvantage is diminished survey radius, as the average human ear can hear quieter sounds - such as distant bird songs - than the average all-weather microphone and recorder (Rempel et al. 2013). Manual analysis of recorded surveys may involve measuring song examples from a species to make inference about the physiology or song evolution, and a handful of studies have also manually sorted song events from a survey to locate an individual in physical space. Automated analysis of recorded surveys may attempt to reduce the survey to a general metric of overall ecosystem health, or it may attempt to identify individuals or species in a survey.

Among the other strengths of recording surveys is the ability to record animal behavior and presence in the absence of a human observer. Wildlife study methods weigh the risk of altering the health or behavior of the subject against the potential benefits of new informa-

ACOUSTIC MONITORING FIELDS OF STUDY

tion. A best case outcome from invasive study methodology would be results biased by the sampling technique, while a worst case outcome would be harm or instability introduced to the population. Passive sampling techniques are therefore preferred when possible, provided sufficiently strong inference can be made using results produced by the passive methods. Unlike active acoustic monitoring (playback and response), passive acoustic monitoring uses the sounds present in the ambient environment to make ecological inference. Passive acoustic monitoring has its own bias in that the extent to which a species can be monitored depends on its acoustic contribution. The most vocal taxa include marine mammals, some terrestrial mammals such as coyotes, elephants and bats, many birds, and many insects. Not surprisingly these same taxa provide fodder for the bulk of the literature in passive bioacoustics.

ACOUSTIC MONITORING FIELDS OF STUDY

In its most elementary form passive bioacoustic research requires multiple components. After the target species emits a signal by vocalizing or stridulating, a microphone must receive the signal, a recording device must store the signal, there must be a method for transforming the recorded signal into useable data, and there must be a method of interpreting the data and learning from it. Automated analysis of recorded surveys is an option available with the advent of digital recorders, as the limitations of cassette technology prevented their use as unaccompanied avian survey tools (Johnson et al. 1981).

Spectrograms made from recordings can be used to statistically analyze the structure and variability of bird songs (Martindale 1980), which can in turn be useful for making behav-

ACOUSTIC MONITORING FIELDS OF STUDY

ioral or environmental comparisons. Adapting to life alongside cities (Slabbekoorn and Peet 2003) and noise from remote industrial sites (Habib et al. 2007, Blickley et al. 2012) is a recurring theme in this vein of study; for example, Luther and Baptista (2010) presented recordings that demonstrated how the frequency of white-crowned sparrow songs in Northern California changed over 30 years to remain audible over the sound of traffic and urban industry.

Systematic recording of non-song calls, particularly night flight calls, holds the potential to monitor migration as it happens. Evans and Mellinger (1999) manually located night flight calls in directional recordings. Their initial research successfully identified to species or species groups the invisible overhead migrants and translated the numbers of calls to a minimum number of individuals passing overhead. They also highlighted the advantages of multiple microphones versus just one microphone for identifying trends in space and making more informed inference about the flight patterns during recording, but noted that the overwhelming volume of data readily available in recordings required more advanced computing capabilities.

To record surveys for most measurements based on spectrograms, a single directional microphone and a recorder can provide sufficient data if the recordist is able to move the microphone location as needed, and enough time has passed to gather clear recordings in good weather. Using multiple microphones requires more specialized equipment but could allow more sophisticated analysis of the results. For example, while measurements of individual songs or syllables may be done by hand, it is now common to use specialized software. The R package *seewave* (Sueur et al. 2008) has a toolbox for sound analysis, as does *Avisoft-SASLab Pro* (Avisoft Bioacoustics e.K. 2014) and *Syrinx-PC* (Burt 2006).

ACOUSTIC MONITORING FIELDS OF STUDY

Another application for passive acoustics with digital recording and processing includes using passive acoustics to track animals as they vocalize in a two-dimensional array of microphones and recorders. This is an appealing concept because it does not require a radio transmitter to be attached to each individual and therefore does not introduce capture and handling stress. Three methods to accomplish this measurement are present in the literature, each taking advantage of a different property of sound waves.

Clark (1980) described a device for locating a vocalizing whale in a marine environment using a hydrophone array and a device that analyzed the phase difference between signals arriving at each hydrophone to produce real-time bearings to the vocalizing individual. This method is specific to low wavelength sounds in low dispersion environments, and therefore has not been applied to terrestrial applications.

For locating birds, McGregor et al. (1997) described a method accurate to within $\sim 4\text{m}$ in open meadows and $\sim 7\text{m}$ in woodlands; the method computed location based on shifts in arrival time at each microphone of an array, and it used the software package Canary to do the beam-forming. This study also highlights the effect of environment on acoustics; while marine environments and line-of-sight terrestrial environments make ideal places to establish accurate synchronized acoustic locating arrays, the physical structure within woodlands adds a fair amount of reverberant noise to the signal, which increases uncertainty by “smearing” arrival time at the microphone. Nine years later Mennill et al. (2006) performed a similar test in dense tropical forests and achieved accuracies of $< 3\text{m}$; the improved accuracy over McGregor et al.’s study was likely due to a combination of advances in Global Positioning System locating technologies as well as advances in analysis software: Mennill et al. manually identified the target calls in the software Syrinx PC and used software modules written in Matlab to perform the beam-forming. This method was successfully

ACOUSTIC MONITORING FIELDS OF STUDY

used to track whale movement through the Stellwagen Bank National Marine Sanctuary and in the Boston shipping channel, a project that is also noteworthy for using cellular data transmission to provide location results in almost real-time (Clark and Peters 2009). In this study a computer made the first-pass detection of whale presence in the array followed up by manual verification.

A third technique, the general signal-strength model, uses the relative peak power (similar to amplitude) of the same signal recorded at different microphones within the array and the spherical spreading rate of sound energy to calculate the location of the signal origin in a plane (Efford et al. 2009). One advantage to this last method is that it does not require strict time synchronization of the recorders; synchronization to the 1/100 of a second sufficed for their measurements and is more practical in an applied study than the method that relies on time differentials between microphones. In all locating arrays, the digital clocks in the recorders diverge over longer deployments and must therefore be repeatedly synchronized if the time is to be trusted to the required precision, a task that requires either manual intervention or a wireless link to satellite or cellular networks.

For bioacoustics applications mentioned thus far, the highest satisfaction with the results have been achieved by having humans scan the recordings from beginning to end and making manual measurements or having humans verify that the correct species is detected. A nascent option is to use computers to scan the recorded surveys and automate the data collection process. The primary drawback of requiring a human listener to evaluate and classify each recorded song is that it can be more time consuming than an *in situ* human survey. The ability to review each survey multiple times is traded against the benefit of visual field cues available to point counters, but generally it is more efficient to have a trained listener do the field assessments directly (Hutto and Stutzman 2009), especially if

the study design does not make numerous repeat visits at each location.

AUTOMATING BIOACOUSTICS

To automate data extraction from a recorded survey there must be a mathematical means of reducing the acoustic data to comparable metrics. Often this involves multiple transformations to the acoustic data - usually Fourier transforms to tease apart the energy at different frequencies - followed by statistical analysis. The recipe of transformations and statistical reduction is termed an analytical algorithm. Many algorithms in automated survey analysis have spun off of artificial speech recognition technology in theoretical literature, but only a handful have been implemented for general use by wildlife biologists. In both commercial and theoretical applications the pool of algorithms includes hidden Markov models based on spectrograms (Song ScopeTM: Wildlife Acoustics 2011 2012), mel-frequency cepstral coefficients (Lee et al. 2008), spectrogram cross-correlation (e.g. Avisoft-SASLab ProTM: Avisoft Bioacoustics e.K. 2014), binary point matching of spectrograms, (e.g. *monitoR*: Katz et al. In prepa). Additional options are linear predictive coding with geometric distance matching (SongID: Boucher and Jinnai 2012), and band-limited energy detection in spectrograms (Raven: Bioacoustics Research Program 2011) with parametric or non-parametric clustering (Ross and Allen 2014). The R package *soundecology* (Villanueva-Rivera and Pijanowski 2014) has functions to implement the Acoustic Complexity Index (Pieretti et al. 2011), the Acoustic Diversity Index (Villanueva-Rivera et al. 2011), and the Acoustic Evenness Index (Villanueva-Rivera et al. 2011), which are indices based on the overall "soundscape" or "biophony" as opposed to the presence of songs from individual species (Pijanowski et al. 2011). These methods are gaining a foothold in the bioacoustics community, and may grow more popular after being the subject of a "News Focus" in a recent issue of Science magazine

AUTOMATING BIOACOUSTICS

(Servick 2014).

As recently as 2007 the (now former) director of the Bioacoustic Research Program (BRP) at the Cornell Laboratory of Ornithology stated that detection and classification of signals is primarily a job for an experienced human operator (Urazghildiiev and Clark 2007). Supporting the BRP's lightly pessimistic view on automatic detection is the fact that few projects have successfully used it un-mediated for species inventory or monitoring, and it is still widely believed that the available software tools are incapable of these functions (Hutto and Stutzman 2009). The statement may not have been intended as dismissive, but instead it may have signified the BRP's intention to dedicate additional funding to algorithm research and analysis techniques. The BRP is one of the pioneering innovators in the field of bioacoustics and have developed the hardware, software, and methodology that form the foundation of most modern implementations. The BRP has also sponsored several innovative projects that demonstrate the capabilities of bioacoustics, including the real-time whale monitoring program in Cape Cod Bay (Clark and Peters 2009), the Elephant Listening Project which uses recordings to estimate forest elephant group size in dense tropical forest (Payne et al. 2003), and the use of automated detectors in the search for the ivory-billed woodpecker (Swiston and Mennill 2009). The report that Ivory-billed Woodpeckers were possibly not extinct but could be living undetected in the cypress swamps of the southeast United States (Fitzpatrick et al. 2005) provided what was perceived to be an ideal presence-absence project to test automated detection, but it became a story of epic failure. When Swiston and Mennill deployed autonomous recording units capable of timed surveys over extended periods, and they compared manual and automated methods of searching through hours of recordings made in the Florida panhandle for pileated woodpecker calls (as a surrogate for ivory-billed woodpecker). They determined that automated scanning using the software package XBAT (Figueroa 2011) was over 25 times faster than

manual browsing using the software package Syrinx PC but produced so many false positive detections and omitted so many detections that it was not appropriate for exclusive field use.

The problem they encountered was probably not the result of using recorded surveys. Celis-Murillo et al. (2009) found that the recorded counts often yielded higher detection probabilities. Acevedo and Villanueva-Rivera (2006) compared the presence data in recorded surveys to 20m fixed-radius point counts and concluded that despite the lack of visual data, the recordings allowed them to mark more species present than the point counts—likely because of the greater temporal coverage at each site in their study. The issue of high error rates in automated detection has roots in limitations of the artificial intelligence required to identify complex sounds in the recording. Objectively evaluating the strengths and weaknesses of each algorithm is difficult, and two studies may close with exactly opposite conclusions. For example, Khanna et al. (1997) were not at all impressed with spectrogram cross correlation as a detection algorithm, but Cortopassi and Bradbury (2000) offered compelling evidence that it could be a useful tool if understood and used correctly.

Regardless of the error rates reported by Swiston and Mennill, one aspect of XBAT that set it apart from other available detection software is that it was a Matlab extension, which meant that its code was open to all users for review and it was itself extensible—although in practice further extensibility was limited by the use of a graphical user interface. In theory XBAT offered the opportunity to understand the spectrogram cross correlation process to the degree Cortopassi and Bradbury suggested was necessary for success, and it is reasonable to conclude that some species are more difficult to detect than others. The short calls of the pileated woodpecker were difficult to detect accurately, but template matching algorithms can provide a useful level of specificity and sensitivity for species with more complex songs (Katz et al. In prep*b*). Nonetheless, the high error rate of these methods has

CORRELATION IN AUTOMATIC DETECTION DATA

inspired some researchers to develop indices to evaluate ecosystem health that don't require the identification of individual species.

The soundscape approach championed by Pijanowski, Villanueva-Rivera, and others is attractive because as a series of indices it cannot be subject to the same type of error rate evaluation. On the other hand, a strength of automated detection versus these indices and versus human detection is that the error rate can be easily quantified, and it will not be subject to inter-observer variation. Projects using computers to convert recordings to data have the potential to amass enormous quantities of detections, and this is an area where automatic detection can really shine. Many fields of study are exploring what “big data” can bring to networks and systems analysis, and one of our goals was to automate not just the detection process but the data management as well: by automating inputs with batch processing and parsing detections directly to a database, we hoped to take advantage of the fast data turnaround and constant error rates of automated detection in a manner that could produce datasets large enough to monitor entire landscapes for extended periods of time, which addressed an aspect of monitoring that had previously been ignored.

CORRELATION IN AUTOMATIC DETECTION DATA

In a template matching algorithm a template is passed over the spectrogram of the survey, scoring a discrete section of the survey for match, shifting one time bin in the spectrogram, and scoring the next discrete section of the survey. Local maxima are identified, and only those signals that exceed a threshold are considered “hits”. The proportion of hits that are correctly identified is known as the positive predictive value (PPV), and the proportion of

CORRELATION IN AUTOMATIC DETECTION DATA

hits that are incorrectly identified is known as the false discovery rate (FDR). The detection rate for the template includes two correct identification types and two error types: sounds can be correctly identified as from the target species or not (“true positives” and “true negatives”), or sounds can be incorrectly attributed to the target species or incorrectly skipped (“false positives” and “false negatives”). In past studies, the error rate of automated methods has usually been expressed as a per song event rate (Swiston and Mennill 2009, Goyette et al. 2011, Digby et al. 2013), while for human observers the error rate is typically expressed as a per survey rate, and often as a species detection probability (Forcey et al. 2006, Kissling and Garton 2006, Campbell and Francis 2011). Inference of species presence from survey results is based on several assumptions, including that the species was correctly identified (Royle and Link 2006). However, from a monitoring perspective, a site should be recorded as “occupied” as long as just one of the recorded hits is correctly identified, regardless of how many false positives were encountered.

Ideally we would like to gauge confidence in the accuracy of detection when given the results from each new survey. To convert the per-event error rate to a per-survey error rate it is tempting to evaluate the proportion of surveys with one or more hits. One approach could be to treat survey presence as a binary metric; an estimate of the true positive and false positive rate is possible with sufficient numbers of surveys, of which all have been manually verified. Another approach could be to base the degree of confidence on the number of detections in a survey, with more detections yielding higher probabilities of presence than single detections.

When the probability of presence is contingent on the number of detections, each additional detection can only result in increased probability of presence—no detection will ever reduce the probability of presence. With sufficient detections the probability of presence will there-

CORRELATION IN AUTOMATIC DETECTION DATA

fore always approach 1. This simple model is fine when detections are correctly identified, but it becomes less meaningful when any substantial number of incorrect detections is encountered in a single survey. In an automated detection algorithm false positives typically cannot be completely eradicated, but they are minimized by applying a threshold binary filter to the scored results. False positives occur when an unintended signal scores high enough to cross the threshold. If all false positives were random events it would be reasonable to expect one or two false positive detections per survey when roughly 150 sounds are evaluated during the survey and the overall false positive rate for the detector is a lowly 0.01. However, it is generally incorrect to assume that all detections in a survey are random, and therefore independent, as a number of factors are shared by any pair of detections in a short to moderate length survey.

The first source of correlation between signals is present in songs or calls issued by the same individual in a survey. The correlation arises due to the bird's singing behavior: many species repeatedly sing from the same perch while counter-singing (Wood and Perkins 2012), or have a favorite perch on territory from which songs are broadcast (Rossell 2001, Bolsinger 2000). Thus for shorter survey durations it is reasonable to expect many songs of similar amplitude, distance to recorder, and originating individual to be recorded. In this scenario the probability of detecting the first signal may be described by the true positive rate of the detector, but the probability that future signals issued by the same individual during the same survey will also be detected may be slightly greater than the true positive rate due to the correlation between the signals. Similarly, the probability that that signals after the first signal from an individual are not detected may be less than the true positive rate. The time that must elapse to regain independence between song events will vary by species; independence can be regained as a result of the individual singing a new song type or moving to a new perch.

CORRELATION IN AUTOMATIC DETECTION DATA

When an individual's song is learned from neighboring singers, it can be reasonable to treat the songs from different individuals of the same species singing during the survey as independent detections, but the presence of multiple singers may introduce a second source of correlation between signals. In areas where the target species occupies adjacent territories, individuals may counter-sing in a call-and-response pattern in which song from one individual will elicit song by another individual. The songs recorded during bouts of counter-singing may be very similar, as some species will attempt to match song type when counter-singing (Beecher et al. 1996). Other patterns beyond song matching may occur during counter-singing. Some species respond to songs of other species in addition to their own species, particularly when both species compete for resources (Martin and Martin 2001, Sedlacek et al. 2006). It is also possible for more than two birds, and even more than two species, to interact via call-and-response to songs of a single species (Moller 1992). For species that engage in counter-singing songs are not issued randomly but are clustered over time, and as a result there is an elevated probability of recording a song from individual "B" in a ten minute survey when individual "A" sings during the survey as well. If an automatic detector is capable of detecting both individuals A and B, then the probability of detecting the first song is again described by the true positive rate while subsequent songs in the interactive session have a slightly higher probability of detection than the true positive rate.

A third source of correlation between signals arises from the use of a biased detector. For example, if two bird species "C" and "D" do not typically occur in the same habitat, but a template designed to detect species C predictably and consistently returns a false positive when a song of species D occurs, then the false positive error of species C and species D are correlated even though the songs of the two species are otherwise independent of one another. If species D is falsely identified as species C many times in a survey, the resulting

CORRELATION IN AUTOMATIC DETECTION DATA

probability of presence for species C will not stand out as being unusually low despite the fact that all detections are in fact false positives.

When an automated detector correctly identifies the species of all singing individuals, the non-random singing may not be a problem—the state of species presence for the survey remains unchanged, even though the estimated probability of species presence is higher than supported by the data. Indeed, image-based automated detection algorithms such as spectrogram cross correlation, binary point matching, and hidden Markov model matching rely on some degree of correlation among songs. The matching algorithm in spectrogram cross correlation directly scores the correlation between a spectrogram of the template and spectrogram segments from a longer survey, and high correlations are considered detections (Mellinger and Clark 1997). Using this algorithm, it should be possible to identify the maximum spectrogram cross correlation between each template and every other sound - both bird songs and other more random noises - that may potentially be encountered in a survey, and to then design a template that has a measurably higher correlation to songs from a particular species than to all other sounds. In practice it is unrealistic to exhaustively test all alternative sounds. To combat the dependence between repeated songs, at least one study has assembled independent surveys from random one-minute recordings spliced together (Wimmer et al. 2013). However, a systematic false positive error that results in correlated error without correlated signal emission may not be fully addressed by rearranging the survey—it might be more practical to handle the issue of song dependence statistically.

Correlated detection is an issue with some wildlife surveys, and abundance models robust to its effects have been proposed (Martin et al. 2011*b*). A model to estimate presence that allows for correlated detections could assume that false positive detections are binomially distributed, with probability equal to the probability of getting a false positive conditional on a signal being detected (the “False Discovery Rate”, distinct from the false positive

CORRELATION IN AUTOMATIC DETECTION DATA

rate). Correlated error is common in regression analysis, where clustering by the shared factor is often done to avoid underestimating overall error. However, the mechanism by which we estimate probability of survey presence with correlated detection error is more similar to the way the structured finance markets assume correlated default risk when rating collateralized debt obligations. Moody's Investors Services has described how correlation effectively fattens the tails of a standard binomial, recognizing that with correlation the improbable outcomes are actually more probable than would be predicted by a standard binomial (Witt 2004). With a single-species detector based on song identification, two assumptions of the binomial are a good fit: there are a known number of detections in each survey, and each detection can have one of two outcomes with a constant probability of success. A mathematically elegant and appropriate model for correlated detector data is the correlated beta-binomial, in which the degree of correlation is integrated into the *alpha* and *beta* shape parameters for conjoined Beta distributions to yield an over-dispersed binomial distribution (Szabo 2013). The overdispersion effectively makes the extreme events in the tails of the binomial distribution more probable. With automated detection we are interested identifying the probability of the most extreme case: when all detections are false positives. That probability, plus the alternative that at least one detection is correct, accounts for all potential outcomes and sums to a probability of one. The probability that at least one detection is correct is the probability at which we are justified declaring the target species was present at the time of survey, although this may prove to be too ambitious for automated detection.

Using a correlated model reduces overestimation of the probability of species presence within a survey when multiple detections are made by a single template, but it has no effect on the probability of survey presence when only one song detection is made in a survey. In surveys where only a single detection is made, the probability it is a true positive detection

CORRELATION IN AUTOMATIC DETECTION DATA

is equivalent to the positive predictive value of the detector. A high quality detector should have a positive predictive value close to 1; however, as mentioned earlier, even low false positive rates can be problematic when hundreds of sounds are evaluated per survey, as a single random false positive will predict the species is present with probability equal to the positive predictive value. A single false positive can also be problematic if the survey is evaluated with a binary presence state, as a single random false positive will allow the presence to be erroneously reported as true. If the mean song rate for the target species is known to be greater than 0.5 songs per minute, there is a good reason to expect more than two songs from the target species in a ten minute survey. Random false positives can be removed by eliminating the first two detections in each survey, and evaluating the probability of presence using the number of remaining detections. The tradeoff is reduced overall sensitivity of the detector, as some proportion of surveys will contain only one or two true positive detections and will thus be indistinguishable from random false positive surveys. For this reason eliminating the first detections will be most useful when the the number of expected songs for the survey duration is greater than the number of detections eliminated, and the practice is probably unsuitable for detection of rare species (those with low expected occupancy probabilities) or elusive species (those with low detection probabilities).

The effect of evaluating wildlife detections using a correlated model versus an uncorrelated model was described by Martin et al. (2011*b*). When detection correlation is the sole violation of an abundance model, the resulting estimate of abundance will be biased high, which is undesirable when working with vulnerable wildlife populations. On the other hand, when correlation is assumed in the model but not present in the detection method, the resulting estimate will likely be biased low—which is also undesirable if funds are allocated for unnecessary restoration efforts.

CORRELATION IN AUTOMATIC DETECTION DATA

Correlated detection error is not used for human point counts because we typically - and sometimes unrealistically - assume a negligible false positive rate: the probability of species presence in a survey is assumed to be 1 when a single detection is made. In contrast, the quantifiable error rates of automated detectors can be used to estimate the probability of species presence in a survey given a) the number of detections in a survey, b) the false positive error rate per-detection, and c) the potential bias introduced when the error rate between calls is not independent. In this framework each detection is similar to a Bernoulli trial of the detection algorithm. The outcome of the test is the number of detections, and we are trying to estimate site occupancy based on the result of our test.

The result of using a correlated likelihood to identify the probability that at least one detection in a survey is correctly identified is the probability that a species was present, taking into account the number of detections observed during the survey. This probability of presence is not an occupancy probability because the parameters that feed it are survey metrics (number of detections) potentially biased by the detector as opposed to site metrics (number of surveys with observations). A popular method of estimating occupancy is to analyze the results of repeated surveys with a multinomial model that allows for false negative error, or imperfect detection, and then predict occupancy at spatial locations between survey points (MacKenzie et al. 2002). Variations of the model extend the sampling across multiple seasons to estimate colonization and extinction rates (MacKenzie et al. 2003), introduce mixtures to the likelihood to allow for false positive errors (Royle and Link 2006) and estimation of abundance (Royle and Nichols 2003), and account for improper species identification that can be verified by alternative methods (Miller et al. 2011). The multinomial occupancy model has become a very useful tool for wildlife managers, primarily due to the ease of data collection, its ability to compensate for some detection error, and the ease of converting the output to spatially explicit forecasts. Aside from a few published

case studies, it is unknown how popular the N-mixture and multi-state variants are, and how many researchers or land-management practitioners are using them.

ESTIMATING OCCUPANCY FROM AUTOMATED DETECTION DATA

The multinomial occupancy framework is shaping up to be one of the most valuable analytical tools in the wildlife monitoring toolkit. General enthusiasm for the model may be based on three factors: it uses easy to collect data to adjust for imperfect detection, it can be spatially and temporally explicit, and dedicated software to implement the analysis is freely available for use by wildlife biologists and conservation practitioners.

Detection probability is the probability that a randomly selected individual of a population is detected by an observer during a survey (Farnsworth et al. 2005), or alternately as a correction factor to account for individuals present but not detected (Thompson 2002). Rosenstock et al. (2002) observed that 95% of 224 unique published landbird studies between 1989 and 1998 used index counts without estimating detection probability, including large projects such as the Breeding Bird Survey. The issue of sub-optimal detection is fundamental to wildlife monitoring where many species are reclusive, rare on the landscape, or nocturnal, and as a result cannot be reliably observed in general surveys. Although it is possible to estimate the ratio of actual observations to expected observations using methods other than the multinomial occupancy model, the extra effort to do so is rarely expended (Farnsworth et al. 2005, McCallum 2005). Expending the extra effort during a bird survey may involve using an extra observer during counts, estimating distance to all

ESTIMATING OCCUPANCY FROM AUTOMATED DETECTION DATA

birds observed, or using a fixed radius count. Of these three options only the first can be used to estimate what fraction of birds are truly present but not observed, while the second two assume that 100% of individuals nearest to the observer are detected.

The multinomial occupancy model uses a binary detection history comprised of presence/absence data from multiple individual surveys, between which it is assumed that the occupancy state at the site remains unchanged and the population remains closed to immigration and emmigration (MacKenzie et al. 2002). In addition to simultaneously estimating occupancy and detection probabilities, the framework of the model can be used to establish how many surveys to perform in order to optimize sampling under a variety of occupancy/detection probability scenarios. For example, the optimum sampling strategy for common species (high expected occupancy probability) is to perform many surveys across a few sites to achieve reasonably accurate detection estimates, while an optimum sampling strategy for rare species (low occupancy probability) is to perform a few surveys across many sites (Mackenzie and Royle 2005). One factor that makes the model attractive is that to collect the binary presence/absence data it is rarely necessary to change the sampling procedure from existing protocol; as a result, it is easily applied to unlimited radius point count data for birds, systematic search data for amphibians, or scat count data for forest mammals. The most popular variants of the model, of those mentioned above and as cited in the literature, appear to be the single-season model to estimate occupancy and detection probabilities (cited approximately 83 times per year in other published papers) and the multi season model which adds estimates for colonization and extinction rates (cited approximately 45 times per year in other published papers).

The flexibility of the model to include separate detection and occupancy covariates allows for two useful extensions beyond simultaneous estimation of occupancy and detection prob-

ESTIMATING OCCUPANCY FROM AUTOMATED DETECTION DATA

abilities. A variety of models can be created with different covariates, and the covariates that are most influential to either occupancy or detection can be identified using a model selection and model averaging framework, such as that described in Anderson (2010) or Burnham and Anderson (2002). This becomes a powerful predictive tool, as once a suite of influential occupancy covariates are identified it is possible to predict the occupancy probability for a species at locations other than those surveyed. With a sufficient number of surveys spread across a landscape, spatially explicit predictive occupancy maps can be used to inform wildlife management at any scale, from local to landscape (Schwenk and Donovan 2011). It is also then possible to predict the effect that different management practices or changes in land use will have on individual species over time (Brown et al. 2014).

The single season and multiseason models are especially easy to implement using the free software PRESENCE (MacKenzie and Hines 2014). PRESENCE remains in a state of active development by the USGS for both stability and modifications to accomodate new analytical goals. Meanwhile, the R package unmarked (Fiske and Chandler 2011) was created as an alternative software package, which also contains functions to accomodate most analytical goals using the multinomial occupancy models with unmarked animals. Having functional software dedicated to analyzing wildlife data has made analysis using the multinomial occupancy model accessible to nearly any wildlife monitoring organization with the means to collect data, and the near ubiquity of the method improves the ability of monitoring organizations to reanalyze or evaluate results from other projects. Without these software options wildlife monitoring projects would be left to code their own analyses, an option that is now possible without additional means using a platform such as R (R Core Team). Of course, the ability to program custom analyses for repeated use or to share with colleagues is a relatively recent development in wildlife conservation, and the fact that it was previously unavailable may be one of the reasons Rosenstock et al. (2002) observed such

ESTIMATING OCCUPANCY FROM AUTOMATED DETECTION DATA

a small fraction of projects doing the extra analysis to convert index counts to measures of density or abundance.

With many projects using similar occupancy models and a collaborative culture building between research entities, it is reasonable that multiple local projects would want to collaborate toward landscape-scale inference. However, and perhaps as testament to the flexibility of the analytical framework, even neighboring projects are unlikely to be using the same data collection process. Since all data collection processes are inherently biased it is questionable whether mixing data from different sources will, in a best case scenario, produce models with greater support for some covariates than any individual model, or in a worst case scenario result in less agreement of which covariates are important. The logical answer is to evaluate all instances of potential data sharing for feasibility independently. Some studies have combined multiple survey methods to increase sample size and reduce sampling bias (Kendall et al. 2008, Boulanger et al. 2004), while others have pointed out that these benefits are realized only when the survey methods sample the same population, or different segments of the population that respond to the same environmental covariates (Graves et al. 2012). When the methods sample populations that respond to different covariates, the effect of important predictive variables may not be evident in the model. This concurs with the statistical sampling philosophy in which data from multiple sources only be considered for pooling when the methods used sample different sections of the same population (Gelman et al. 1995).

The ultimate goal of automated acoustic monitoring is to collect data that can be used to identify trends that would be difficult to detect using other data collection methods. Monitoring species occupancy as a parameter of interest has been suggested due to the efficiency of data collection and the flexibility of the results for evaluating trends beyond occupancy (Noon et al. 2012). In addition to the alternative uses for occupancy models listed earlier,

ESTIMATING OCCUPANCY FROM AUTOMATED DETECTION DATA

occupancy as a monitored system state can be directly used to identify synchronized local changes in breeding area across widely separated areas. With anything short of high spatial sampling resolution with a large extent it would be difficult to identify the synchronicity of the local changes and thus conclude that a trend is forming. Another example would be changes to migration timing, which would be most evident in long-term data sets that would ideally maintain a constant error rate. For each of these examples, the parameter of interest would be daily occupancy probability estimates so that we can ask in which areas has the area of high occupancy probability grown vs shrank indicating a range shift, or on which date did occupancy probability rise indicating spring arrival? One method could be to use survey occupancy as a metric, and another could be to use a moving window history of presence/absence observations.

Automated acoustic monitoring has the potential to sample an area continuously for species presence, but to rely completely on automated detection methods to convert the surveys to data would be putting all of the metaphorical monitoring eggs into a single basket, and the relatively high error rates of automated detection methods would be analogous to that basket having holes big enough to lose a few eggs. Ideally the entire survey set would be converted to presence data by a human, but that method is too inefficient to be considered a viable alternative. Second best would be to rely on an automated detection algorithm for the majority of detections and manually validate a fraction of the results. It is known that manually validated data does not have an error rate of zero (Fritsch and Bruckner 2014), but a process to combine the results of the two methods would be more useful than simply acknowledging the different error rates.

DEVELOPING VALIDATION THEORY FOR AUTOMATED ACOUSTIC MONITORING

Validation theory makes extensive use of the fundamentals of statistical inference, but the goal of results validation changes as the process being validated moves from production to detection. While production or manufacturing processes use validation to identify faults and evaluate the need for corrective action, detector validation differs in that there is rarely corrective action available—the test is presumably already as accurate as possible, and if not it is replaced with a new test. With automated bioacoustic monitoring it is possible to re-make a template or hidden Markov model recognizer and subsequently re-analyze historical data using the newest test, but as monitoring histories become deeper a program will face the decision to abandon full historical results comparability, apply a corrective coefficient to old results, or reanalyze archived surveys with a new detector. This represents a non-trivial decision for long-term monitoring programs, as creating a new detector will require validation efforts to be repeated to ensure that accuracy has truly improved. Rather than use the validation process to weigh whether to take corrective measures, detector validation is typically used to characterize overall detector error. From the overall error rate confidence in a single positive detection can be deduced, and the error itself will often influence management decisions.

The degree of confidence in a single positive test is based on measurements of variation when a sample of tests is compared against known outcomes. Detector validation refers to any method that links the test to known outcomes. It is most common in the medical field, where tests often do not directly identify disease state but instead identify the risk of disease state (Mossman and Berger 2001). The value of a test is its predictive value weighed against the

*DEVELOPING VALIDATION THEORY FOR AUTOMATED ACOUSTIC
MONITORING*

cost of validation; most tests are prone to small amounts of error, but information from the results can influence the decision for more costly or more invasive validation. The process of validating tests is typically much more costly than the test itself, and therefore as validation is ideally performed on the smallest sample that still allows inference in the majority of test cases. Given how new automated monitoring is to natural resources management, it is not surprising that there is little information available to monitoring programs to guide the validation of automated detection results.

REFERENCES

- Acevedo, M. A., and L. J. Villanueva-Rivera. 2006. Using Automated Digital Recording Systems as Effective Tools for the Monitoring of Birds and Amphibians. *Wildlife Society Bulletin* **34**:211–214.
- Amar, A., C. M. Hewson, R. M. Thewlis, K. W. Smith, R. J. Fuller, J. A. Lindsell, G. Conway, S. Butler, and M. A. MacDonald. 2006. What’s happening to our woodland birds? Long-term changes in the populations of woodland birds. A re-survey of breeding bird populations of broadleaved and mixed woodlands in 2003-04 to examine changes since the 1960s, 1970s and 1980s and test these changes against a range of hypotheses for their causes. *RSPB Research Report* **19**:i–vi, 1–204.
- Anderson, D. R. 2010. *Model Based Inference in the Life Sciences: A Primer on Evidence*. Springer. New York.
- Avisoft Bioacoustics e.K., 2014. Avisoft-SASLab Pro version 5.2 [Computer Software]. URL <http://www.avisoft.com/>.
- Barbraud, C., and H. Weimerskirch. 2006. Antarctic birds breed later in response to climate change. *Proceedings of the National Academy of Sciences of the United States of America* **103**:6248–6251.

REFERENCES

- Bart, J., and J. D. Schoultz. 1984. Reliability of Singing Bird Surveys: Changes in Observer Efficiency with Avian Density. *The Auk* **101**:307–318.
- Beecher, M., P. Stoddard, S. Campbell, and C. Horning. 1996. Repertoire matching between neighbouring song sparrows. *Animal Behaviour* **51**:917–923.
- Bioacoustics Research Program, 2011. Raven Pro: Interactive Sound Analysis Software [Computer Software]. URL <http://www.birds.cornell.edu/raven>.
- Blickley, J. L., D. Blackwood, and G. L. Patricelli. 2012. Experimental Evidence for the Effects of Chronic Anthropogenic Noise on Abundance of Greater Sage-Grouse at Leks. *Conservation Biology* **26**:461–471.
- Bolsinger, J. S. 2000. Use of two song categories by Golden-cheeked Warblers. *Condor* **102**:539–552.
- Both, C., S. Bouwhuis, C. M. Lessells, and M. E. Visser. 2006. Climate change and population declines in a long-distance migratory bird. *Nature* **441**:81–83.
- Boucher, N. J., and M. Jinnai, 2012. SoundID Documentation: A Revolutionary Sound Recognition System. URL <http://www.soundid.net>.
- Boulanger, J., S. Gordon, and R. Munro. 2004. Sources of Heterogeneity Bias When DNA Mark-Recapture Sampling Methods Are Applied to Grizzly Bear (*Ursus arctos*) Populations. *Journal of Mammalogy* **85**:618–624. URL <http://www.jstor.org/stable/1383780>.
- Brewster, J. P., and T. R. Simons. 2009. Testing the importance of auditory detections in avian point counts. *Journal of Field Ornithology* **80**:178–182.
- Brown, M. L., T. M. Donovan, W. S. Schwenk, and D. M. Theobald. 2014. Predicting impacts of future human population growth and development on occupancy rates of forest-dependent birds. *Biological Conservation* **170**:311–320.

REFERENCES

- Burnham, K. P., and D. R. Anderson. 2002. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. Springer. New York.
- Burt, J., 2006. Syrinx-PC. URL <http://www.syrinxpc.com/>.
- Butler, C. J. 2003. The disproportionate effect of global warming on the arrival dates of short-distance migratory birds in North America. *Ibis* **145**:484–495.
- Campbell, M., and C. M. Francis. 2011. Using stereo-microphones to evaluate observer variation in North American Breeding Bird Survey point counts. *Auk* **128**:303–312.
- Carson, R. 1987. Silent Spring. 25th anniversary edition edition. Houghton Mifflin, Boston.
- 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.
- Clark, C. W. 1980. A real-time direction finding device for determining the bearing to the underwater sounds of Southern Right Whales, *Eubalaena australis*. *The Journal of the Acoustical Society of America* **68**:508–511.
- Clark, C. W., and D. B. Peters. 2009. Acoustic System Monitors and Mitigates Harm to Marine Mammals in Real Time. *Sea Technology* **50**:10–14.
- Cortopassi, K. A., and J. W. Bradbury. 2000. The comparison of harmonically related rich sounds using spectrographic cross-correlation and principal components analysis. *Bioacoustics* **11**:89–127.
- Digby, A., M. Towsey, B. D. Bell, and P. D. Teal. 2013. A practical comparison of manual and autonomous methods for acoustic monitoring. *Methods in Ecology and Evolution* **4**:675–683.

REFERENCES

- Donovan, T. M., and C. H. Flather. 2002. Relationships among north American songbird trends, habitat fragmentation, and landscape occupancy. *Ecological Applications* **12**:364–374.
- Efford, M. G., D. K. Dawson, and D. L. Borchers. 2009. Population density estimated from locations of individuals on a passive detector array. *Ecology* **90**:2676–2682.
- Evans, W. R., and D. K. Mellinger. 1999. Monitoring grassland birds in nocturnal migration. *Studies in Avian Biology* **19**:219–229,.
- Farmer, R. G., M. L. Leonard, J. E. M. Flemming, and S. C. Anderson. 2014. Observer aging and long-term avian survey data quality. *Ecology and Evolution* **4**:2563–2576.
- Farmer, R. G., M. L. Leonard, and A. G. Horn. 2012. Observer effects and avian-call-count survey quality: rare-species biases and overconfidence. *Auk* **129**:76–86.
- Farnsworth, G. L., J. D. Nichols, J. R. Sauer, S. G. Fancy, K. H. Pollock, S. A. Shriner, and T. R. Simons, 2005. Statistical Approaches to the Analysis of Point Count Data: A Little Extra Information Can Go a Long Way. USDA Forest Service General Technical Report PSW-GTR-191, 736–743. URL http://www.fs.fed.us/psw/publications/documents/psw_gtr191/Asilomar/pdfs/736-743.pdf.
- Figuerola, H., 2011. XBAT [Computer Software]. Bioacoustics Research Program. URL [Notavailable](#).
- Fiske, I., and R. Chandler. 2011. unmarked: An R Package for Fitting Hierarchical Models of Wildlife Occurrence and Abundance. *Journal of Statistical Software* **43**:1–23. URL <http://www.jstatsoft.org/v43/i10/>.
- Fitzpatrick, J. W., M. Lammertink, M. D. Luneau, T. W. Gallagher, B. R. Harrison, G. M. Sparling, K. V. Rosenberg, R. W. Rohrbaugh, E. C. H. Swarthout, P. H. Wrege, S. B. Swarthout, M. S. Dantzker, R. A. Charif, T. R. Barksdale, J. V. Remsen, S. D. Simon,

REFERENCES

- and D. Zollner. 2005. Ivory-billed woodpecker (*Campephilus principalis*) persists in continental North America. *Science* **308**:1460–1462.
- Forcey, G. 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.
- Forero-Medina, G., J. Terborgh, S. J. Socolar, and S. L. Pimm. 2011. Elevational Ranges of Birds on a Tropical Montane Gradient Lag behind Warming Temperatures. *Plos One* **6**:e28535.
- Fritsch, G., and A. Bruckner. 2014. Operator bias in software-aided bat call identification. *Ecology and Evolution* **4**:2703–2713.
- Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 1995. *Bayesian Data Analysis*. Chapman and Hall. New York, NY.
- Gibbs, J. P., H. L. Snell, and C. E. Causton. 1999. Effective monitoring for adaptive wildlife management: Lessons from the Galapagos Islands. *Journal of Wildlife Management* **63**:1055–1065.
- 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.
- Graves, T. A., J. A. Royle, K. C. Kendall, P. Beier, J. B. Stetz, and A. C. Macleod. 2012. Balancing Precision and Risk: Should Multiple Detection Methods Be Analyzed Separately in N-Mixture Models? *Plos One* **7**. URL [GotoISI>://WOS:000313236200015](https://doi.org/10.1371/journal.pone.0003132).
- Habib, L., E. M. Bayne, and S. Boutin. 2007. Chronic industrial noise affects pairing success and age structure of ovenbirds *Seiurus aurocapilla*. *Journal of Applied Ecology* **44**:176–184.

REFERENCES

- Harris, J. B. C., D. D. Putra, S. D. Gregory, B. W. Brook, D. M. Prawiradilaga, N. S. Sodhi, D. Wei, and D. A. Fordham. 2014. Rapid deforestation threatens mid-elevational endemic birds but climate change is most important at higher elevations. *Diversity and Distributions* **20**:773–785.
- 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.
- Heldbjerg, H., and T. Fox. 2008. Long-term population declines in Danish trans-Saharan migrant birds. *Bird Study* **55**:267–279.
- Hobson, K. A., R. S. Rempel, H. Greenwood, B. Turnbull, and S. L. V. 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.
- Johnson, L., R. J. Camp, K. W. Brinck, and P. C. Banko. 2006. Long-Term Population Monitoring: Lessons Learned From an Endangered Passerine in Hawai'i. *Wildlife Society Bulletin* **34**:1055–1063.
- Johnson, R. R., B. T. Brown, L. T. Haight, and J. M. Simpson, 1981. Playback recordings as a special avian censusing technique, volume 6 of *Studies in Avian Biology*, Book section estimating relative abundance - part ii, pages 68–75 . Cooper Ornithological Society.
- Katz, J., S. Hafner, and T. Donovan. In prep^a. Avian acoustic monitoring with the R package `monitoR` .
- Katz, J., S. Hafner, and T. Donovan. In prep^b. Evaluating error rates of the R package `monitoR` .

REFERENCES

- Kendall, K. C., J. B. Stetz, D. A. Roon, L. P. Waits, J. B. Boulanger, and D. Paetkau. 2008. Grizzly Bear Density in Glacier National Park, Montana. *The Journal of Wildlife Management* **72**:1693–1705. URL <http://www.jstor.org/stable/40208452>.
- Kendall, W. L., B. G. Peterjohn, and J. R. Sauer. 1996. First-time observer effects in the North American Breeding Bird Survey. *Auk* **113**:823–829.
- Kepler, C. B., and J. M. Scott. 1981. Reducing bird count variability by training observers. *Studies in Avian Biology* **6**:366–371.
- Khanna, H., S. L. L. Gaunt, and D. A. McCallum. 1997. Digital spectrographic cross-correlation: tests of sensitivity. *Bioacoustics* **7**:209–234.
- 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. Estimaciones de la Densidad y de las Probabilidades de Detección a Partir de Muestreos Utilizando Conteos en Puntos: Una Combinación de Muestreos de Distancia y de Doble Observador. **123**:735–752.
- Lee, C.-H., C.-C. Han, and C.-C. Chuang. 2008. Automatic Classification of Bird Species From Their Sounds Using Two-Dimensional Cepstral Coefficients. *IEEE Transactions on audio speech and language processing* **16**:1541–1550.
- Legg, C. J., and L. Nagy. 2006. Why most conservation monitoring is, but need not be, a waste of time. *Journal of Environmental Management* **78**:194–199.
- Luther, D., and L. Baptista. 2010. Urban noise and the cultural evolution of bird songs. *Proceedings of the Royal Society B: Biological Sciences* **277**:469–473.
- MacKenzie, D., and J. Hines, 2014. PRESENCE [Computer Software]. URL <http://www.mbr-pwrc.usgs.gov/software/presence.html>.

REFERENCES

- Mackenzie, D., and J. Royle. 2005. Designing occupancy studies: general advice and allocating survey effort. *Journal of Applied Ecology* **42**:1105–1114.
- MacKenzie, D. I., J. D. Nichols, J. E. Hines, M. G. Knutson, and A. B. Franklin. 2003. Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly. *Ecology* **84**:2200–2207.
- MacKenzie, D. I., J. D. Nichols, G. B. Lachman, S. Droege, J. A. Royle, and C. A. Langtimm. 2002. Estimating site occupancy rates when detection probabilities are less than one. *Ecology* **83**:2248–2255.
- Martin, J., P. L. Fackler, J. D. Nichols, B. C. Lubow, M. J. Eaton, M. C. Runge, B. M. Stith, and C. A. Langtimm. 2011*a*. Structured decision making as a proactive approach to dealing with sea level rise in Florida. *Climatic Change* **107**:185–202.
- Martin, J., J. A. Royle, D. I. Mackenzie, H. H. Edwards, M. Kery, and B. Gardner. 2011*b*. Accounting for non-independent detection when estimating abundance of organisms with a Bayesian approach. *Methods in Ecology and Evolution* **2**:595–601.
- Martin, P., and T. Martin. 2001. Behavioral interactions between coexisting species: Song playback experiments with wood warblers. *Ecology* **82**:207–218.
- Martindale, S. 1980. A Numerical Approach to the Analysis of Solitary Vireo Songs. *The Condor* **82**:199–211.
- McCallum, D. A., 2005. A Conceptual Guide to Detection Probability for Point Counts and Other Count-based Survey Methods. USDA Forest Service General Technical Report PSW-GTR-191, 754–761. URL http://www.fs.fed.us/psw/publications/documents/psw_gtr191/Asilomar/pdfs/754-761.pdf.
- McGregor, P. K., T. Dabelsteen, C. W. Clark, J. L. Bower, J. P. Tavares, and J. Holland.

REFERENCES

1997. Accuracy of a passive acoustic location system: empirical studies in terrestrial habitats. *Ethology Ecology & Evolution* **9**:269–286.
- Mellinger, D. K., and C. W. Clark. 1997. Methods for automatic detection of mysticete sounds. *Marine and Freshwater Behaviour and Physiology* **29**:163–181.
- Mennill, D. J., J. M. Burt, K. M. Fristrup, and S. L. Vehrencamp. 2006. Accuracy of an acoustic location system for monitoring the position of duetting songbirds in tropical forest. *Journal of the Acoustical Society of America* **119**:2832–2839.
- Miller, D. A., J. D. Nichols, B. T. McClintock, E. H. C. Grant, L. L. Bailey, and L. A. Weir. 2011. Improving occupancy estimation when two types of observational error occur: non-detection and species misidentification. *Ecology* **92**:422–428.
- Moller, A. P. 1992. Interspecific Response To Playback of Bird Song. *Ethology* **90**:315–320.
- Moore, J. E., D. M. Scheiman, and R. K. Swihart. 2004. Field comparison of removal and modified double-observer modeling for estimating detectability and abundance of birds. *Auk (American Ornithologists Union)* **121**:865–876.
- Mossman, D., and J. O. Berger. 2001. Intervals for Post-Test Probabilities: A Comparison of Five Methods. *Medical Decision Making* **21**:498–507.
- Nichols, J. D., and B. K. Williams. 2006. Monitoring for conservation. *Trends in Ecology & Evolution* **21**:668–673.
- Noon, B. R., L. L. Bailey, T. D. Sisk, and K. S. McKelvey. 2012. Efficient Species-Level Monitoring at the Landscape Scale. *Conservation Biology* **26**:432–441.
- Parker, T. A. 1991. On the use of tape-recorders in avifaunal surveys. *Auk* **108**:443–444.
- Payne, K. B., M. Thompson, and L. Kramer. 2003. Elephant calling patterns as indicators

REFERENCES

- of group size and composition: the basis for an acoustic monitoring system. *African Journal of Ecology* **41**:99–107.
- Pieretti, N., A. Farina, and D. Morri. 2011. A new methodology to infer the singing activity of an avian community: The Acoustic Complexity Index (ACI). *Ecological Indicators* **11**:868–873.
- Pijanowski, B. C., A. Farina, S. H. Gage, S. L. Dumyahn, and B. L. Krause. 2011. What is soundscape ecology? An introduction and overview of an emerging new science. *Landscape Ecology* **26**:1213–1232.
- R Core Team, 2012. R: A language and environment for statistical computing. URL <http://www.R-project.org/>.
- Rappole, J. H., and M. V. McDonald. 1994. Cause and effect in population declines of migratory birds. *Auk* **111**:652–660.
- 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.
- Rodenhouse, N. L., S. N. Matthews, K. P. McFarland, J. D. Lambert, L. R. Iverson, A. Prasad, T. S. Sillett, and R. T. Holmes. 2008. Potential effects of climate change on birds of the Northeast. *Mitigation and Adaptation Strategies for Global Change* **13**:517–540.
- Rosenstock, S. S., D. R. Anderson, K. M. Giesen, T. Leukering, and M. F. Carter. 2002. Landbird counting techniques: Current practices and an alternative. *AUK* **119**:46–53.
- Ross, J. C., and P. E. Allen. 2014. Random Forest for improved analysis efficiency in passive acoustic monitoring. *Ecological Informatics* **21**:34–39.

REFERENCES

- Rossell, C. R. 2001. Song perch characteristics of Golden-winged Warblers in a mountain wetland. *Wilson Bulletin* **113**:246–248.
- 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., and J. D. Nichols. 2003. Estimating abundance from repeated presence-absence data or point counts. *Ecology* **84**:770–790.
- Runge, M. C., J. R. Sauer, M. L. Avery, B. F. Blackwell, and M. D. Koneff. 2009. Assessing Allowable Take of Migratory Birds. *Journal of Wildlife Management* **73**:556–565.
- Sanderson, F. J., P. F. Donald, D. J. Pain, I. J. Burfield, and F. P. J. van Bommel. 2006. Long-term population declines in Afro-Palearctic migrant birds. *Biological Conservation* **131**:93–105.
- Sauer, J. R., B. G. Peterjohn, and W. A. Link. 1994. Observer differences in the North-American Breeding Bird Survey. *Auk* **111**:50–62.
- Schwenk, W. S., and T. M. Donovan. 2011. A Multispecies Framework for Landscape Conservation Planning. *Conservation Biology* **25**:1010–1021.
- Sedlacek, O., B. Cikanova, and R. Fuchs. 2006. Heterospecific rival recognition in the Black Redstart (*Phoenicurus ochruros*). *Ornis Fennica* **83**:153–161.
- Servick, K. 2014. Eavesdropping on Ecosystems. *Science* **343**:834–837.
- Slabbekoorn, H., and M. Peet. 2003. Ecology: Birds sing at a higher pitch in urban noise. *Nature* **424**:267–267.
- Sueur, J., T. Aubin, and C. Simonis. 2008. Seewave: a free modular tool for sound analysis and synthesis. *Bioacoustics* **18**:213–226.

REFERENCES

- Swiston, K. A., and D. J. Mennill. 2009. Comparison of manual and automated methods for identifying target sounds in audio recordings of Pileated, Pale-billed, and putative Ivory-billed woodpeckers. *Journal of Field Ornithology* **80**:42–50.
- Szabo, A., 2013. CorrBin: Nonparametrics with clustered binary and multinomial data. R package version 1.4. URL <http://CRAN.R-project.org/package=CorrBin>.
- Thompson, W. L. 2002. Towards Reliable Bird Surveys: Accounting For Individuals Present But Not Observed. *The Auk* **119**:18–25.
- Urazghildiiev, I. R., and C. W. Clark. 2007. Detection performances of experienced human operators compared to a likelihood ratio based detector. *Journal of the Acoustical Society of America* **122**:200–204.
- Villanueva-Rivera, L. J., and B. C. Pijanowski, 2014. soundecology: Soundscape ecology. R package version 1.1.1. URL <http://CRAN.R-project.org/package=soundecology>.
- Villanueva-Rivera, L. J., B. C. Pijanowski, J. Doucette, and B. Pekin. 2011. A primer of acoustic analysis for landscape ecologists. *Landscape Ecology* **26**:1233–1246.
- Wildlife Acoustics, 2011. Song Scope [Computer Software]. URL <http://www.wildlifeacoustics.com>.
- Wimmer, J., M. Towsey, P. Roe, and I. Williamson. 2013. Sampling environmental acoustic recordings to determine bird species richness. *Ecological Applications* **23**:1419–1428.
- Wintle, B. A., M. C. Runge, and S. A. Bekessy. 2010. Allocating Monitoring Effort In The Face Of Unknown Unknowns. *Ecology Letters* **13**:1325–1337.
- Witt, G. 2004. Moody’s Correlated Binomial Default Distribution. Working Paper (Moody’s Investors Service: Structured Finance Rating Methodology). URL: http://fcic-static.law.stanford.edu/cdn_media/fcic-testimony/2010-0602-exhibit-binomial.pdf.

REFERENCES

- Wood, P. B., and K. A. Perkins. 2012. Behavioral Activities of Male Cerulean Warblers in Relation to Habitat Characteristics. *Wilson Journal of Ornithology* **124**:497–505.
- Yoccoz, N. G., J. D. Nichols, and T. Boulinier. 2001. Monitoring of biological diversity in space and time. *Trends in Ecology & Evolution* **16**:446–453.

CHAPTER 2

TOOLS FOR AUTOMATED ACOUSTIC MONITORING WITHIN THE R PACKAGE `monitoR`

SUMMARY

Tools for automated acoustic monitoring within the R package **monitoR**

Jonathan Katz

Vermont Cooperative Fish and Wildlife Research Unit,
University of Vermont, USA (jonkatz4@gmail.com)

Sasha D. Hafner

University of Southern Denmark,
Odense, Denmark (saha@kbm.sdu.dk, sdh11@cornell.edu)

Therese Donovan

U.S. Geological Survey,
Vermont Cooperative Fish and Wildlife Research Unit,
University of Vermont, USA (tdonovan@uvm.edu)

SUMMARY

The R package **monitoR** contains tools for managing survey metadata, template creation and manipulation, automated detection, and results management. These tools are scalable for use with small projects as well as larger long-term projects and those with expansive spatial extents. Here we describe typical workflow when using the automated tools in **monitoR**. Typical workflow utilizes a generic sequence of functions, with the option for either binary point matching or spectrogram cross correlation.

Prepublication Disclaimer

This draft manuscript is distributed solely for purposes of scientific peer review. Its content is deliberative and predecisional, so it must not be disclosed or released by reviewers. Because the manuscript has not yet been approved for publication by the U.S. Geological Survey (USGS), it does not represent any official USGS finding or policy.

INTRODUCTION

Repeated surveys to monitor wildlife presence supply the data from which broader species occupancy can be extrapolated (MacKenzie et al. 2002), and they may also serve to fulfill broader management objectives (Yoccoz et al. 2001, Legg and Nagy 2006, Nichols and Williams 2006). Human surveillance is typically used to gather species occurrence data because of our high mobility and perceived low error rate, but there is increasing recognition that the non-zero error rate has some effect on overall data quality (Kepler and Scott 1981, Bart and Schoultz 1984, Sauer et al. 1994, Kendall et al. 1996, Farmer et al. 2012). For the occasions where mobility is not a requirement monitoring programs have a variety of weatherproof hardware options with which to record surveys as digital audio files. Having an audio recording of the survey allows the presence of reclusive species to be evaluated by extensive and repeated review, it allows surveys to be archived for future reanalysis (Parker 1991, Haselmayer and Quinn 2000, Hobson et al. 2002, Acevedo and Villanueva-Rivera 2006), and it allows for simultaneous surveys at numerous sites. The potential disadvantages are 1) diminished survey radius, as the average human ear can hear quieter sounds - such as distant bird songs - than the average all-weather microphone and recorder; 2) high start-up costs for equipment and training; and 3) commitment to only auditory detections, which rules out as many as 3% of potential bird detections (Brewster and Simons 2009). However, when recorded surveys are analyzed automatically, a second set of advantages are realized: a) the capacity for rapid data availability and b) detections are unbiased by observer experience or ability. Both of these advantages are specific to automated detection methods, and both have positive consequences in wildlife management decision making.

Many aspects of automated survey analysis have spun off of artificial speech recognition technology, but there are only a handful of algorithms that have been implemented for

INTRODUCTION

general use by wildlife biologists. In both commercial and theoretical applications the pool of algorithms includes hidden Markov models based on spectrograms (Song Scope: Wildlife Acoustics 2011 2011), mel-frequency cepstral coefficients (Lee et al. 2008), spectrogram cross-correlation (e.g. Avisoft SAS-Lab Pro Avisoft Bioacoustics e.K. 2014 2014), binary point matching of spectrograms (Towsey et al. 2012), linear predictive coding with geometric distance matching (SongID: Boucher and Jinnai 2012), and band-limited energy detection in spectrograms (Raven: Bioacoustics Research Program 2011 2011) with parametric or non-parametric clustering (Ross and Allen 2014). In addition to the named commercial software there are a handful of R packages dedicated to acoustic analysis: **seewave** (Sueur et al. 2008) has a toolbox for sound analysis, and **soundecology** (Villanueva-Rivera and Pijanowski 2014) has functions to implement the Acoustic Complexity Index (Pieretti et al. 2011), the Acoustic Diversity Index (Villanueva-Rivera et al. 2011), and the Acoustic Evenness Index (Villanueva-Rivera et al. 2011), which are indices based on the overall “soundscape” or “geophony” as opposed to the presence of songs from individual species. We developed the R package **monitoR** to detect individual species in recorded surveys, take advantage of the ability to insert detections directly into a database, and avoid interobserver variation with detection probability, all while efficiently managing input metadata and the volume of data returned.

PACKAGE GOALS

As the name **monitoR** suggests, this package is designed to accomplish more than just automated acoustic detection. We placed a strong emphasis on rapid data availability while maintaining high data throughput, both of which are valuable features of long-term monitoring programs in an adaptively managed system, and both of which require a high degree

INTRODUCTION

of integration between inputs, detector, and data storage. We identified three primary objectives for automated acoustic monitoring:

1. Automatic detection of animal vocalizations in recorded audio surveys.
2. Automatic identification of detections to species.
3. Automatic parsing of survey metadata and detections with identification to a data repository.

A template matching system fulfills the first two primary objectives simultaneously; **monitor** includes two template matching algorithms with different properties (spectrogram cross-correlation and binary point matching), only one of which is selected per survey. These two algorithms can be demonstrated to produce nearly identical detection results; cross-correlation can be used when correlation is required, and binary point matching can be used when an intuitive scoring unit is required. The third primary objective is fulfilled by inclusion of a MySQL schema from which a database can be built easily, plus a variety of pre-written queries by which the database will be populated as new detections are encountered.

Additional software objectives were identified to bolster the program objectives:

1. Easily create and manipulate templates.
2. Efficient manual verification of identifications.
3. Visual and aural spectrogram browsing.
4. Manual annotation of song events in spectrograms.

OVERVIEW OF FUNCTIONS AND TERMINOLOGY

Inclusion of these secondary objectives allows **monitoR** to be a full-featured acoustic detection package useful for monitoring programs of all extents, e.g. detection of a single species in a handful of surveys as well as detection of many species in many repeated surveys.

The **monitoR** package provides functions for two kinds of template detection: spectrogram cross correlation and binary point matching. Here we provide an overview of the detection process and its functions, and we describe the role of the acoustics database.

OVERVIEW OF FUNCTIONS AND TERMINOLOGY

Documentation for **monitoR** refers to recordings made in the field as recording files, and they become surveys after they are copied to their storage destination and renamed with a site and date code using either function **fileCopyRename** or **mp3Subsamp**. These functions also have arguments that predict renaming by subsequent file conversion (e.g. from **wac** to **wav**) or subsampling for surveys (e.g. drawing short surveys from extended MP3 recording files). Use of these functions is encouraged, as they are the only automated means to preserve absolute survey times and to collect survey metadata for subsequent upload to the user's acoustics database.

A MySQL schema for an acoustics database is available for download at the **monitoR** website (<http://www.uvm.edu/rsenr/vtcfwru/R/?Page=monitoR/monitoR.htm>; accessed 08/01/2014); the database is also available as a MySQL Workbench file which can be merged into an existing database. Connection to the database is accomplished using package **RODBC**. The schema can be loaded to an active MySQL database instance using the function **dbSchema**, and most

OVERVIEW OF FUNCTIONS AND TERMINOLOGY

data transfers have pre-written queries in the functions `dbUploadSurvey`, `dbUploadTemplate`, `dbUploadResult`, `dbDownloadSurvey`, `dbDownloadTemplate`, and `dbDownloadResult`.

Template matching is a simple process in which the template is repeatedly scored for similarity against a moving window of the survey, and a go-no-go score cutoff is applied to distinguish between acceptable and unacceptable matches. We refer to the transition threshold as the score cutoff. Examples of songs from target species are referred to as song clips, and they are made into templates with the interactive functions `makeCorTemplate` or `makeBinTemplate`. Templates are list objects that include the score cutoff used to filter detections, a name to be assigned to each detection, and a comment field to annotate the origin or effectiveness in addition to the path to the song clip and spectrogram data and signal parameters. The signal component of templates in `monitoR` are more sophisticated than a fourier-transformed song clip: both types of templates allow inclusion or rejection of regions of the spectrogram (individual syllables) in the final signal to be matched.

The functions that perform the matching, `corMatch` and `binMatch`, both produce an object of class `templateScores`. This object contains the score for each time bin in the survey and each template, plus it also contains all the templates and the spectrogram data.

The maximum score for each sound event (each “peak”) occurs when the sound event in the survey reaches maximum alignment with the template, and it is identified with the function `findPeaks`. This function also reads the score cutoff from each template and filters out peaks that fail to exceed the score cutoff, leaving only “detections”. The `templateScores` object is packaged with all peaks and detections in a `detectionList` object. There are several potential fates for `detectionList` objects: the peaks or detections can be extracted

ACOUSTIC DETECTION WORKFLOW WITH LOCAL DATA STORAGE

from this object with `getPeaks` or `getDetections` for writing to a file, they can be plotted on the survey spectrogram at once with the `plot` method, they can be individually plotted on the survey spectrogram for manual verification with `showPeaks`, or they can be uploaded to an acoustics database with `dbUploadResults`.

Spectrograms of audio files can be interactively viewed/browsed, annotated to a csv file, and played with the function `viewSpec`. Short clips from sound files can be extracted and saved from within this function for making templates. Audio playback calls the `play` function from package `tuneR` (Ligges et al. 2014), for which a player must be specified. We use the cross-platform player SoX (Bagwell et al. 2013), but the native Windows Media Player (`wmplayer.exe`) works in Windows and `afplay` works on MacOS.

ACOUSTIC DETECTION WORKFLOW WITH LOCAL DATA STORAGE

For each template matching method `monitoR` contains two core functions that are used in sequence (a `___Match` function followed by `findPeaks`), a variety of functions for creating, reading, writing, and manipulating templates, a function for manual result verification, two getter functions for extracting data frames or lists of detections, and a detection plot method. The flow-chart in Figure 2.1 represents common usage for users working from a local data repository. Recording files can be used directly as surveys if their modification date has not changed since initial recording.

ACOUSTIC DETECTION WORKFLOW WITH DATABASE STORAGE

Inclusion of the acoustics database into the workflow replaces the read/write template functions with upload/download functions, and functions to upload and download results are available as well. The functionality to pull detections from the database and return them to **detectionList** objects allows past detections to be manually verified with **showPeaks**, or plotted using the class method. When downloading results, the resulting **detectionList** object will necessarily be incomplete as it will lack the complete **templateScores** object, plus it may lack peaks below the score cutoff if only detections were originally saved to the database. Surveys remain as locally stored audio files, but a data frame of metadata is uploaded to the database to allow detections to relate to their originating survey. Figure 2.2 illustrates this workflow.

MANAGING INPUTS

SURVEYS

monitorR relies on the package **tuneR** to read wave files and decode MP3 files. Package **tuneR** offers what it describes as a “bare bones” MP3 decoder, which decodes the entire MP3 file to a **Wave** object. Compared to MP3 files, the wave format is bulkier to store but easier to work with, as specific portions of a file can be accessed quickly by specifying the start and end points to the sample, second, or minute. Both supported formats allow a

MANAGING INPUTS

maximum of two channels per recording; detection defaults to the left channel unless the right channel is actively selected in advance using the `channel` function in package `tuneR`.

The function `readMP3` in `monitoR` masks the version in `tuneR`. The version of `readMP3` in `monitoR` helps to make MP3 files nearly as easy to navigate as wave files; this function calls the third-party software `mp3splt` to extract segments of MP3 recordings without first decoding them. The program `mp3splt` and `libmp3splt` for GNU Linux, *BSD, MacOS X, BeOS, and Windows can be downloaded at http://mp3splt.sourceforge.net/mp3splt_page/home.php (accessed 08/01/2014) and must currently be installed outside of R.

We are not aware of a standard recording format among bioacoustic monitoring programs, although we are aware of proprietary compression formats in use by some commercial recorders. Without existing support in R for alternate formats (such the compressed WAC format from Wildlife Acoustics™ or alternate PCM formats such as AIFF), we anticipate that users will likely work with wave files as opposed to converted `Wave` objects.

Detections for each template are reported by both relative time (since the beginning of the file) and “absolute” time (time of day when the recording was made). Absolute times are computed from file modification times, which are coded into the survey file name when files are transferred from removable media using either `fileCopyRename` or `mp3Subsamp`. In addition to coding the file modification time into the file name, these functions gather metadata about each survey and make a copy of the survey in a target folder. While copying and renaming survey files is the primary purpose of `fileCopyRename`, the function `mp3Subsamp` will call the third-party software `mp3splt` to isolate short surveys in long MP3 recordings without the delay of decoding the long file. This function may be useful if using

MANAGING INPUTS

a recorder with limited scheduling capabilities. A user could, for example, use this function to isolate and store a 10 minute survey from each hour of an 8 hour recording. Each 10 minute MP3 survey could then be decoded in seconds when needed, rather than waiting the many minutes required to decode the full length recording only to immediately discard the majority of the audio.

TEMPLATES

As noted earlier, templates are the union of the spectrogram data and parameters, the score cutoff, and a flexible comment field. Templates are contained in **corTemplateList** or **binTemplateList** objects. Templates may not make use of the entire sound clip spectrogram, so the sound clip that the template is based on is required for plotting. A binary point matching template and correlation template are plotted in Figure 2.3. The static path to the sound clip stored in each **templateList** object must therefore remain valid if templates are to be plotted. Although there is no function to update the path in the **templateList** object, the path can be revised in the template file after it is written to disk with either **writeCorTemplate** or **writeBinTemplate**. These two functions write text files (with respective extensions *ct* and *bt*) that can be opened and edited by any text editor. Template files are read back into **monitoR** with either **readCorTemplates** or **readBinTemplates**.

Both correlation and binary point template lists are uploaded to the acoustics database with **dbUploadTemplate**. In the database, the file path can be updated in *'tblTemplate'. 'fldClipPath'*. Template lists are downloaded back to **monitoR** with the function **dbDownloadTemplate**.

MANAGING OUTPUTS

Three levels of detections are produced during the detection process: the raw scores, maximum scores for each sound event (peaks), and detections (peaks greater than the score cutoff). Depending on the sample rate of the survey and the window length for the fast Fourier transform there may be several hundred raw scores evaluated per second. For any given sound event in the survey only one score will represent the maximum score as the sound event reaches maximum alignment with the template. There are thus many fewer peaks than raw scores, and for many species there are fewer detections than peaks. It is assumed that it will not be efficient for an established monitoring program to store raw scores, but either peaks or detections may be stored locally or in the database. To write results locally use either **getPeaks** or **getDetections** to extract the results as either a data frame or list, and then write the results to a csv file or an Rda file using generic functions. To upload peaks or detections to an acoustics database use the function **dbUploadResult** and adjust the **which.one** argument to select either peaks or detections. Detections are contained in **detectionList** objects, which can be plotted as results superimposed over spectrograms (Figure 2.4).

If it is impractical to archive entire surveys, audio for detections may be extracted using either **bindEvents** or **collapseClips**. These functions can produce a text file documenting the template used to detect each event, and they can be used to archive detections derived from manual annotation as well as automated identification.

MANIPULATING TEMPLATES

In a multi-species monitoring program it is likely that more than one target species will be searched for in each survey, and for each target species it is likely that more than one template will be used for more complete detection. Duplicate results arising from multiple templates for a single target species can be eliminated using the `timeAlign` function, which will keep only the highest scoring detection for a given time window. Rather than specify each template to be matched in the matching function (`corMatch` or `binMatch`), templates are aggregated into `corTemplateList` or `binTemplateList` objects which contain one or more complete objects of class `binTemplate` or one or more objects of class `corTemplate` (but never mixing the two classes in a single `templateList`), and the entire `templateList` object is specified in the match function. `templateList` objects containing one or more templates can be combined using `combineCorTemplates` or `combineBinTemplates`.

All template manipulation is performed to `templateList` objects. Available manipulation includes accessing/replacing the template name (with function `templateNames`), the score cutoff (with function `templateCutoff`), and the template comment (with function `templateComment`). `templateList` objects can be plotted with the default plot method, and they can be written to or read from a ct or bt file using `writeCorTemplates`, `writeBinTemplates`, `readCorTemplates`, and `readBinTemplates`.

MANUAL SPECTROGRAM ANNOTATION

While the primary objective of **monitoR** is to form the scaffold for an unsupervised monitoring system, there are at least two occasions when manual annotation is required. First, example sound clips for target species must initially be manually identified in existing recordings and either saved as independent files or the start and end times noted. Second, users are likely to attempt to estimate the accuracy of their monitoring by comparing the results from **monitoR** to what is detectable in the audio surveys to a human observer. For these occasions the **viewSpec** function can be used. **viewSpec** has three modes of use: non-interactive spectrogram viewing, interactive viewing without annotation, and interactive viewing with annotation. When called to display a non-interactive spectrogram **viewSpec** will create the spectrogram and exit immediately. In interactive viewing mode the function prints a series of command options to the console. The options allow the user to page through the spectrogram, play the visible time, zoom on the time and frequency axes, and save the visible time to a wave file. In interactive and annotation mode an additional option offers the ability to click in the active graphics device to draw rectangles around song events and name each event in the console. Switching between the console and the graphics device is minimized by recycling the previous event name if a new name is not supplied. Deleting annotations is done by exiting annotation mode, entering delete mode, and drawing a rectangle around all annotations to be deleted.

After annotation of an audio file is complete a prompt to save the annotated events appears in the console and a csv file is written containing the time, frequency, and event name. If a name is not provided annotations are saved in a temporary file to prevent inadvertent data loss. The temporary file remains available until overwritten by the next call to **viewSpec** with annotation. Existing annotation files can be displayed on the survey spectrogram

MANUAL RESULTS VERIFICATION

within `viewSpec` by using the `anno` argument to point to a csv file of annotated events.

MANUAL RESULTS VERIFICATION

Results for a given template can be manually verified by sequentially plotting a spectrogram of each using the function `showPeaks`. When the `verify` argument is set to `TRUE` each peak or detection can be labeled as a true or false positive detection in the console.

An alternate method of verification is to annotate all target song events in a survey, run the template matching functions for a survey, and then compare the manually annotated events to the detected events using the function `eventEval`.

THE ACOUSTICS DATABASE

This package has the potential to extract an enormous volume of data from digital recordings. The acoustics database provided for download is designed primarily to store the observations generated from `monitoR`, but it contains tables to store environmental data as well, including data such as daily temperature or precipitation figures. The ready-to-use schema can be used as the primary data storage facility for new monitoring programs, while monitoring programs with existing databases can opt to incorporate the database into their current system using the MySQL Workbench file.

To set up the acoustics database the schema should be uploaded to an existing database

TOWARD AN AUTOMATED MONITORING FRAMEWORK

instance using the function `dbSchema`. Once the tables are created all hardware should be assigned a unique identifier, which is entered into the tables `tblCard`, `tblCardRecorder`, and `tblRecorder`. Similarly, information about the survey locations any locations at which recordings are made (such as for template creation) are stored in `tblLocation`. The table `tblSurvey` stores survey file paths and recording metadata; its relationship to `tblCardRecorder` allows each survey to be traced back to the location, recorder, and digital media card upon which it was originally recorded. The tables `tblTemplate` and `tblSpecies` store the templates and the species associated with each template, respectively. The relationship between `tblTemplate` and `tblLocation` allow the original location of the sound clip for each template to be traced. The tables `tblResult` and `tblArchive` store the output of the detection and classification process, which are detections and file paths to archived events. The remainder of the tables, `tblPerson`, `tblPersonContact`, `tblOrganization`, `tblProgram`, and `tblProject`, store programmatic data that allow device installations, templates, and results to be traced back to individual analysts and the acoustic monitoring programs with which they are affiliated. A summary of the tables, including several not described here, are listed in Table 2.1.

TOWARD AN AUTOMATED MONITORING FRAMEWORK

The tools described here offer ample opportunity to share individual observations, tables of observations, and entire templates between monitoring programs. The use of a diverse suite of templates has the potential to enhance the robustness of detection, although templates must be vetted carefully to minimize the inclusion of templates prone to excessive false-positive detections.

ACKNOWLEDGEMENTS

A complete framework for sample design remains to be proposed; ideally such a framework would specify survey duration, daily timing, and potentially annual timing. A benefit to adhering to an established framework is the ease with which shared data among participating organizations can be compared without any need for corrective measures before analysis.

ACKNOWLEDGEMENTS

Support for this project was provided by National Park Service cooperative agreement P10AC00288. We thank _____ for their critical review of this manuscript. The Vermont Cooperative Fish and Wildlife Research Unit is jointly supported by the U.S. Geological Survey, the University of Vermont, the Vermont Department of Fish and Wildlife, and the Wildlife Management Institute. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

REFERENCES

- Acevedo, M. A., and L. J. Villanueva-Rivera. 2006. Using Automated Digital Recording Systems as Effective Tools for the Monitoring of Birds and Amphibians. *Wildlife Society Bulletin* **34**:211–214.
- Avisoft Bioacoustics e.K., 2014. Avisoft-SASLab Pro version 5.2 [Computer Software]. URL <http://www.avisoft.com/>.
- Bagwell, C., R. Sykes, and P. Giard, 2013. SoX: Sound eXchange. Version 14.4.1. URL <http://sox.sourceforge.net/Main/HomePage>.
- Bart, J., and J. D. Schoultz. 1984. Reliability of Singing Bird Surveys: Changes in Observer Efficiency with Avian Density. *The Auk* **101**:307–318.
- Bioacoustics Research Program, 2011. Raven Pro: Interactive Sound Analysis Software [Computer Software]. URL <http://www.birds.cornell.edu/raven>.
- Boucher, N. J., and M. Jinnai, 2012. SoundID Documentation: A Revolutionary Sound Recognition System. URL <http://www.soundid.net>.
- Brewster, J. P., and T. R. Simons. 2009. Testing the importance of auditory detections in avian point counts. *Journal of Field Ornithology* **80**:178–182.

REFERENCES

- Farmer, R. G., M. L. Leonard, and A. G. Horn. 2012. Observer effects and avian-call-count survey quality: rare-species biases and overconfidence. *Auk* **129**:76–86.
- 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.
- Hobson, K. A., R. S. Rempel, H. Greenwood, B. Turnbull, and S. L. V. Wilgenburg. 2002. Acoustic surveys of birds using electronic recordings: new potential from an omnidirectional microphone system. *Wildlife Society Bulletin* **30**:709–720.
- Kendall, W. L., B. G. Peterjohn, and J. R. Sauer. 1996. First-time observer effects in the North American Breeding Bird Survey. *Auk* **113**:823–829.
- Kepler, C. B., and J. M. Scott. 1981. Reducing bird count variability by training observers. *Studies in Avian Biology* **6**:366–371.
- Lee, C.-H., C.-C. Han, and C.-C. Chuang. 2008. Automatic Classification of Bird Species From Their Sounds Using Two-Dimensional Cepstral Coefficients. *IEEE Transactions on audio speech and language processing* **16**:1541–1550.
- Legg, C. J., and L. Nagy. 2006. Why most conservation monitoring is, but need not be, a waste of time. *Journal of Environmental Management* **78**:194–199.
- Ligges, U., S. Krey, O. Mersmann, and S. Schnackenberg, 2014. tuneR: Analysis of music. URL <http://r-forge.r-project.org/projects/tuner/>.
- MacKenzie, D. I., J. D. Nichols, G. B. Lachman, S. Droege, J. A. Royle, and C. A. Langtimm. 2002. Estimating site occupancy rates when detection probabilities are less than one. *Ecology* **83**:2248–2255.
- Nichols, J. D., and B. K. Williams. 2006. Monitoring for conservation. *Trends in Ecology & Evolution* **21**:668–673.

REFERENCES

- Parker, T. A. 1991. On the use of tape-recorders in avifaunal surveys. *Auk* **108**:443–444.
- Pieretti, N., A. Farina, and D. Morri. 2011. A new methodology to infer the singing activity of an avian community: The Acoustic Complexity Index (ACI). *Ecological Indicators* **11**:868–873.
- Ross, J. C., and P. E. Allen. 2014. Random Forest for improved analysis efficiency in passive acoustic monitoring. *Ecological Informatics* **21**:34–39.
- Sauer, J. R., B. G. Peterjohn, and W. A. Link. 1994. Observer differences in the North-American Breeding Bird Survey. *Auk* **111**:50–62.
- Sueur, J., T. Aubin, and C. Simonis. 2008. Seewave: a free modular tool for sound analysis and synthesis. *Bioacoustics* **18**:213–226.
- Towsey, M., B. Planitz, A. Nantes, J. Wimmer, and P. Roe. 2012. A toolbox for animal call recognition. *Bioacoustics-the International Journal of Animal Sound and Its Recording* **21**:107–125.
- Villanueva-Rivera, L. J., and B. C. Pijanowski, 2014. soundecology: Soundscape ecology. R package version 1.1.1. URL <http://CRAN.R-project.org/package=soundecology>.
- Villanueva-Rivera, L. J., B. C. Pijanowski, J. Doucette, and B. Pekin. 2011. A primer of acoustic analysis for landscape ecologists. *Landscape Ecology* **26**:1233–1246.
- Wildlife Acoustics, 2011. Song Scope [Computer Software]. URL <http://www.wildlifeacoustics.com>.
- Yoccoz, N. G., J. D. Nichols, and T. Boulinier. 2001. Monitoring of biological diversity in space and time. *Trends in Ecology & Evolution* **16**:446–453.

TABLE AND FIGURE LEGEND

Table 2.1: Descriptions of tables in the supplied MySQL “acoustics” database schema, to which survey metadata and resulting detections are saved.

Figure 2.1: Workflow with inputs and results stored locally. File inputs are ovals, rectangles are functions, and hexagons are object classes.

Figure 2.2: Workflow with input metadata and results stored in the supplied MySQL “acoustics” database. File inputs are ovals, rectangles are functions, database tables are parallelograms, and hexagons are object classes.

Figure 2.3: A binary point template (a) and correlation template (b), plotted using the default `plot` method for `templateList` objects. These templates are from the same black-throated green warbler (*Setophaga virens*) song clip. Binary point templates have “on” and “off” points, which are plotted in yellow and blue respectively, while correlation templates have only included points and ignored points, and the included points are plotted in yellow. The default `plot` method superimposes the template over a spectrogram of the song clip from which it is made.

Figure 2.4: Three binary point matching detections of black-throated green warbler (tall rectangles) and two detections of ovenbird (*Seiurus aurocapillus*, short rectangles) in a 30 second spectrogram of a recorded survey. Detections are plotted using the default `plot` method for `detectionList` objects in `monitor`. A running score for each template is plotted

TABLE AND FIGURE LEGEND

beneath the template detections and spectrogram, and the dashed line indicates the score cutoff for each template. The time and score of each detection is marked here with an open circle.

TABLE AND FIGURE LEGEND

Table 2.1: Descriptions of tables in the supplied MySQL “acoustics” database schema, to which survey metadata and resulting detections are saved.

Acoustics Database Tables

TABLE NAME	REMARKS
tblAnnotations	For annotated song events in surveys.
tblArchive	For archiving sound clips extracted from surveys.
tblCard	This table stores information about memory cards.
tblCardRecorder	Track survey, recorder, and memory card links.
tblCovariate	Describe covariates and types of enviromental data collected.
tblEnvironmentalData	Non-acoustic data: environmental covariates.
tblLocation	Information about about locations for surveys and templates.
tblOrganization	Store the organization name and contact info here.
tblPerson	Names of people in the monitoring program.
tblPersonContact	Contact info, including Cell/Work Phone and email.
tblProject	Store the names of multiple projects per organization here.
tblRecorder	This table stores information about recording units.
tblResult	Table to store the results of findPeaks().
tblResultSummary	Store probability of survey presence.
tblSpecies	Store BBL codes or other 4, 6, or 8 character codes.
tblSpeciesPriors	Store site and species specific priors here.
tblSurvey	This table stores attributes of the survey recording.
tblTemplate	Store templates and template metadata.
tblTemplatePrior	Store beta parameter estimates for error rates.

TABLE AND FIGURE LEGEND

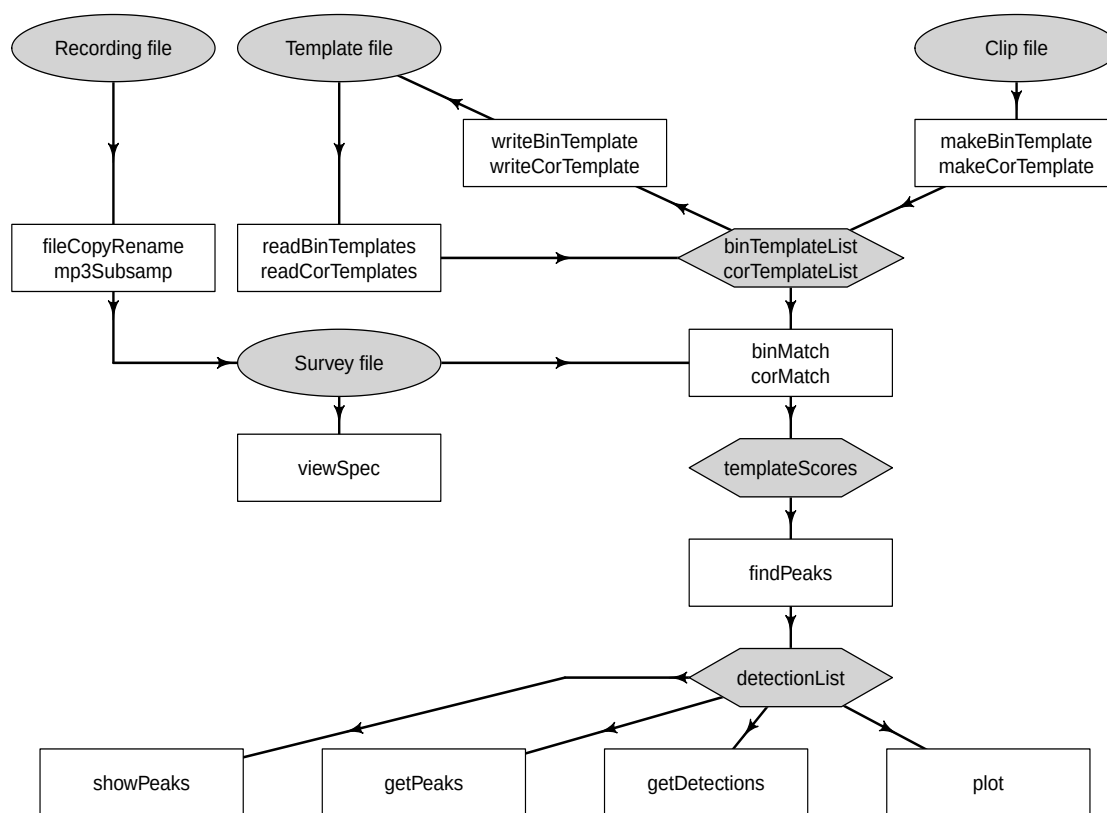


Figure 2.1

TABLE AND FIGURE LEGEND

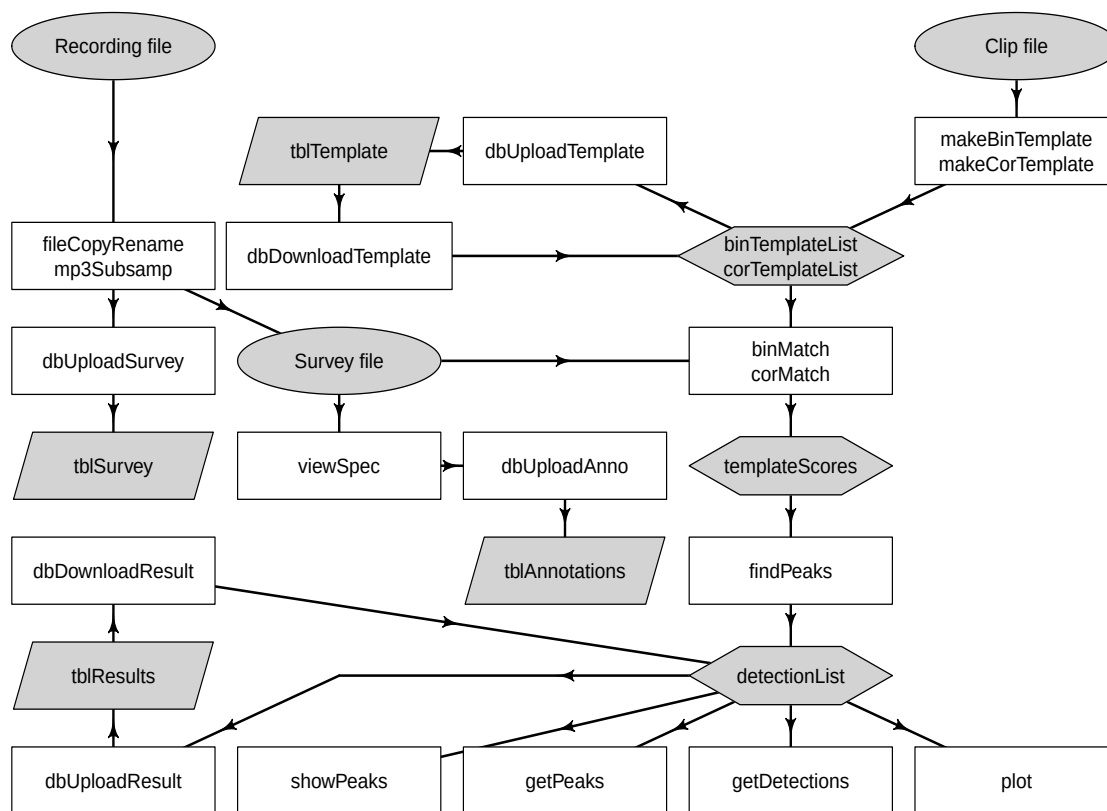


Figure 2.2

TABLE AND FIGURE LEGEND

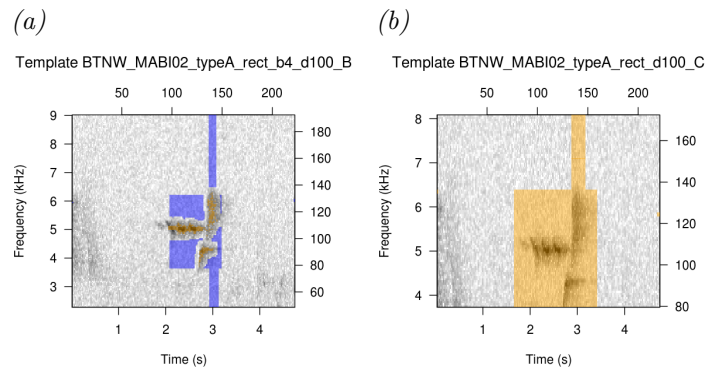


Figure 2.3

TABLE AND FIGURE LEGEND

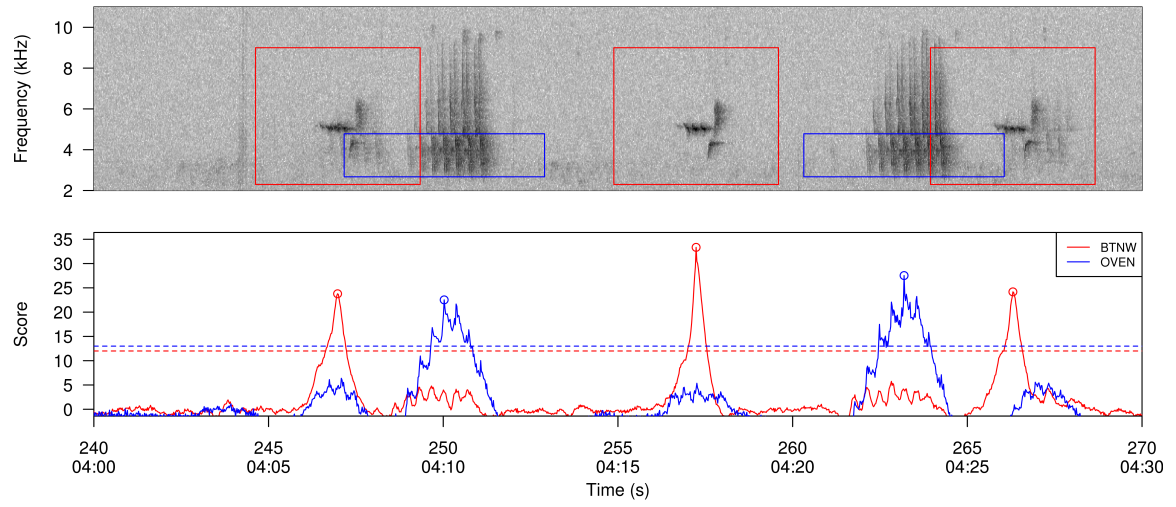


Figure 2.4

CHAPTER 3

ASSESSMENT OF ERROR RATES IN ACOUSTIC MONITORING WITH THE R PACKAGE MONI- TOR

Assessment of Error Rates in Acoustic Monitoring with the R package `monitoR`

Jonathan Katz

Vermont Cooperative Fish and Wildlife Research Unit
Rubenstein School of Environment and Natural Resources
University of Vermont, 81 Carrigan Dr, Burlington VT, 05405

Sasha Hafner

Hafner Consulting,
University of Odense
Odense, Denmark

Therese Donovan

U.S. Geological Survey
Vermont Cooperative Fish and Wildlife Research Unit
Rubenstein School of Environment and Natural Resources
University of Vermont, 81 Carrigan Dr, Burlington VT, 05405

Prepublication Disclaimer:

This draft manuscript is distributed solely for purposes of scientific peer review. Its content is deliberative and predecisional, so it must not be disclosed or released by reviewers. Because the manuscript has not yet been approved for publication by the U.S. Geological Survey (USGS), it does not represent any official USGS finding or policy.

ABSTRACT

ABSTRACT

Climate change coupled with land-use change will likely alter habitats and affect state parameters of the animal populations that dwell in them. Affected parameters are anticipated to include site occupancy and abundance, population range, and phenophase cycles (e.g., arrival dates on breeding grounds for migrant bird species). Detecting these changes will require monitoring many sites for many years, a process that is well suited for an automated system. We developed and tested *monitoR*, an R package that is designed for long-term, multi-taxa automated passive acoustic monitoring programs. We demonstrate that its results can provide reliable site occupancy data, as well as reliable counts of individual song events, by comparing the results from *monitoR* to a “silver” standard for two northeastern songbird species: black-throated green warbler and ovenbird. Unlike a “gold” standard, which contains every song event a human could detect within the surveys, the silver standard contains complete song events of a single song type with a signal:noise ratio of at least 10 dB (166 out of 439 songs for black-throated green warbler, 502 out of 990 songs for ovenbird). *monitoR* correctly identified presence for black-throated green warbler and ovenbird in 64% and 72% of the 52 surveys using binary point matching, respectively, and 73% and 72% of the 52 surveys using spectrogram cross-correlation, respectively. Of individual black-throated green warbler song events, 73% of 166 black-throated green warbler songs and 69% of 502 ovenbird songs were identified by binary point matching. Spectrogram cross correlation identified 64% of 166 black-throated green warbler songs and 64% of 502 ovenbird songs. False positive rates were <1% for song event detection.

INTRODUCTION

The fundamental objective of wildlife inventory and monitoring programs is to estimate the state of a system at one or more points in time to fulfill scientific or management objectives (Yoccoz et al. 2001, Legg and Nagy 2006, Nichols and Williams 2006). System states traditionally include the abundance of a target species or the probability of occurrence of a species, but climate change has brought species range and phenophase timing to the forefront as well (Barbraud and Weimerskirch 2006, Both et al. 2006, Lemoine et al. 2007, Devictor et al. 2008). Monitoring is also indispensable for assessing the results of management activities and modeling the effects of planned activities (Gibbs et al. 1999, Johnson et al. 2006, Runge et al. 2009, Martin et al. 2011). Whatever the system state or assessment target, there is an overt understanding that monitoring programs aim to maximize precision while minimizing bias and cost.

Wildlife surveys are usually performed with classic human-observer methods. Monitoring programs such as the Breeding Bird Survey (Sauer et al. 2003) and the North American Amphibian Monitoring Program (Weir et al. 2005) are examples of large monitoring programs that use volunteer human observers to collect data on species presence, absence, or abundance, but it is more common for monitoring to be done at a local level by paid biologists or technicians. Several shortcomings limit the strength of inference from such programs. First, budgetary constraints limit the number of field observers, which ultimately constrains the number of sites or overall area that can be monitored. Second, human observers vary in their ability to see, hear, and concentrate. This variation can result in two kinds of errors. False negative errors occur when an observer fails to detect a species that is present. False positive errors occur when an observer erroneously records the presence of a species when the species is absent. Failure to account for these errors during data analysis can adversely

INTRODUCTION

affect the accuracy of inference (Kepler and Scott 1981, Bart and Schoultz 1984, Sauer et al. 1994, Kendall et al. 1996, Farmer et al. 2012); the goal is maximize both the true positive rate (i.e., sensitivity) and the true negative rate (i.e., specificity). In general practice, extensive data validation is performed to limit additional human errors introduced during data entry, which delays analysis of the results to ensure the data are accurate. These delays may be more than inconvenient. If an adaptive management approach is used, they may shift the data-gathering phase out of sync with the decision-making phase (Gibbs et al. 1999).

One avenue for minimizing human error, bias, and cost is to adopt passive automated monitoring methods. In the case of birds or amphibians, passive bioacoustics monitoring can use audio recorders to make ambient recordings that either supplement or replace human observer surveys (Parker 1991, Haselmayer and Quinn 2000, Hobson et al. 2002, Acevedo and Villanueva-Rivera 2006). Autonomous microphones and recording devices are readily available for this purpose, and less expensive equipment are relatively simple to construct (Tierney et al. 2013). The use of autonomous field recordings for monitoring is appealing because it increases the detection rate of rare species and vocal species in locations where visibility is limited, potentially decreasing false negative errors. Moreover, equipment costs for multi-year projects and the cost of deployment and collection compare favorably to the cost of several seasons of human field visits, and the ability to schedule recordings for multiple surveys per day permits surveying diurnal, crepuscular, and nocturnal taxa with the same equipment.

As with human surveyors, there are several impediments to using audio recorders for conducting surveys. The primary challenge remains the conversion of raw recordings to a monitoring dataset. With audio recordings, data storage and access can be an issue as even a small monitoring effort can result in terabytes of storage needs. In addition to data

INTRODUCTION

storage, acoustic signals need to be matched to the target species. Converting the recorded audio files to spectrograms for visual analysis by humans has been attempted (Celis-Murillo et al. 2009, Dawson and Efford 2009), but this approach does not remove human error. An alternative method uses computer algorithms to match a signals from a field recording with a signal known to be issued by the target species (i.e., a template). A number of algorithms have been employed for this purpose, many of which have been tested with positive (Cor-topassi and Bradbury 2000, Griebmann and Naguib 2002) or mixed results (Khanna et al. 1997, Swiston and Mennill 2009).

Across all automated methods, each sound event in the survey is reduced to metrics that are scored based on the similarity to those from a template. An analyst must choose a threshold which converts each score exceeding the cutoff to a species detection, or “hit”. Analogues of this process exist in commercial speech recognition and music recognition software, but complications are introduced in wildlife monitoring such as fluctuating signal:noise ratios due to variation in ambient noise and distance to signal source. Another notable set of hurdles is that the vocalizations of some species are known to vary among regions and individuals and it is possible for two or more species issue similar vocalizations. These challenges directly affect the accuracy and precision of automated acoustic monitoring, and they introduce error which can limit inference. However, a benefit of audio recordings over human observations is that the error rates can be directly estimated by analyzing the recordings in multiple ways.

We assessed the performance of *monitoR*, an automated acoustic detection and monitoring package for R in terms of true positive, false positive, true negative, and false negative classification rates. R is open source statistical computing and graphics software available for Windows, MacOS, and UNIX-like platforms (R Core Team). We evaluated detection of two species in a series of recorded audio surveys collected at 10 sites in Vermont and

METHODS

New York, USA using both template matching algorithms that are available in *monitoR*: spectrogram cross-correlation and binary template matching.

Our objectives in this paper were to 1) set up a monitoring database to store data from all aspects of an acoustic monitoring program; 2) collect acoustic surveys with an autonomous recording unit at each of 10 sites in VT and NY and input the logistics data (to track recorder and memory card deployment) and recordings into the monitoring database, 3) develop “gold” and “silver” standard assessments of the bird vocalizations within 60 ten-minute surveys, where the gold standard documents every occurrence of the target species, and the silver standard documents all occurrences that are completely visible in the spectrogram and are of the same song type as the template; 4) develop templates of both target species in the R package *monitoR*, and 5) use *monitoR* to detect the target species, and to evaluate classification rates (true positive, false positive, true negative, false negative) of signals within the surveys with two template matching methods (spectrogram cross-correlation and binary template matching).

METHODS

DATABASE DEVELOPMENT (OBJECTIVE 1)

monitoR includes an optional MySQL database schema which can be used to track all phases of an acoustical monitoring program (Katz et al. In prep). We entered each recording device, sound card, and personnel information into the database tables, **tblRecorder**, **tblCard**, **tblPerson**, **tblPersonContact** respectively. The database assigned a unique

METHODS

identifier for each person and hardware element, which was used for tracking field operations (Objective 2).

FIELD METHODS (OBJECTIVE 2)

Monitoring stations were installed at six sites in Vermont and four sites in New York. Sites were selected near vernal pools, permanent ponds, and other waterbodies to facilitate multi-taxa monitoring (see also Brauer et al. in review). Two monitoring stations in Vermont were in large mixed northern-hardwood/balsam fir forests at approximately 480 m elevation. Four monitoring stations in Vermont were in northern-hardwood forest fragments separated by pastures at approximately 335 m elevation. Four monitoring stations in New York were at the edge of pastures set within pine-oak forest fragments at approximately 90 m elevation. All monitoring stations were selected to be aurally isolated from one another either by distance or topography. Information about each station was entered into the database table, **tblLocation**.

Each monitoring station was outfitted with a Wildlife Acoustics SM1 digital audio recorder strapped to a tree four feet above ground level. Trees no larger than 23 cm dbh in diameter were selected to promote omnidirectional recordings. Surveys were recorded within five hours after sunrise on May 28, 2010, June 11, 2010, and June 25, 2010 to coincide with typical visitation times for human surveys. We set each SM1 to record in Wildlife Acoustics' proprietary WAC0 format (lossless compression) to SDHC memory cards. The pairing of a recorder with a card was logged in the database table, **tblCardRecorder**. Cards were collected several days after each survey date and downloaded using the `monitoR` function `fileCopyRename`. All recordings were converted to stereo WAV format using the `beta`

METHODS

Wildlife Acoustics Kaleidoscope version 0.9.0 (Wildlife Acoustics 2012). Two consecutive ten-minute surveys separated by 1 minute were extracted from the recording on each survey date at each location using the wave I/O capabilities of package tuner (Ligges 2011). Ten minute surveys were intended to be a conservative length, as traditional survey duration varies widely among projects (Briggs et al. 2012, Wimmer et al. 2013).

Study Species

To test `monitoR`'s performance we evaluated detection of two bird species that breed in Vermont and New York: black-throated green warbler (*Setophaga virens*) and ovenbird (*Seiurus aurocapilla*). These focal species were selected due to their dissimilar vocalization types, frequencies, and song rates. Common species were chosen in an attempt to provide reasonable estimates of the capabilities of the detection methods. Both species were considered common in the habitats surveyed and were therefore believed to be present in many of the recorded surveys. Information about each species was entered into the database table, **tblSpecies**.

The black-throated green warbler sings two song types, "A" and "B". Only the type B song was included in this analysis for simplicity. The type B song consists of five or six notes: three or four notes at the same narrow frequency band (~ 5.0 - 5.2 kHz), followed by a lower note (~ 3.5 - 4.5 kHz) and a higher slurred buzz (~ 5.5 - 7.0 kHz); the total song length is typically 1.5 - 2.0 seconds. This song is phonetically printed zee-zee-zee-zoo-zee (Morse and Poole 2005). The waveform of the song is plotted in Figure 4.1a. Figure 4.1b illustrates the spectrogram that results from the Fourier transform, with digital bandpass between 2 kHz and 9 kHz.

METHODS

The ovenbird has one primary song type consisting of a series of broad-frequency staccato notes emitted in rapid repeat for 2.5-4.0 seconds. The song has a familiar mnemonic of loud, repetitive “teacher teacher teacher” notes. These notes are typically rendered on a spectrogram as a series of near-vertical lines. The frequencies covered by the song are often between 3 kHz and 5.5 kHz, although in higher amplitude recordings the notes of the song often blend seamlessly with harmonics up to 8 kHz (Porneluzi et al. 2011). Figure 4.1e depicts the waveform of this song, and Figure 4.1f illustrates the spectrogram from the Fourier transform with digital bandpass between 2 kHz and 12 kHz.

CREATING GOLD AND SILVER STANDARDS (OBJECTIVE 3)

The “gold” standard is a dataset which we considered to be “truth” in terms of the song events recorded in each survey, including the time of each song. Each 10 minute survey was thoroughly reviewed by J. Katz using high-quality headphones (Grado SR225i) while simultaneously browsing the spectrogram with `monitoR`’s `viewSpec` function. Identifications were made either by audio alone, spectrogram image alone, or a combination of the two methods. The recording was stopped and portions replayed as necessary to identify and annotate all instances of target species song events. Extra effort was made to identify events that were too quiet to be visible on the spectrogram but which remained audible for inclusion in the gold standard because they could potentially be identified in a human point count survey. Figure 3.2a depicts the spectrogram from a 30 second composite sound clip in which 5 instances of black-throated green warbler (songs I, II, V, VII, and IX) and 4 instances of ovenbird (songs III, IV, VI, and VIII) would be included in the gold standard.

The “silver” standard is a subset of the gold standard that differs in three ways: first,

METHODS

song events of both species that were less than 10 dB louder than ambient noise were not included (such as VI in Figure 3.2a). Second, black-throated green warbler song events of song type B but which were issued with the initial syllables at a frequency other than 5.0-5.2 kHz were not included (I in Figure 3.2a), and third, the black-throated green warbler song type A was not included (IX in Figure 3.2a). Eliminating quiet events is analogous to performing a fixed-radius point count, as quiet events are more likely to be farther from the recorder. Since we are primarily interested in quantifying detection error of *monitoR*, i.e., the difference between what is detected by *monitoR* and what we could reasonably expect to be detected given a template and an understanding of the detection algorithms, we do not consider eliminating other song types an unrealistic usage bias. In a monitoring situation we expect each song variant would be detected with a unique template, and there would be little to no overlap in which template detects each song variant.

The process of creating the gold standard data set took roughly three times longer than the real time survey durations. Ten-minute recorded surveys with many song events took as long as 40 minutes to listen to and annotate, while those with very few song events were completed in as few as 10 minutes. Either the gold or silver dataset could be used to quantify the number of songs missed or misclassified by the automated detection and classification methods to estimate the false negative and false positive rate.

TEMPLATES (OBJECTIVE 4)

Template Development. To evaluate survey presence of the two target species templates were created from song events selected from within one of the recordings. In *monitoR*, templates are created from a Fast Fourier Transform (FFT) of the recording. We used a

METHODS

512 point FFT window length with no overlap and a Hanning window function to create our templates.

MonitoR can perform two methods of template matching (binary point matching and spectrogram cross-correlation), and each requires a method-specific template. The `monitoR` functions `makeCorTemplate` and `makeBinTemplate` were used to create correlation and binary point matching templates, respectively, for each species. Binary point matching templates are based on a map of anticipated regions of signal within a spectrogram, designated on points, and regions of anticipated non-signal, designated off points; all other values are ignored. During point selection, groups of on or off points can assume regular or irregular shapes in the resulting template. An amplitude cut-off is used to quickly isolate potential on points during template creation. In contrast to the binary point template, a correlation template is based on a matrix of amplitudes, and the values for amplitudes are copied directly out of the template spectrogram. A rectangular selector tool in `monitor` allowed manual selection of the syllables within a spectrogram that are used or ignored in a template.

The binary point template for black-throated green warbler was constructed from a high signal:noise song example from within the survey recordings. A plot of the template points superimposed upon the spectrogram of the original sound clip is displayed in Figure 4.1c. The amplitude cutoff was set at ~ 26 dB to minimize the contribution from signal that has reflected off of leaves, tree trunks, and other structures in the area. A four time-bin and frequency-bin buffer was specified around all on points (85 ms and 188 Hz) to prevent most higher-amplitude cells from bleeding into the off-points. The template included the song and at least 50% of the song duration as leading and trailing ambient noise. Off-points were strategically located around the on-points to minimize false positives from broadband frequency noises such as microphone clicks and squirrel vocalizations. We constructed an

METHODS

ovenbird template in a similar manner to the black-throated green warbler template, using all of the same parameters but with an amplitude cutoff of -14 dB to limit the on points to the highest amplitude syllables (Figure 4.1g).

Correlation templates for black-throated green warbler and ovenbird were constructed using the same song clips used for binary point templates (Figure 4.1d, h). Areas within the spectrogram were chosen to minimize false positives from broadband frequency noises such as microphone clicks and squirrel vocalizations.

We treated the process of template development as an iterative one in which a template would be constructed and tested on a short recording from the same sites but outside of the survey set, and the results assessed using the `monitor` function `viewPeaks`. The final templates were uploaded to the MySQL database table, **tblTemplate**, using the function `dbUploadTemplate`.

HITS AND CLASSIFICATION RATES OF HITS (OBJECTIVE 5)

To analyze the acoustic recordings from the 10 study sites with `monitor` we used the functions `binMatch` and `corMatch`. These functions create spectrograms of each survey with the same parameters as the templates, and will recycle the spectrogram between successive templates to save time if possible. We downloaded all four templates (Figure 4.1c, d, g, h) from the database with the `monitor` function `dbDownloadTemplates`, which automatically aggregates both binary point templates into a `binTemplateList` object and both correlation templates into a `corTemplateList` object. The function `binMatch` scores each time bin in the survey as the difference in mean on point and off point amplitudes. The binary

METHODS

point matching method implemented in *monitoR* is a modification of the implementation described in Towsey et al. (2012). Matching is based on the equation:

$$score = \frac{\sum amp_{on}}{N_{on}} - \frac{\sum amp_{off}}{N_{off}} \quad (3.1)$$

in which amp_{on} and amp_{off} represent the amplitudes of spectrogram cells anticipated to be signal and noise, respectively, and N is the count of cells (see Appendix A). Scores theoretically range between 0 and the absolute-value of the noise-floor of the recording equipment, although due to the averaging in the above formula scores typically range between 0 and 30. The amplitudes compared are from the survey itself; the template only guides the selection of points for comparison. The score cutoff is a signal:noise ratio threshold below which there are no detections, thus this method is inherently biased toward higher amplitude signals.

For correlation templates, the function *corMatch* scores each frame with the Pearson's correlation coefficient. The process implemented in *monitoR* is based on Mellinger and Clark (1997). Matching is based on the equation:

$$r_{T,S} = \frac{n \sum TS_i - \sum T \sum S_i}{\sqrt{n \sum T^2 - (\sum T)^2} \sqrt{n \sum S_i^2 - (\sum S_i)^2}} \quad (3.2)$$

in which n is the number of amplitude values in the template, T are the amplitudes values in the template, S_i are the amplitude values in the survey, and $r_{T,S}$ is the Pearson correlation coefficient for the appropriate template and survey matrix (see Appendix A). This algorithm scores a known template against each potential song event in the survey in a moving frame analysis. Scores range between 0 and 1, where 0 indicates the amplitudes in the two spectrograms are uncorrelated, and 1 indicates a perfect time-frequency-amplitude

METHODS

alignment. In theory, this method will match low-amplitude and high-amplitude signals to a template equally well, but in practice matches to low amplitude signals appear more susceptible to perturbation from random noise than those to high amplitude signals.

After scoring each frame the function `findPeaks` was used to identify local maxima, or peaks, within the score vector. We used the default frame time duration which sets the frame duration equal to the template duration; increasing the frame duration ignores peaks farther apart, and decreasing the frame duration risks detecting multiple peaks generated by the same song event (e.g. the multiple peaks of song VIII in Figure 3.2b).

The resulting scores were uploaded to the database table **tblResults** with the function `dbUploadResult`. We exercised the option in this function to upload all peaks since our goal was to quantify the error rates of the process, including how many song events go undetected (false negatives).

Peaks greater than or equal to the user-defined threshold are considered species detections, or “hits”. The score cutoff adjusts the overall error rate of the analysis; lowering the score cutoff produces more positive classifications (identifications and misidentifications) but reduces false negatives (failures to classify a song). Conversely, raising the score cutoff reduces false positives but increases false negatives. Choosing a score cutoff that optimizes these error rates is a user-driven process determined by project-specific objectives. We generated hits for each survey across a range of cutoffs, from 0 to 30 for binary point matching and from 0 to 1 for correlation approaches.

We used a Dell Precision T7500 Workstation with six 2.93 GHz dual-core processors, 48GB

METHODS

RAM and running Ubuntu 12.04 “Precise Pangolin” to make templates and perform the detection and classification analysis. Using a non-Windows operating system allowed the use of the parallel computation in the template matching functions `binMatch` and `corMatch` and in `findPeaks` to improve speed performance. All detections (hits) for each template and score cutoff were verified by comparison with the silver standard dataset using the function `eventEval`. Events were assumed to be correctly detected and identified by `monitoR` if any time within one second of the detection center overlapped with the center of the event time in the silver standard dataset. The error margin of one second was chosen due to it being roughly half the duration of black-throated green warbler song, although in trial-and-error testing with the longer-duration ovenbird song one second remained the value that produced the most accurate assessment. Each event detected in the surveys was classified as either a true positive detection (in which the score for a song event matched or exceeded the score cutoff), false negative (in which either the score for a song event was less than the score cutoff or the song event was not scored at all), a false positive (in which the score for a non-song event matched or exceeded the score cutoff), or a true negative (in which the score for a non-song event was less than the cutoff).

The error rates of the results were calculated to evaluate the efficacy of each template under each score cutoff with spectrogram cross-correlation and binary point matching. The probability of a song event being classified as a false positive, false negative, true positive, or true negative was computed for each species and analysis algorithm as the rate of events of each classification per number of events total. We were interested in identifying the positive predictive value (PPV) of each template (the conditional probability of a song event being identified correctly given that `monitoR` identified it as a hit) and the false negative rate (the conditional probability that a song event is not detected given that it occurred).

RESULTS

DATABASE POPULATION (OBJECTIVE 2)

The ten sites, three survey dates, and paired surveys should have yielded 60 surveys, but eight surveys were lost due to corrupt media cards resulting in a total of 52 surveys. It is unknown if this failure rate is typical of the media card brand we purchased, of the recording hardware, if it was a product of our formatting and use methods, or if it was due to an incompatibility between the card brand and the recorder model.

GOLD AND SILVER STANDARD (OBJECTIVE 3)

A total of 439 black-throated green warbler song events were manually identified in 19 of the 52 surveys (Table 3.1; Gold Standard). Of those song events, 166 were retained in the silver standard at 11 of the 52 surveys (Table 3.1; Silver Standard). The loss of 8 surveys from the silver standard was a combination of some surveys containing only a different song type and others containing only low amplitude songs. Song events dropped from the gold standard included 51 type “A” song events, 138 type “B” song events at a different frequency (primarily the frequency shown in song I in Figure 3.2a), and 84 type “B” song events too low amplitude to produce recognizable spectrograms. The remaining song events were in surveys from four of the ten sites throughout the sampling duration. A total of 990 ovenbird song events were manually identified in 35 of the 52 surveys (Table 3.1; Gold Standard). Of those song events, 532 were retained in the silver standard at 29 of the 52 surveys. The remaining 458 song events were dropped because they were too low amplitude to produce recognizable spectrograms (such as song VI in Figure 3.2a).

RESULTS

The song events were in surveys from five of the ten sites throughout the sampling duration.

DETECTIONS AND CLASSIFICATION RATES OF DETECTIONS (OBJECTIVE 5)

The binMatch analysis for each ten-minute survey was completed in roughly 15 seconds, and the scans of all 52 surveys using both templates took 14 minutes for binary point matching and 15 minutes for correlation matching, each using the parallel processing option.

Scores for binMatch analyses of both species ranged from 0 to about 40 dB, and scores for cross correlation analysis of both species ranged from ~ 0 to near 1. Figure 3.2a, b, and c illustrate detections similar to plotting the results from *monitoR* for the binary point template (Figure 3.2b) and the correlation template (Figure 3.2c). Figure 3.2a depicts 30 seconds of a composite assembled from three survey spectrograms; song event I is from one survey, song events II:VIII are from a second survey, and song event IX is from a third survey. The nearest ovenbird to the microphone in this composite (III in Figure 3.2a) is apparently counter-singing with both a nearby (IV in Figure 3.2a) and a more distant individual (VI in Figure 3.2a; the more distant individual was actually several minutes later than the first counter singing event and may be the same counter-singer in a different location). Figure 3.2b is a plot of scores at each time bin for both binary point templates. Detections identified by *findPeaks* are marked with points; the labels correspond to the song events in Figure 3.2a. Neither binary point matching nor spectrogram cross-correlation (Figure 3.2b and c) match the black-throated green warbler song type “B” high variant (I in Figure 3.2a) nor the song type “A” (IX in Figure 3.2a), but songs which coincide with other noises are occasionally matched at the dashed-line detection level (V in Figure 3.2a is not a detection

RESULTS

while VIII is).

At least one example of a difference in detection between binary point matching and spectrogram cross-correlation is visible between these panels: song event V in Figure 3.2 has a score peak that could have been detected with the binary point template with a slight reduction in the score cutoff, but in the spectrogram cross-correlation analysis, the same song event has no distinct peak and cannot be detected without introducing many false positives. It is unknown how often similar situations arose in the surveys, but among the 52 surveys there was little overall difference in either survey presence or raw song events between the two template matching methods for both species (Table 3.1). At a score cutoff of 12, binary point matching correctly detected black-throated green warbler presence in 7 of the 11 surveys from the silver standard (64%), incorrectly identified it as present in 2 surveys (5%), and failed to identify presence at 5 surveys. Most of the false positives were in the survey from site GIFF2 on 25-Jun; these were all songs of black-throated blue warbler (*Setophaga caerulescens*). The song of the black-throated green warbler is more similar to that of the black-throated blue warbler than any other sympatric member of the *Setophaga* genus, although most humans can distinguish the two species due to the fewer number of notes and the more buzzy quality to the black-throated blue warbler song. At a score cutoff of 0.35, spectrogram cross-correlation correctly identified presence in 8 surveys (73%), incorrectly identified it as present in 3 surveys (7%), and failed to identify presence at 2 surveys. As with binary point matching, all of the false positives in the survey from site GIFF2 on 25-Jun were songs of black-throated blue warbler. The similarity between results of the two methods is expected as both templates were based on the same original song clip.

Ovenbird detection results were similar to black-throated green warbler results. At a score cutoff of 17, binary point matching correctly identified presence in 21 of the 29 surveys

RESULTS

(72%), incorrectly identified it as present in 3 surveys (13%), and failed to detect presence in 7 surveys. With a cutoff of 0.55, spectrogram cross-correlation correctly identified it as present in 22 surveys (76%), incorrectly identified it as present in 2 surveys (9%), and failed to detect it as present in 6 surveys. The majority of the false positive detections in the survey from site GIFF2 on 25-Jun were songs of common yellowthroat (*Geothlypis trichas*).

There was a small divergence between spectrogram cross-correlation and binary point matching in terms of raw number of song events detected, with the former yielding more true events for black-throated green warbler: 122 vs. 107 events out of 166 events in the silver standard (73% vs 64%). Similarly, for ovenbird, there was strong similarity between results of the two methods, although the cross-correlation method achieved both more accurate presence metrics and more true raw detections than binary point matching: 369 vs. 342 events of 532 events in the silver standard (69% vs 64%).

The results for the score cutoffs in Table 3.1 are set to maximize the sensitivity (or true positive rate, given a detection) while keeping the specificity (or true negative rate, given a miss) close to 1. Thus, for black-throated green warbler at the selected cutoffs, the specificity for binary point matching is 0.99 and the corresponding sensitivity is 0.71 (Figure 3.3a, circle and square points), and the specificity for cross-correlation is 0.99 with a corresponding sensitivity of 0.83 (Figure 3.3b, circle and square points). The ovenbird specificity for binary point matching was 0.99 with a sensitivity of 0.62 (Figure 3.3a, diamond and triangle points), and the cross-correlation specificity was 0.99 with a sensitivity of 0.65 (Figure 3.3b, diamond and triangle points).

The true positive rates shown in Table 3.1 can be increased by reducing the score cutoff, at

RESULTS

the expense of increasing the false positive rate (Figure 3.3a, b). Lowering the cutoff will decrease the specificity but will increase the sensitivity. Examining the change in sensitivity and specificity as a function of the score cutoff can illuminate which cutoff value will best balance the rates of the two types of errors (Figure 3.3a, b). For example, for ovenbird with binary point matching, a score cutoff of 5 will set the two error rates to roughly equal values. However, considering only the rates disguises the total number of events: with roughly 40 true negative events for each true positive event, the number of false positives has the potential to grow much faster than the number of true positives.

Another common method for exploring how changes in the cutoff affect the true positive rate and false positive rate is the Receiver Operating Characteristic (ROC), which plots the true positive rate as a function of the false positive rate (which is 1-specificity; Figure 3.3c). A detector that is equivalent to a random guess results in a line with a slope of 1 up the diagonal of the ROC plot; lines above the diagonal line are progressively better detectors than random, and lines below the diagonal are progressively worse detectors. According to the ROC plot, both template matching methods are very good detectors provided that the limitations are understood. In this case, the limitations are those acknowledged in the silver standard: weak signals are not expected to be detected and only signals very similar to the template will score as detections.

Yet another way of examining monitoR's efficacy is to graph the fraction of all detections that are truly song events and what fraction are errors (Figure 3.4a). At the score cutoffs chosen for Table 3.1, a PPV of over 90% means that percentage of hits represent true song events (Figure 3.4b and c, circle and square points), and a false negative rate of less than 10% means that percentage of song events were overlooked (Figure 3.4b and c, diamond and triangle points). Note that these plots are a response to a subtly different question than

DISCUSSION

the ROC plots. The ROC plots in Figure 3.3 display the proportion of the songs detected and the proportion of non-songs ignored, while the PPV plots in Figure 3.4 display the predictive value of the detector for assessing species presence.

DISCUSSION

By necessity, monitoring programs balance the need for knowledge (unbiased and precise estimates) with the time and expense of gathering data. Large-scale monitoring programs aimed at detecting shifts in species distributions or phenophase tend to minimize costs by enlisting an army of volunteer observers (e.g. the North American Breeding Bird Survey). Aside from volunteer-based efforts, there have been few occasions where the cost and effort of gathering data across vast spatial extents or ensuring data consistency in long-term monitoring was justifiable without automated systems. An automated program such as *monitoR* has the ability to process large amounts of data of consistent quality, plus a means of organizing those data to quickly recognize trends. Both of these elements are needed to effectively document changes in wildlife populations as they confront new climate patterns and increased habitat conversion pressures.

We have assessed the performance of *monitoR* with respect to true and false positive error rates, and have shown that both template matching methods are very good detectors provided that the limitations are understood. When evaluating the detection rates presented here, it is important to consider how the limitations of the method could be practically addressed in a monitoring setting. First, it is likely that the area sampled with passive acoustic monitoring was significantly less than the area that would be sampled by humans.

DISCUSSION

It is widely acknowledged that humans have a much wider hearing range than a typical microphone for sampling biodiversity, and our silver standard dataset included only those signals that were picked up clearly by the microphone. Although untested, we believe it may be possible to calculate the maximum listening radius from the output of the `monitoR` function `binMatch`. Binary point matching has an intuitive scoring system: the scores are equivalent to signal:noise ratios, and when noise is limited to microphone and amplifier self-noise, the score cutoff represents the additional amplitude required to qualify as a detection. The maximum listening radius at each survey could be calculated as follows. The Wildlife Acoustics SM1 User Manual (Wildlife Acoustics 2009) states that the microphone has a sensitivity of -35 dBV/Pa at 1 kHz ($\pm 4 \text{ dB}$), which corresponds to a value of 94 dB SPL . With the recommended default amplifier setting of $+42 \text{ dB}$ and the default 16-bit sample depth, the quietest noise the recorder and microphone are theoretically capable of responding to is -41 dB (after setting the amplifier to map a value of 0.5 Pa to full scale (-6 dB relative to 1 Pa), $-35 \text{ dB} - 6 \text{ dB} = -41 \text{ dB}$). This corresponds to a theoretical minimum recordable volume of 88 dB SPL . The nearest gain setting of $+42 \text{ dB}$ maps the microphone sensitivity value to near zero full scale ($-41 \text{ dB} + 42 \text{ dB} = 1 \text{ dB}$), and it reduces the minimum recordable volume to 46 dB SPL ($88 \text{ dB} - 42 \text{ dB} = 46 \text{ dB}$). We know that signal may be audible in the recording at 46 dB SPL , but it cannot be detected with binary point matching until it is 12 dB (i.e. the score cutoff) above the microphone and amplifier self-noise. The required signal is thus $46 \text{ dB} + 12 \text{ dB} = 58 \text{ dB SPL}$. The attenuation of sound as it travels through air (known as terrestrial sound transmission loss, TL) can be modeled using a spherical spreading equation, based on the ratio of sound intensity at one meter from the source to that at the source:

$$TL \text{ (dB)} = 20 \log_{10}(r) \tag{3.3}$$

DISCUSSION

where r in meters is distance to the one-meter reference point. For forested sites containing large trees and thick foliage, the scattering of sounds $>2\text{kHz}$ by trees must be accounted for as well. Scattering loss (SL) can be estimated using the equation:

$$SL \text{ (dB)} = 8.5 + 0.12 * d \quad (3.4)$$

where d in meters is the distance in meters sound travels through forests (Bullen and Fricke 1982). A loud forest songbird such as the winter wren (*Troglodytes hiemalis*) has a peak amplitude of about 100 dB SPL (Brackenbury 1979); to calculate at what distance this value will diminish to the minimum detectable value, we subtract the loss due to spreading and the loss due to scattering to yield a total loss of 58 dB at ~ 100 m (± 20 m depending on true microphone sensitivity). It is therefore reasonable to assume that the maximum listening radius of a recorder at a forested site with no topographical obstructions is 80-120m, considerably less than the distance many songbirds are audible to human surveyors (Wolf et al. 1995). This value may fluctuate with fluctuations in background noise, such as leaf movement in a breeze, and topographical effects may render the area non-circular.

Measuring actual microphone sensitivity would reduce the error to a useable value, and daily fluctuations in background noise can be adjusted for by setting binMatch to report the on and off point values in addition to the scores. Such information could be beneficial for studies that use microphone arrays to track the location and movements of singers on a territory (Bower and Clark 2005, Mennill et al. 2006, Efford et al. 2009), or to establish multiple-microphone survey stations to mimic to area that would be sampled by human surveys.

DISCUSSION

Second, the silver standard against which we measured success is not general practice; our goal was to test the accuracy rate of monitoR for signals that it had a probability greater than 0 of detecting. By comparing to the silver standard, our true classification rates were higher than if we had compared monitoR's performance against the gold standard. For example, for black-throated green warblers, we focused on a single song type (type B: zoo zoo zo zee). Our template would have missed this species' alternative song type (zoo zee zoo zoo zee), which would have increased the false negative rate if we had compared monitoR's hits to the gold standard. This song variation within a species - either by region, among individuals, or within the repertoire of an individual - is perhaps the most difficult limitation to address in acoustic monitoring programs.

One strategy for automated detection of species with multiple songs types or other predictable variation in songs is to develop a template for each observed each song type or variant. For presence analysis, when using a suite of templates for each species, emphasis shifts from the error rate of any single template to the error rate of the entire suite, and the suite can theoretically expand to encompass the full range of vocal patterns for a target species, including distinctive calls. Beyond simple presence data, the use of multiple templates may allow more sophisticated analysis of song rate throughout the breeding cycle, which has been demonstrated to vary with changes in breeding status for some warblers including ovenbird, Kentucky Warbler (*Geothlypis formosa*), and American redstart (*Setophaga ruticilla*) (Gibbs and Wenny 1993, Staicer et al. 2006), or analysis of regional fluctuations in song type prevalence.

Third, the results of our assessment would differ had we selected different templates for evaluation. Template development is a source of subjective decisions including which example songs to base the template on, which syllables to incorporate in the template, and what

DISCUSSION

score cutoff will be used. Although an ideal template has a high true positive classification rate and a low false-negative rate, creating a template that meets these criteria can be a challenge as the sensitivity of the detector is typically increased at the expense of some specificity. For instance, in selecting an ovenbird song that could be used as a template, the location of the “teacher” notes in the most crowded songbird frequency band, plus the overall lack of shape to the notes, were anticipated to result in low sensitivity (low true positive rate). Preliminary study of the ovenbird song suggested that the individual notes were only apparent in high amplitude recordings; in lower amplitude recordings, the spectrogram smeared, presumably as the sounds reflected off leaves and trees in the understory. As a result of these challenges, we treated template development as an iterative process, and we envision that most users of *monitoR* will adopt this practice. If patterns of failure were observed, a new template was constructed and the process was repeated. The template creation process took between 30 minutes and 2 hours to complete for each species, plus time to locate suitable sound clips upon which to base them and test them. Time for the latter process should not be underestimated, as it can be a several-fold increase to the time needed to actually create the template.

It is important to note that template matching is unlikely to be suitable for species with apparently infinite variation or those that specialize in acoustic mimicry such as brown thrasher (*Toxostoma rufum*) or northern mockingbird (*Mimus polyglottos*). It is conceivable that templates could be made to identify some signature sounds of these species, such as the tell-tale “meowing” of the gray catbird (*Dumetella carolinensis*).

Our results potentially have several implications for monitoring and research. We have demonstrated that a template can be evaluated with respect to its Positive Predictive Value (PPV), given its cutoff score. The PPV metric summarizes the per-call classification

ACKNOWLEDGEMENTS

rate. However many monitoring programs track site occupancy (presence-absence) as a state variable. Thus, for species that call or sing repeatedly in a survey, only one signal must be correctly identified to confirm presence. Estimating the probability of presence will be a function of the true number of individuals present, their calling rates, and the template and cutoffs used in analysis. In a companion paper, we demonstrate how to “aggregate” multiple hits across a survey to infer the probability of species presence at a monitoring site (Katz et al. In *prepa*). Given a site is occupied by a target species, this approach generates an increasing probability of presence as more and more “hits” are encountered. In a separate paper we show how this probability can be combined with an occupancy estimate predicted by environmental covariates to monitor populations through time (Katz et al. In *prepb*).

ACKNOWLEDGEMENTS

We thank _____ for their critical review of this manuscript. The Vermont Cooperative Fish and Wildlife Research Unit is jointly supported by the U.S. Geological Survey, the University of Vermont, the Vermont Department of Fish and Wildlife, and the Wildlife Management Institute. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

REFERENCES

- Acevedo, M. A., and L. J. Villanueva-Rivera. 2006. Using Automated Digital Recording Systems as Effective Tools for the Monitoring of Birds and Amphibians. *Wildlife Society Bulletin* **34**:211–214.
- Barbraud, C., and H. Weimerskirch. 2006. Antarctic birds breed later in response to climate change. *Proceedings of the National Academy of Sciences of the United States of America* **103**:6248–6251.
- Bart, J., and J. D. Schoultz. 1984. Reliability of Singing Bird Surveys: Changes in Observer Efficiency with Avian Density. *The Auk* **101**:307–318.
- Both, C., S. Bouwhuis, C. M. Lessells, and M. E. Visser. 2006. Climate change and population declines in a long-distance migratory bird. *Nature* **441**:81–83.
- Bower, J., and C. Clark. 2005. A field test of the accuracy of a passive acoustic location system. *Bioacoustics-The International Journal of Animal Sound and Its Recording* **15**:1–14.
- Briggs, F., B. Lakshminarayanan, L. Neal, X. Z. Fern, R. Raich, S. J. K. Hadley, A. S. Hadley, and M. G. Betts. 2012. Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach. *Journal of the Acoustical Society of America* **131**:4640–4650.

REFERENCES

- 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.
- Cortopassi, K. A., and J. W. Bradbury. 2000. The comparison of harmonically related rich sounds using spectrographic cross-correlation and principal components analysis. *Bioacoustics* **11**:89–127.
- Dawson, D. K., and M. G. Efford. 2009. Bird population density estimated from acoustic signals. *Journal of Applied Ecology* **46**:1201–1209.
- Devictor, V., R. Julliard, and F. Jiguet. 2008. Distribution of specialist and generalist species along spatial gradients of habitat disturbance and fragmentation. *Oikos* **117**:507–514.
- Efford, M. G., D. K. Dawson, and D. L. Borchers. 2009. Population density estimated from locations of individuals on a passive detector array. *Ecology* **90**:2676–2682.
- Farmer, R. G., M. L. Leonard, and A. G. Horn. 2012. Observer effects and avian-call-count survey quality: rare-species biases and overconfidence. *Auk* **129**:76–86.
- Gibbs, J. P., H. L. Snell, and C. E. Causton. 1999. Effective monitoring for adaptive wildlife management: Lessons from the Galapagos Islands. *Journal of Wildlife Management* **63**:1055–1065.
- Gibbs, J. P., and D. G. Wenny. 1993. Song Output as a Population Estimator: Effect of Male Pairing Status. *Journal of Field Ornithology* **64**:316–322.
- Grießmann, B., and M. Naguib. 2002. Song Sharing in Neighboring and Non-Neighboring Thrush Nightingales (*Luscinia luscinia*) and its Implications for Communication. *Ethology* **108**:377–387.

REFERENCES

- 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.
- Hobson, K. A., R. S. Rempel, H. Greenwood, B. Turnbull, and S. L. V. Wilgenburg. 2002. Acoustic surveys of birds using electronic recordings: new potential from an omnidirectional microphone system. *Wildlife Society Bulletin* **30**:709–720.
- Johnson, L., R. J. Camp, K. W. Brinck, and P. C. Banko. 2006. Long-Term Population Monitoring: Lessons Learned From an Endangered Passerine in Hawai'i. *Wildlife Society Bulletin* **34**:1055–1063.
- Katz, J., S. Hafner, and T. Donovan. In prep^a. Assessing survey presence with automated detection. .
- Katz, J., S. Hafner, and T. Donovan. In prep^b. Automatic detection results in an objective Bayesian format. .
- Katz, J., S. Hafner, and T. Donovan. In prep^c. Avian acoustic monitoring with the R package *monitoR* .
- Kendall, W. L., B. G. Peterjohn, and J. R. Sauer. 1996. First-time observer effects in the North American Breeding Bird Survey. *Auk* **113**:823–829.
- Kepler, C. B., and J. M. Scott. 1981. Reducing bird count variability by training observers. *Studies in Avian Biology* **6**:366–371.
- Khanna, H., S. L. L. Gaunt, and D. A. McCallum. 1997. Digital spectrographic cross-correlation: tests of sensitivity. *Bioacoustics* **7**:209–234.
- Legg, C. J., and L. Nagy. 2006. Why most conservation monitoring is, but need not be, a waste of time. *Journal of Environmental Management* **78**:194–199.

REFERENCES

- Lemoine, N., H.-C. Schaefer, and K. Boehning-Gaese. 2007. Species richness of migratory birds is influenced by global climate change. *Global Ecology and Biogeography* **16**:55–64.
- Ligges, U., 2011. tuneR: Analysis of music. URL <http://r-forge.r-project.org/projects/tuner/>.
- Martin, J., P. L. Fackler, J. D. Nichols, B. C. Lubow, M. J. Eaton, M. C. Runge, B. M. Stith, and C. A. Langtimm. 2011. Structured decision making as a proactive approach to dealing with sea level rise in Florida. *Climatic Change* **107**:185–202.
- Mennill, D. J., J. M. Burt, K. M. Fristrup, and S. L. Vehrencamp. 2006. Accuracy of an acoustic location system for monitoring the position of duetting songbirds in tropical forest. *Journal of the Acoustical Society of America* **119**:2832–2839.
- Morse, D. H., and A. F. Poole, 2005. Black-throated Green Warbler (*Setophaga virens*).
- Nichols, J. D., and B. K. Williams. 2006. Monitoring for conservation. *Trends in Ecology & Evolution* **21**:668–673.
- Parker, T. A. 1991. On the use of tape-recorders in avifaunal surveys. *Auk* **108**:443–444.
- R Core Team, 2012. R: A language and environment for statistical computing. URL <http://www.R-project.org/>.
- Runge, M. C., J. R. Sauer, M. L. Avery, B. F. Blackwell, and M. D. Koneff. 2009. Assessing Allowable Take of Migratory Birds. *Journal of Wildlife Management* **73**:556–565.
- 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.
- Sauer, J. R., B. G. Peterjohn, and W. A. Link. 1994. Observer differences in the North-American Breeding Bird Survey. *Auk* **111**:50–62.

REFERENCES

- Staicer, C. A., V. Ingalls, and T. W. Sherry. 2006. Singing behavior varies with breeding status of American Redstarts (*Setophaga ruticilla*). *Wilson Journal of Ornithology* **118**:439–451.
- Swiston, K. A., and D. J. Mennill. 2009. Comparison of manual and automated methods for identifying target sounds in audio recordings of Pileated, Pale-billed, and putative Ivory-billed woodpeckers. *Journal of Field Ornithology* **80**:42–50.
- Tierney, G., B. Mitchell, A. Miller-Rushing, J. Katz, E. Denny, C. Brauer, T. Donovan, A. Richardson, M. Toomey, A. Kozlowski, J. Weltzin, K. Gerst, E. Sharron, O. Sonnentag, C. Teshera-Sterne, and F. Dieffenbach, 2013. Phenology monitoring protocol: Northeast Temperate Network. Report Natural Resource Report NPS/NETN/NRR-2013/681. Published Report-2197242.
- Towsey, M., B. Planitz, A. Nantes, J. Wimmer, and P. Roe. 2012. A toolbox for animal call recognition. *Bioacoustics-the International Journal of Animal Sound and Its Recording* **21**:107–125.
- Weir, L. A., J. A. Royle, P. Nanjappa, and R. E. Jung. 2005. Modeling Anuran Detection and Site Occupancy on North American Amphibian Monitoring Program (NAAMP) Routes in Maryland. *Journal of Herpetology* **39**:627–639.
- Wildlife Acoustics, 2012. Kaleidoscope [Computer Software]. URL <http://www.wildlifeacoustics.com>.
- Wimmer, J., M. Towsey, P. Roe, and I. Williamson. 2013. Sampling environmental acoustic recordings to determine bird species richness. *Ecological Applications* **23**:1419–1428.
- Yoccoz, N. G., J. D. Nichols, and T. Boulinier. 2001. Monitoring of biological diversity in space and time. *Trends in Ecology & Evolution* **16**:446–453.

TABLE LEGEND

TABLE LEGEND

Table 3.1. Black-throated green warbler and ovenbird song events per survey as determined for the gold and silver standard datasets, by binary point matching (BPM) at score cutoffs of 12 and 17 for the respective species, and by spectrogram cross-correlations (SCC) at score cutoffs of 0.35 and 0.55 for the respective species. Differences between the gold and silver standards represent limitations of Fourier transform-based analysis. Differences in detections between the Silver Standard and monitor represent detection errors; these numbers vary depending on the cutoff used. False survey presence is noted in bold; false absence is in parentheses. Missing values (-) represent lost recordings due to corrupt media cards.

FIGURE LEGEND

Figure 4.1. a) Waveform of a black-throated green warbler song. b) Spectrogram of the same song. Spectrograms are matrices of time x frequency x amplitude, and they are created by Fourier transforming the vector of discrete audio samples in the digital recording. c) Binary point template for black-throated green warbler overlaid on a spectrogram of the sound clip. Orange regions are “on” points and blue regions are “off” points. Transparent regions are ignored. d) Correlation template for black-throated green warbler overlaid on a spectrogram of the sound clip. Orange regions are the irregular template matrix. Transparent region is ignored. e) Waveform of an ovenbird song. f) Spectrogram of the same song. g) Binary point template for ovenbird overlaid on a spectrogram of the sound clip. Orange regions are on points and blue regions are off points. Transparent regions are ignored. h) Correlation template for ovenbird overlaid on a spectrogram of the sound clip. Orange

FIGURE LEGEND

regions are the irregular template matrix. Transparent region is ignored.

Figure 3.2. a) Spectrogram scores by both template matching methods for 30 seconds of a composite survey. black-throated green warbler detections and scores are in thin lines and ovenbird detections and scores are in thick lines. b) scores for spectrogram cross-correlation, with detections marked by points. Times corresponding to song events labeled in (a) are in numerals. Score cutoffs are dotted lines of matching weight to the respective species. Refer to Figure 4.1c and Figure 4.1g for images of the templates. This composite survey is composed of song events from three surveys: event I is from survey one, events II:VIII are from survey two, and event IX is from survey three. Event I is the same black-throated green warbler song type that the template was designed around, but is 600 Hz higher than the template song and is not detected (compare to frequencies of event II). Event II is a good match for the template is and detected. Event III is a loud ovenbird song and is detected. This ovenbird is counter-singing with a nearby individual that is also well represented in the spectrogram (IV), but is not detected, likely due to interference from event III. Event V is a black-throated green warbler song that was not detected because of interference from event IV. Event VI is a distant ovenbird that was not detected. Events VII and VIII are a case where both the ovenbird and the black-throated green warbler were detected despite mutual interference. Event IX represents the second song type of black-throated green warbler that is not detected. The silver standard was created by dropping events I, VII, and IX from the gold standard because they stand a near-zero chance of being detected using automated methods.

Figure 3.3. a) Sensitivity and specificity curves for binary point matching at varying score cutoffs. Sensitivity is the proportion of true positives out of all song events; specificity is the proportion of true negatives out of all non-song events. b) Sensitivity and specificity

FIGURE LEGEND

curves for spectrogram cross-correlation at varying score cutoffs. c) ROC analysis for each species and each template matching method. The true positive rate is proportion of true positives in all song events; the false positive rate is proportion of false positives in all non-song events. A perfect detector would reach into the upper-left corner, and a random guess would plot a line up the diagonal. The upper point identifies the performance at a black-throated green warbler score cutoff of 12 for binary point matching (bpm) and 0.35 for spectrogram cross-correlation. The lower point identifies the performance at an ovenbird score cutoff of 13 for binary point matching and 0.55 for spectrogram cross-correlation.

Figure 3.4. a) Each song event may be detected or missed depending on how it scores; likewise, each detected event may either be a song event or not a song event. The four possible outcomes are defined in this matrix. The lines in plots b and c represent the joint probabilities of the columns in the matrix. b) Joint probability of a black-throated green warbler (BTNW; heavy lines) and an ovenbird (OVEN; light lines) hit being true given that it scored high enough to be a hit with binary point matching (positive predictive value; solid lines), and the joint probability of a song event being missed (false negative rate; dotted lines). c) Same plot as b for spectrogram cross-correlation.

FIGURE LEGEND

Table 3.1: Black-throated green warbler and ovenbird song events per survey as determined for the gold and silver standard datasets.

Site	Date: Survey ID:	black-throated green warbler						ovenbird					
		28-May		11-Jun		25-Jun		28-May		11-Jun		25-Jun	
		A	B	A	B	A	B	A	B	A	B	A	B
GIFF1	Gold Std	0	0	0	0	0	0	0	0	0	0	1	5
	Silver Std	0	0	0	0	0	0	0	0	0	0	0	0
	BPM	0	0	0	0	0	0	0	0	0	0	0	0
	SCC	0	0	0	0	0	0	0	0	0	0	0	0
GIFF2	Gold Std	31	64	0	0	0	0	0	0	0	0	0	0
	Silver Std	4	52	0	0	0	0	0	0	0	0	0	0
	BPM	(0)	18	0	0	7	1	0	0	0	0	1	0
	SCC	1	30	0	0	5	1	0	0	0	0	13	8
MABI1	Gold Std	0	6	0	0	38	33	59	83	32	36	61	26
	Silver Std	0	0	0	0	1	1	24	50	17	13	30	22
	BPM	0	0	0	0	(0)	(0)	14	2	(0)	(0)	10	9
	SCC	0	0	0	0	(0)	(0)	14	4	(0)	(0)	20	19
MABI2	Gold Std	25	69	21	18	22	37	61	48	19	18	22	51
	Silver Std	0	1	19	18	20	37	25	21	5	7	14	40
	BPM	0	(0)	14	17	21	33	20	18	5	6	6	31
	SCC	0	1	19	16	21	30	22	19	5	6	9	36
MABI3	Gold Std	13	12	14	15	8	2	14	9	3	4	27	24
	Silver Std	3	1	0	0	0	0	6	2	0	0	7	8
	BPM	3	1	0	0	0	0	1	(0)	0	0	(0)	(0)
	SCC	3	1	0	0	0	0	2	(0)	0	0	(0)	(0)
MABI4	Gold Std	0	0	11	1	0	0	49	52	22	8	41	46
	Silver Std	0	0	0	0	0	0	34	37	8	6	18	27
	BPM	0	0	0	0	0	0	27	36	3	(0)	9	26
	SCC	0	0	0	0	0	0	29	36	2	(0)	10	26
SARA1	Gold Std	0	0	0	0	0	0	31	30	30	6	17	20
	Silver Std	0	0	0	0	0	0	31	30	30	4	6	10
	BPM	0	0	0	0	0	0	31	30	28	2	(0)	(0)
	SCC	0	0	0	0	0	0	33	29	29	4	4	1
SARA2	Gold Std	–	–	0	0	–	–	–	–	8	27	–	–
	Silver Std	–	–	0	0	–	–	–	–	0	0	–	–
	BPM	–	–	0	0	–	–	–	–	0	0	–	–
	SCC	–	–	1	0	–	–	–	–	0	0	–	–
SARA3	Gold Std	0	0	0	0	0	0	0	0	0	0	0	0
	Silver Std	0	0	0	0	0	0	0	0	0	0	0	0
	BPM	0	0	0	0	0	0	0	0	1	0	0	0
	SCC	0	0	0	0	0	0	0	0	0	0	0	0
SARA4	Gold Std	–	–	0	0	–	–	–	–	0	0	–	–
	Silver Std	–	–	0	0	–	–	–	–	0	0	–	–
	BPM	–	–	0	0	–	–	–	–	0	0	–	–
	SCC	–	–	0	0	–	–	–	–	0	0	–	–

FIGURE LEGEND

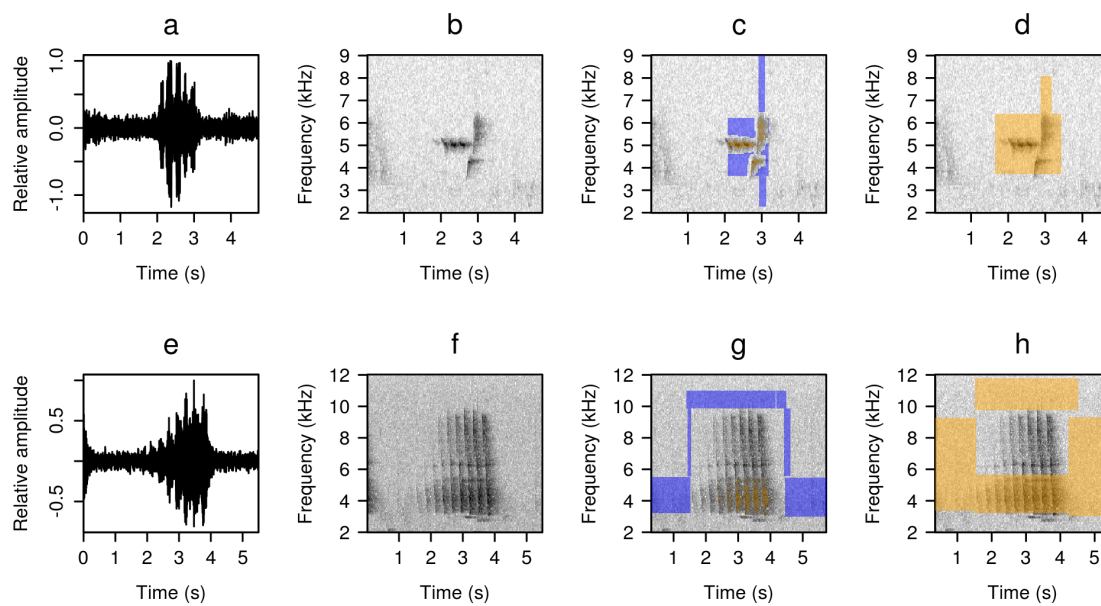


Figure 3.1

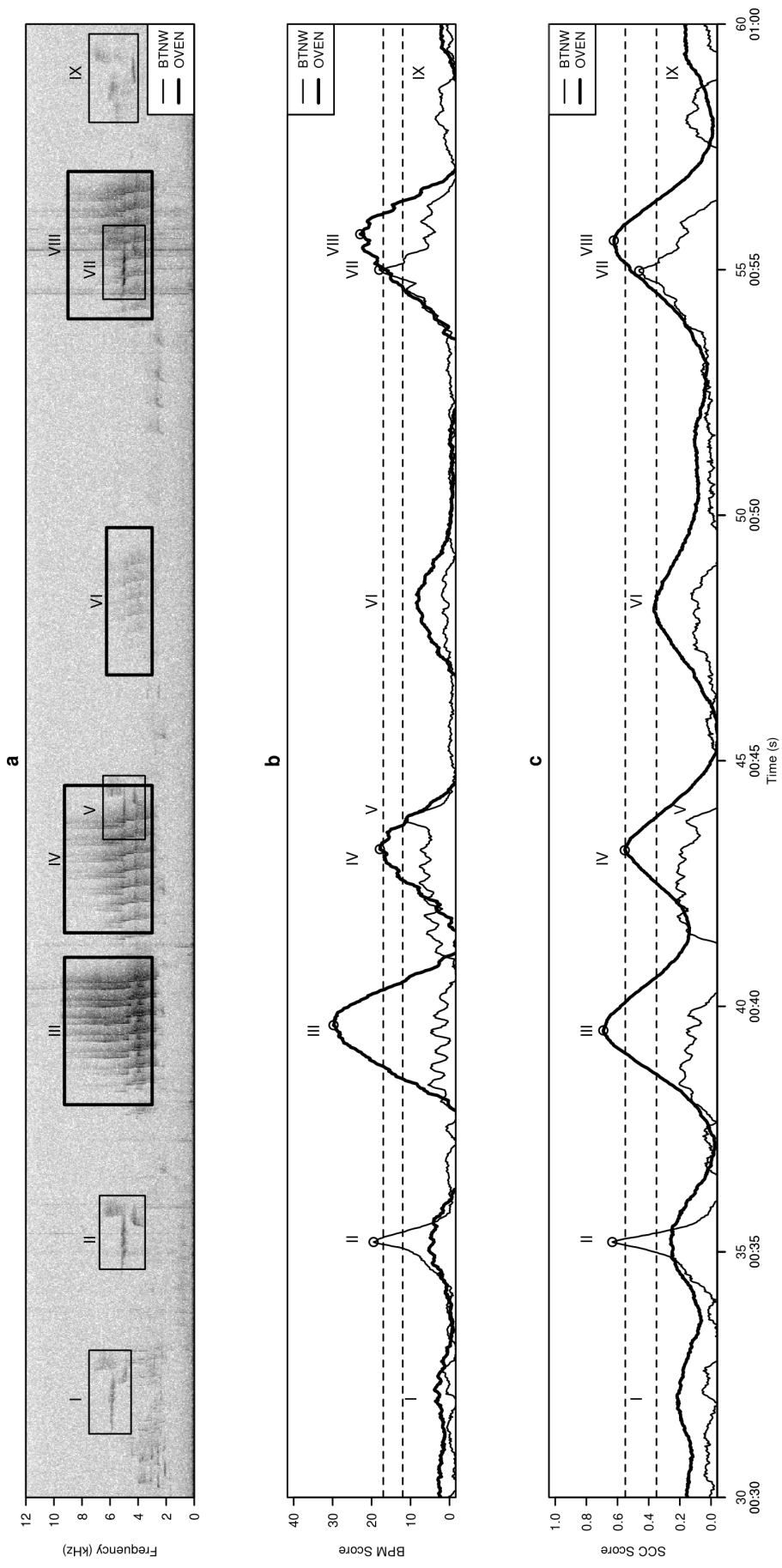


Figure 3.2

FIGURE LEGEND

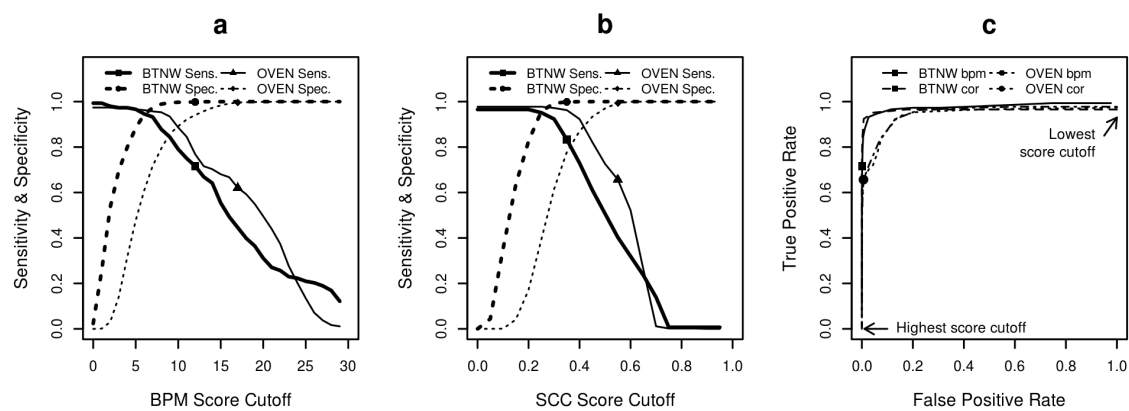


Figure 3.3

FIGURE LEGEND

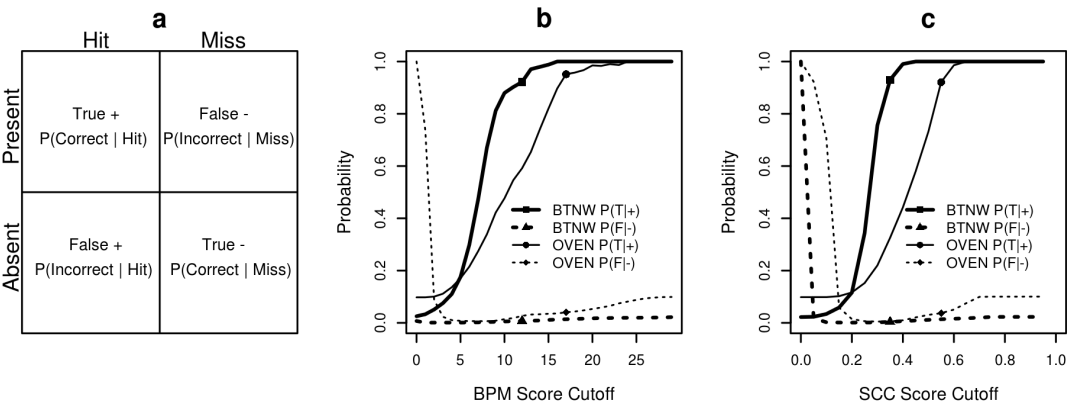


Figure 3.4

CHAPTER 4

ESTIMATING PRESENCE WITH CORRELATED AUTOMATIC DETECTION RESULTS

Estimating survey presence with automatic detection results

Jonathan Katz

Vermont Cooperative Fish and Wildlife Research Unit
Rubenstein School of Environment and Natural Resources
University of Vermont, Burlington VT

Sasha Hafner

Hafner Consulting,
University of Odense
Odense, Denmark

Therese Donovan

U.S. Geological Survey
Vermont Cooperative Fish and Wildlife Research Unit
Rubenstein School of Environment and Natural Resources
University of Vermont, Burlington VT

Ruth Mickey

College of Engineering and Mathematical Sciences
University of Vermont, Burlington VT

Prepublication Disclaimer:

This draft manuscript is distributed solely for purposes of scientific peer review. Its content is deliberative and predecisional, so it must not be disclosed or released by reviewers. Because the manuscript has not yet been approved for publication by the U.S. Geological Survey (USGS), it does not represent any official USGS finding or policy.

ABSTRACT

ABSTRACT

We describe a method to identify the probability of survey presence in a template-based automated detection system using known false positive rates for each template. True and false positive detection rates were observed in 146 training surveys. These probabilities were used in a Bayesian approach that discriminates between detections in occupied surveys and unoccupied surveys. We evaluated this approach in 146 test surveys. A total of 1142 Black-throated green warbler (*Setophaga virens*) songs were observed in the training surveys and test surveys, which we attempted to locate with 3 different binary point matching templates. When only posterior probabilities greater than 0.5 were considered detections, the average ratio of accurate identifications of survey presence to false positive identifications in 500 bootstrapped samples improved from 1.2:1 using a standard score cutoff approach to 2.8:1 using all 3 templates and a likelihood-based discriminator. With the selected score cutoffs the average true positive and false positive rates for the combined three templates were 0.18 and 0.002, respectively. In addition to offering the ability to reduce the false positive rate, this method uses a prior probability of presence which can potentially add another discriminating factor to the analysis.

INTRODUCTION

Wildlife inventory and monitoring programs typically attempt to estimate the state of a system at one or more points in time to fulfill scientific or management objectives (Yoccoz et al. 2001, Legg and Nagy 2006, Nichols and Williams 2006). A common system state of interest is species presence at a study site, which is traditionally determined via repeat surveillance by one or more human observers. Monitoring programs such as the Breeding Bird Survey (Sauer et al. 2003) and the North American Amphibian Monitoring Program (Weir et al. 2005) are examples of large monitoring programs that use volunteer human observers to collect data on species presence, absence, or abundance.

Inferring presence-absence for species from survey results demands acceptance of a variety of assumptions, including that the species was correctly identified (Royle and Link 2006). A handful of studies suggest that this assumption is commonly violated, with the degree of violation varying by observer (Kepler and Scott 1981, Bart and Schoultz 1984, Sauer et al. 1994, Kendall et al. 1996, Farmer et al. 2012). The probability of committing a false-positive error (i.e., recording the presence of a species when it is in fact absent) is often discounted, and human observers continue to dominate data collection for monitoring and inventory due to a) their perceived maximal precision in comparison to a range of automated alternatives, b) the widespread use of analytical methods that correct for false-negative errors (MacKenzie et al. 2002), and c) the difficulty of estimating actual false positive rates in a monitoring program.

Potentially viable alternatives to human observers for monitoring animal populations are maturing. Most alternatives begin with passive monitoring methods which consist of an au-

INTRODUCTION

tonomous microphone and recorder to collect surveys (Parker 1991, Haselmayer and Quinn 2000, Hobson et al. 2002, Acevedo and Villanueva-Rivera 2006, Katz et al. In prep^b). Although recorded surveys can be scrutinized by humans to ascertain presence or absence of a species, some monitoring programs use specialized software to search for acoustic signals from a target species according to temporal and frequency metrics. Given information about a target species’ song, such as a song “template”, a variety of matching algorithms can be used to scan a field recording including hidden Markov models based on spectrograms (e.g. Song Scope: Wildlife Acoustics 2011), mel-frequency cepstral coefficients (Lee et al. 2008), spectrogram cross-correlation (e.g. Avisoft SAS-Lab Pro Avisoft Bioacoustics e.K. 2014), binary point matching of spectrograms (Towsey et al. 2012, Katz et al. In prep^b), linear predictive coding with geometric distance matching (e.g. SongID: Boucher and Jin-nai 2012), and band-limited energy detection in spectrograms (e.g. Raven: Bioacoustics Research Program 2011) with parametric or non-parametric clustering (e.g. Ross and Allen 2014).

Regardless of the algorithm used, the output is a series of scores per survey which measure how close a set of unknown signals matches a vocalization known to be issued from a target species (i.e. the template). The user then must set a threshold value in which those signals whose score exceeds the threshold represent a “hit”, or species detection. This process can result in four types of classifications. When the signal is below the threshold, the conclusion is that the signal is not from the target species, which could be true or false. When the signal is above the threshold, the conclusion is that the signal is from the target species, which could also be true or false. Determining the optimal threshold to use is important because it influences the proportional relationship between all four detection outcomes (true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN)). Indeed, automated acoustic monitoring methods are typically dismissed as inefficient monitoring

INTRODUCTION

tools due to higher error rates (false negative and false positive rates) relative to human observers. However, the error rates of automated methods are usually expressed as a per-song event rate, while for human observers it is expressed as a per-survey rate.

Moreover, computer matching algorithms are potentially handicapped because they make assessments in the absence of the spatial and temporal cues that humans readily use when conducting field surveys. For example, a field observer can use habitat cues such as plant composition, plant structure, and elevation, in combination with temporal cues such as time of day or time of year, to help distinguish between similar sounding species. For template matching algorithms, a “hit” may carry more weight if the *a-priori* probability of presence is high (i.e., the habitat and temporal conditions are suitable for species presence), but carry less weight if the *a-priori* probability of presence is low.

When using acoustic monitoring as the primary survey method for monitoring programs, a constant stream of “hits” and “non-hits” (or detections and non-detections) constitute the survey’s output, where each hit and non-hit may be true or false. A major challenge involves drawing inferences regarding presence or absence of the target species from a stream of detections and non-detections in a survey, given that each signal may be erroneous (MacKenzie et al. 2002, Miller et al. 2011). A second challenge centers around the fact that species often issue multiple, distinct songs or calls, and thus a method is needed for aggregating results from multiple streams of detections produced by multiple templates aimed at detecting different vocalizations. Yet a third challenge is incorporating *a-priori* information regarding the probability of presence for the target species.

In this paper, we outline a process of inferring survey-level presence-absence of a singing

INTRODUCTION

target species, black-throated green warbler (*Setophaga virens*), from automated acoustic detections within recorded surveys. The process involves setting a prior probability of presence for the target species, and manually estimating the proportion of the four possible outcomes (TP, FN, TN, FP) of sounds evaluated in a subset of surveys. Bayes' Theorem is then used to update the prior probability based on the observed hits and non-hits throughout the survey, given the template's detection rates (true positive and false positive rates, or TPR and FPR). We analyze this with a multi-template approach that combines the results of several templates. The data were recorded at 10 sites in Vermont and New York, USA. For all analyses we use *monitoR* (Katz et al. In prep^b), a sound analysis and monitoring package for the R computing software (R Core Team).

Our objectives were to 1) collect acoustic data and manually identify all song events for the target species in the recordings, 2) use *monitoR*'s automatic detection functions to locate song events for three alternative templates and combine them into a single detection result per survey, 3) evaluate *monitoR*'s false positive and true positive classification rates (FPR, TPR) for each template and the combined approach at a variety of score cutoffs, 4) use objective priors to estimate the posterior probability of presence, given both the number of automated detections identified in the survey and the combined template detection rates, and 5) compare inferences about survey presence between the Bayesian method with standard cutoff approaches.

METHODS

OBJECTIVE 1: FIELD RECORDINGS AND MANUAL IDENTIFICATION OF SONG EVENTS

We placed autonomous recording units (Wildlife Acoustics SM1) at 10 locations across Vermont and Eastern New York in the northeast USA during the spring and summer of 2010. Recorders were located near permanent and ephemeral wetlands to offer the opportunity for nocturnal frog and toad monitoring as well (see Brauer et al. In Revision). A total of 292 10 minute surveys were recorded from the 10 sites. Field recordings were timed to coincide with those that would have been made by a hypothetical human observer, as they were collected in the five hours after sunrise. Metadata pertaining to each survey, site, recorder, and digital media card was stored in the default MySQL database provided with the R package *monitoR*. All recordings were made in Wildlife Acoustic's proprietary WAC0 file format; files were converted to WAV format using the beta version of Wildlife Acoustics' software Kaleidoscope (Wildlife Acoustics 2012) and analyzed manually and with *monitoR* in WAV format.

All song events for black-throated green warbler were manually located in the spectrograms of the 292 surveys via a combination of visual and aural inspection using the *monitoR* function `viewSpec`. After initially identifying the time of each song event within the survey, each event was reviewed a second time for identification accuracy. The timestamp of confirmed detections on each recording was stored in the MySQL database.

METHODS

OBJECTIVE 2: AUTOMATIC SONG DETECTION WITH MONITOR

The `monitoR` function `makeBinTemplate` was used to create several binary point matching templates that could be used to automatically locate black throated green warbler songs within a recording. The black-throated green warbler sings two songs types, commonly known as type A and type B (Morse and Poole 2005). During the manual verification process (Objective 1) song type B was issued at a number of frequencies; songs with a middle frequency of either 4.8 kHz, 5.0 kHz, and 5.8 kHz were most common. Separate binary point matching templates were made for song type B issued at each of the higher frequencies, and a single binary point matching template was made for song type A (Figure 4.1). A template for song type B issued at 4.8 kHz was not made for lack of a high quality recording upon which to base it.

The `monitoR` function `binBatch` scanned each survey with the three templates; this function calls the matching function `binMatch` and the filtering function `findPeaks` to identify signals within a survey. The results of spectrogram cross correlation and binary point matching were previously demonstrated to be similar (Katz et al. In prep^b), so only binary point matching was used for detection in this paper. The primary advantage of binary point matching over spectrogram cross correlation is that the score for binary point matching has units that are easily interpreted (dB difference between the on points and the off points), whereas spectrogram cross correlation uses a more abstract correlation coefficient. For example, a binary point matching score of 10 indicates that mean signal (on points) amplitude is 10dB greater than mean background noise (off points) amplitude. This method potentially limits the radius that songs will be detected for species with quiet vocalizations, but the simplicity of the algorithm permits rationalization as to how ambient noise such as wind through leaves may directly affect detection.

METHODS

All sounds scored in each survey by the template matching function (“peaks”) were stored in our acoustics database so that multiple score cutoffs could be applied *a-posteriori*. The peaks for each survey were subsequently filtered for “hits” at score cutoffs between 0dB and 15dB in increments of 1dB. At a score cutoff of 10dB, for example, all peak scores ≥ 10 dB were considered to be a species hit - and could be true or false positive - while scores < 10 dB were considered non-detections, and could be true or false negative.

To illustrate how hits and non-hits are distinguished using template matching in `monitoR` we plotted results from a single survey at site MABI02 on June 6, 2010. The scores from `binMatch` with the three templates in Figure 4.1 are illustrated in Figure 4.2a, b, and c. In this example, we used a score cutoff of 12dB for the song type A and song type B at 5 kHz templates, and 10.5dB for the song type B at 5.8 kHz template. Matching the template for song type B (5kHz) against this 600 second survey resulted in 116 evaluations (peaks), of which 12 surpassed the score cutoff and were logged as hits. Matching the template for song type A against the survey resulted in 124 peaks, of which 22 surpassed the score cutoff and were logged as hits. Matching the template for song type B (5.8kHz) against the survey resulted in 131 peaks, of which 8 surpassed the score cutoff and are logged as hits.

In the combined template analysis, scores for each peak were temporally aligned with the `monitoR` function `timeAlign`, and only the highest score among any overlapping scores was retained. In this way, redundant scores for each sound were removed to minimize double-counting of hits and sounds evaluated. To illustrate how hits and non-hits are tallied when using multiple templates we plotted the same results from Figure 4.2a, b, and c after all hits and sounds evaluated were filtered through the `timeAlign` function (Figure 4.2d). After

METHODS

filtering out duplicates, 244 sounds were evaluated by all three templates, of which 42 surpassed the score cutoff. The `timeAlign` function therefore removed 127 duplicated sounds.

OBJECTIVE 3: ESTIMATING DETECTION RATES IN TRAINING SURVEYS

We estimated TPR and FPR detection rates for the individual templates and the combined templates at score cutoffs between 1 and 15. We expected the highest TPR to be observed with the combined templates, so further analysis with individual templates was not pursued once we were able to demonstrate the TPR and FPR.

To estimate TPR and FPR detection rates for the combined template approach we used a bootstrap approach. A total of 500 bootstrap trials were evaluated; in each bootstrap trial, half of the surveys were randomly selected to be “training” surveys, by which the true positive and false positive detection rates were assessed, and the remainder were designated “test” surveys, in which the posterior probability of target species presence would be evaluated (Objective 4). To control for bias stemming from unequal proportion of surveys occupied between training and test surveys, the surveys were divided into two strata before random assignment to test or training categories: those with verified presence, and those without verified presence. An equal number of surveys was selected from each strata to ensure that the proportion of surveys occupied remained equal between training and test groups.

For each bootstrap trial, we used the `monitoR` function `eventEval` to classify each hit in

METHODS

the 146 training surveys for a given score cutoff as either TP, FN, TN, or FP detections. This function compares the time of each manually annotated song to the time of each automatically detected hit to evaluate in which category each song detection or non-detection belongs.

For each of the three individual templates plus the combined approach, the TPR was calculated as the proportion of correctly identified hits to the total number of songs by the target species, and the FPR was calculated as the proportion of incorrectly identified hits to the total number of sounds not of the target species evaluated by the template matching algorithm.

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

This process was repeated with score cutoffs ranging from 0 to 15, in increments of 0.5. Receiver operating characteristic curves were created for each template by plotting the TPR against the FPR for the full range of score cutoffs.

METHODS

OBJECTIVE 4: ESTIMATE POSTERIOR PROBABILITY OF PRESENCE USING OBJECTIVE PRIORS.

For each bootstrap trial, we estimated the posterior probability of presence in each test survey for the combined template. Hits and the number of sounds evaluated in the survey (n) were tallied for each template (Figure 4.2a-c). The combined template result was calculated as the sum of the number of hits and sounds scored by `monitoR` for all three templates, aligned with the `monitoR` function `timeAlign`, and retaining only the highest score (Figure 4.2d).

In each of the 146 test surveys, the likelihood of observing the test survey results, given the species was present ($L(data|P)$) was calculated as the binomial probability of the observed number of hits out of the total number of sounds evaluated in the test survey (n), with the probability of success set to the TPR as estimated by the training surveys.

$$L(data|P) = \text{Binomial}(n, TPR)$$

Similarly, the likelihood of observing the test survey results, given the species was absent ($L(data|A)$) was calculated as the binomial probability of the observed number of hits out of the total number of sounds evaluated in the survey (n), with the probability of success set to the FPR as estimated by the training surveys.

$$L(data|A) = \text{Binomial}(n, FPR)$$

Our process required a prior value for ψ , the probability of presence, to compute the poste-

METHODS

rior probability of presence conditioned on the observed data. One minus ψ represents the probability of absence. The source of the prior ψ may be derived from occupancy models capable of predicting ψ at each recorder, from expert opinion, or it may be derived from prior monitoring. In any case, ψ may vary by location, which will allow the most informative method of distinguishing surveys with from those without songs of the target species. In the absence of a useful prior occupancy value, ψ may be set at the maximum uncertainty value: a value of $\psi = 0.50$ offers equal weight to the possibility that either occupancy state is possible.

We used an objective prior ($\psi = 0.5$) to update the probability of occupancy for each survey given monitoR's results. Given the prior, Bayes' Theorem was used to evaluate the alternative hypotheses (presence, P , versus absence, A), conditional on the observed data from monitoR.

$$Posterior(P) \propto \frac{Prior(P) * L(data|P)}{Prior(P) * L(data|P) + Prior(A) * L(data|A)}$$

OBJECTIVE 5. COMPARE INFERENCE ABOUT SURVEY PRESENCE BETWEEN THE BAYESIAN METHOD AND A STANDARD CUTOFF APPROACH.

Because monitoring programs will most likely focus on surveys where hits have been recorded, we evaluated TP and FP records for the combined template approach. For test surveys with detections, results are summarized as the number of surveys that contain songs (i.e. true positives) of the black-throated green warbler relative to the number of surveys that do not contain songs (false positives). These odds are similar to estimates of positive predictive

RESULTS

value because they are conditional on detection. Like positive predictive value, the observed odds imply a fixed proportion of surveys occupied, which we accomplished by stratifying the training and test survey groups. The reported odds thus describe the outcomes of our methods relative to one another, but caution must be exercised before attempting to extrapolate outside of our test as the differing proportion of surveys occupied is likely to render a summary such as this incomparable. For each bootstrap sample that included detections we used the combined template results and computed the ratio of TP surveys to FP surveys.

RESULTS

OBJECTIVE 1: FIELD RECORDINGS AND MANUAL IDENTIFICATION OF SONG EVENTS

We obtained a total of 48.66 hours of recordings across study sites. A total of 1142 black-throated green warbler songs were manually identified and annotated in 64 of the 292 surveys. Songs were not evenly distributed among surveys; the number of songs per 10 minute survey ranged from a low of 1 to a high of 69, with a mean of 18 and a standard deviation 14.5.

OBJECTIVE 2: AUTOMATIC SONG DETECTION WITH MONITOR

Three binary point templates were created for the black-throated green warbler (Figure 4.1). Score cutoffs of 10-12 are most appropriate for typical templates (Katz et al. In prep^b); at

RESULTS

a score cutoff of 12dB for the type A and type B 5.0 kHz templates and 10.5dB for the type B 5.8 kHz, in one iteration of our 500 trial bootstrap analysis the total number of surveys with hits was greater for the combined template analysis than with any single template—37 with all 3 templates versus 21 with the template for song type B (5.8kHz), 11 with the template for song type A, and 11 for song type B (5kHz). Multiple templates had hits in the same survey on 10 occasions.

OBJECTIVE 3: ESTIMATING MONITOR’S DETECTION RATES IN TRAINING SURVEYS

For each of the 500 bootstrap trials we set aside a stratified random sample of 146 10-minute training surveys to estimate the TPR and FPR of each template. The remaining 146 10-minute test surveys were evaluated in Objective 4. Of the 292 surveys analyzed, 64 contained songs of the black-throated green warbler and 228 did not. These two groups were divided equally between training surveys and test surveys. An average of 473 songs were in the 32 training surveys with confirmed songs (SE 56.8), and an average of 466 were in the 32 test surveys with confirmed songs (SE 56.8).

We evaluated TPR and FPR at score cutoffs between 0 and 15. For each template and score cutoff we compared the manually verified records against the test survey results, and counted the number of TP, FP, TN and FN detections. These in turn were used to estimate the TPR and FPR for each template and cutoff combination, plotted as a receiver operating characteristic curve (Figure 4.3a). As the score cutoff increased from 0 to 15, the FPR decreased while the TPR increased. The FPR was close to 0 at score cutoffs above 10 for all templates, while the TPR was between 0.05 and 0.20 depending on the template.

RESULTS

When all three templates were combined, the TPR and FPR were generally higher at a given score cutoff than any of the component templates. In terms of an absolute count of true positive detections, the number of true positives decreased as the score cutoff increased (Figure 4.3b). At score cutoffs less than 10, the negative slope of the line indicates that template matching is sensitive to changes in score cutoff, and as the score cutoff increases from 0 to 10 an increasing number of true song detections are not scored as hits. At score cutoffs greater than 10, the shallower negative slope indicates that there is little change in the total number of true detections when score cutoff increases, suggesting that cutoffs greater than 10 do not unnecessarily increase the number of false negatives. The absolute number of false positives was very high for cutoffs less than 10 (Figure 4.3c), but dropped to near 0 between a score cutoff of 12 and 13 for most templates.

OBJECTIVE 4: OBJECTIVE PRIORS TO ESTIMATE THE POSTERIOR PROBABILITY OF PRESENCE

The prior probability of presence in a survey was set to 0.5 for all test surveys. Given the results from Objective 3, we used a score cutoff of 12dB for the song type A and song type B at 5 kHz templates, and a score cutoff of 10.5dB for the song type B at 5.8 kHz template. We then calculated the likelihood of observing the survey data under the hypothesis that the species was absent, and the likelihood of observing the survey data under the hypothesis that the species was present. For each survey, we combined the results of the three templates and updated the prior probability using Bayes' Theorem, conditioning on the total number of hits in the survey from `monitoR` and the detection rates (TPR and FPR) of the combined template from Objective 3.

RESULTS

OBJECTIVE 5. COMPARE INFERENCE ABOUT SURVEY PRESENCE BETWEEN THE BAYESIAN METHOD AND A STANDARD CUTOFF APPROACH

When considering only those surveys where detections were made, the posterior probability of presence differed between surveys where the black throated green warblers were known to be singing compared to surveys where the species was known to not be singing. An example from a single bootstrap trial is shown in Figure 4.4, with the results from the individual templates displayed alongside the combined templates for comparison. Regardless of whether the posterior was from individual template results (Figure 4.4a-c) or the combined results of all templates (Figure 4.4d), most of the estimated posteriors fell near the extremes of 0 or 1. These results, however, varied by template and between individual templates and the combined approach. Differences in the distribution of posteriors computed from data by the three templates is due to the differing TPR and FPR among the templates and differences in how many songs each template detected. The template for song type B (5.8kHz) detected songs in more surveys than either of the other two templates, possibly due to its lower score cutoff, while the template for song type A ultimately produced low probabilities of species presence for the majority of its true hits.

The prior probability of presence for all test surveys was set at 0.5. For the song type B at 5 kHz template (Figure 4.4a), with a score cutoff of 12dB, hits were logged in 15 of the 146 surveys, of which 11 were in surveys with songs confirmed to be absent and 4 were in surveys with songs confirmed to be present. Of the 11 absent surveys, 9 had a posterior probability of presence near 0, 1 had a posterior probability of presence of 0.4, and 1 had a posterior probability of presence near 1. Of the 4 present surveys, 2 resulted in a high

RESULTS

posterior probability of presence (near 1), and 2 resulted in a low posterior probability of presence (near 0). Hits were not logged in any of the remaining 131 surveys, which included 103 surveys without confirmed songs and 28 surveys with confirmed songs.

For the same bootstrap trial, the song type A template logged hits in 14 surveys, which included 11 surveys with songs confirmed to be present and 3 surveys where songs were confirmed to be absent (Figure 4.4b). Of the 11 surveys with hits and confirmed songs, only 2 resulted in high posterior probabilities of presence, which is consistent with this template's relatively low TPR (Figure 4.3a). For the 3 surveys without confirmed songs, only 1 resulted in a high posterior probability of presence. The remaining 132 survey logged no hits with this template, including 111 with no confirmed songs and 21 surveys with confirmed songs.

For the same bootstrap trial, the song type B at 5.8 kHz template logged hits in 21 surveys, including 9 in which songs were confirmed to be present and 13 in which songs were confirmed to be absent (Figure 4.4c). Of these, 5 surveys with songs confirmed to be present and 2 with songs confirmed to be absent had high posterior probabilities of species presence.

For the same bootstrap trial, the combined template approach had hits in 37 surveys, including 19 in surveys in which songs were confirmed to be present and 18 in which songs were confirmed to be absent (Figure 4.3d). Of these, 6 surveys with songs confirmed to be present and 1 with songs confirmed to be absent had high posterior probabilities of species presence.

For this bootstrap trial, the odds of one or more hits identifying the actual occupancy state was roughly 1:1 using a standard score cutoff approach. That is, of the 36 hits, for each

DISCUSSION

true positive detection made a false detection was made. If posteriors less than 0.5 are not considered high enough to classify as detections, the Bayesian updating method improved these odds to 6:1 (for each 6 true detections made, 1 false detection was made).

Resampling the surveys 500 times for training and test group membership indicates that the difference between the ratio of TP:FP surveys in the score cutoff approach and the Bayesian approach in this trial was greater than the average observed in the bootstrap analysis (Figure 4.5). Applying only the score cutoff to filter detections produced an average of 1.2 true positive surveys for each false positive survey in the 146 test group ($sd = 0.24$). Combining the results from the three templates yielded odds of 2.8:1 ($sd = 1.7$). Overwhelming odds ($TP > 0$ & $FP = 0$) were observed in 14 iterations of the combined method. The odds values reported here depend on the number of occupied surveys in the training and test groups and therefore can not be applied outside of this example; however, the consistent increase in the odds relative to the score cutoff alone suggests that a second opportunity to discriminate between occupied surveys and unoccupied surveys measurably improves detection accuracy.

DISCUSSION

We have demonstrated a method for estimating or updating occupancy rates at a site when only a single survey is performed. Our approach entailed estimating the TPR and FPR for each template, counting the number of detections made by each template in a survey, and using a binomial distribution to estimate the conditional probability of the survey being occupied relative to the probability that it is unoccupied. This is done for all templates of the same species, and the probability that all detections are false positive for each tem-

DISCUSSION

plate is combined with any prior knowledge of the actual occupancy state. This approach moves automated monitoring from a per-signal error rate to something that natural resource managers are familiar with: occupancy probability. The occupancy model based on repeat surveys (MacKenzie et al. 2002, Miller et al. 2011) is robust to variable detection probabilities, and the most accurate estimates of ψ are likely to be produced when the fewest false positive results are introduced. This method achieves that goal by introducing the opportunity to employ a second cutoff stage to separate the surveys with detections and a high posterior probability of presence from those with detections but with a low posterior probability of presence.

The reduction of false positives comes with an increase in false negatives, or reduced sensitivity. In a monitoring program, low sensitivity is less of a problem than low specificity for three reasons. First, the previously mentioned occupancy models are sensitive to false positives but robust to false negatives. Second, monitoring programs using autonomous recorders and automated detection software can often compensate for lower detection probability to some degree by increasing the survey frequency. For species where there is some expectation of occupancy ($\psi > 0$), more surveys yields an increased probability that the species will be detected if it is present (i.e. higher sampling power). Third, monitoring programs using automatic detection software to determine occupancy, or phenological attributes such as arrival date, will likely attempt to manually review detections to establish a firm date of first detection. Checking spurious data wastes resources, but as long as some true positive detections are made monitoring can achieve its stated goals. Reducing the number of false positives, while retaining some true positives, improves the overall efficiency of these monitoring programs without substantially diminishing the results.

The TPR values observed for each individual template in this example were low, likely

DISCUSSION

because the black-throated green warbler sings two song types with slight variation in each type. The 3 templates used represented 3 unique components of the total songs, and we know there were song variants our templates could not detect. For example, the template for song type B at 5 kHz generally did not detect song type A and *vice-versa*, so we did not expect any template to have a TPR value near 1.0. By definition, the TPR reflects both the number of songs each template detected relative to the number of songs available to detect, which should be proportional to the total number of songs that match the type and variant of the song used to make the template.

This Bayesian approach can be expanded and improved upon in several ways. First, the number of true positive, false positive, true negative, and false negative detections, and hence the true positive and false positive rates, could be modeled via multinomial logistic regression as a function of daily covariates. The result would be unique detection rates that are dependent on environmental or temporal variables. Covariates such as wind speed, time of day, time of year, and precipitation have rational effects on detection accuracy and can be easily incorporated into a multinomial logistic regression framework. Binary point matching is a particularly useful detection algorithm in this scenario as the mechanism that links the chosen covariates to the predicted outcome is more visible. In the forest, for example, wind increases leaf noise and decreases the signal:noise ratio, thus potentially reducing the total number of detections (Gonzalo-Turpin et al. 2008). Traditional human survey protocols would call for surveillance to abort if the wind speed was too great; however, automated surveys scheduled in advance with no prior consideration for the weather may be susceptible to spurious detections. On days with high winds, analysis of a survey likely won't produce usable data, but when applying this method we could employ an additional failsafe against producing poor quality data in poor survey conditions. The poor survey conditions will increase the false positive rate for the template-survey combination, which

DISCUSSION

the framework described here uses to unweight the contribution of any single detection in that survey. As a result more detections are required to raise the probability of presence. The effect of a variety of covariates has been well established in other studies: temperature is speculated to influence song rate (Hoileitner et al. 1993), repertoire composition, (Mennill 2014), and song structure (Aronov and Fee 2012) in various other species, all of which will affect the relative contributions from each template used for a species and potentially the false positive rate of each template. Season and time of day may affect detection rates for various species, either jointly or independently (Harms and Dinsmore 2014, Lehmicke et al. 2013), and would also be likely candidates for inclusion as covariates.

Second, we used a maximum uncertainty estimate of the current occupancy state probability. Our prior of 0.5 allotted equal weight to the possibility that the site was either occupied or unoccupied at the time the survey was recorded. If some prior information of actual occupancy state could be obtained and used, the prior could potentially discriminate between similar song structures produced by different species that are present only in specific environments. Swamp sparrow (*Melospiza georgiana*) and dark-eyed junco (*Junco hyemalis*), for example, both sing songs with similar structure that may confuse a template matching algorithm. However, during breeding season the two species are unlikely to occupy the same habitat as the swamp sparrow breeds in wet shrubland and the junco breeds in montane forests.

Third, detection rates for most taxa will improve with multiple templates, as multiple templates can detect multiple song variants. High true positive and low false positive rates are indicative of strong templates, which suggests that individual templates - as well as template lists - may be ranked by these detection rates to objectively optimize detection with regards to sensitivity and specificity. Some care must be exercised in this process due to the

DISCUSSION

manner in which the true positive and false positive rates for different templates combine. For example, if two templates have identical true positive rates and each produces a high score for a distinct song type, the count of true detections - and thus the true positive rate - will be higher when they are used together than if either template is used individually. If detections from the two templates are not independent and they produce high scores for the same song type, the count of true detections - and thus the true positive rate - will not increase when the templates are used together. Detection rates for individual templates may thus be compared to those for groups of templates, but they can not be compared against those for individual templates used as part of group.

Fourth, the observed detection rates were specific to the region in which the recordings were made. Monitoring programs that span large geographic areas may use different templates to detect the same species, or perhaps all templates would be used at all locations with region-specific detection rates. The detection rates we employed here were for each individual song. There are several ways to scale these rates up to the survey level. In Katz et al. (In *prepa*) we demonstrate that how these parameters can be estimated at the survey level from observed data using the multiple state occupancy estimator described by Miller et al. (2011).

Lastly, because the posterior probability of occupancy increases with the number of detections, the song rate (number of songs per minute) for a target species will directly affect the posterior estimate for a fixed survey duration. This method could be extended to determine optimal survey length for a given species. Species with distinct songs that can be consistently identified by *monitoR* could be surveyed with short recordings, as could species that either sing frequently or occur at high densities in the sample area. Conversely, longer survey durations may be required to survey for species with high false positive rates. A number of studies with singing warblers have demonstrated that song rate varies with

ACKNOWLEDGEMENTS

breeding status (Gibbs and Wenny 1993, Staicer et al. 2006, Ritchison 1995), so it is possible that the optimum survey length may change throughout the year.

Species that have diverse repertoires, such as song sparrows (*Melospiza melodia*) and the wood thrush (*Catharus mustelina*), would require more templates than those with only a few song types. Subsequently a large number of song examples and training surveys may be necessary to accurately estimate detection rates for templates of the less common song types. In general, template matching may be most successful for species whose songs have multiple defined syllables, for whom identification is not based on subjective evaluations (e.g. buzziness), and whose repertoire does not include the songs of other species (e.g. mimics).

ACKNOWLEDGEMENTS

We thank _____ for their critical review of this manuscript. The Vermont Cooperative Fish and Wildlife Research Unit is jointly supported by the U.S. Geological Survey, the University of Vermont, the Vermont Department of Fish and Wildlife, and the Wildlife Management Institute. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

REFERENCES

- Acevedo, M. A., and L. J. Villanueva-Rivera. 2006. Using Automated Digital Recording Systems as Effective Tools for the Monitoring of Birds and Amphibians. *Wildlife Society Bulletin* **34**:211–214.
- Aronov, D., and M. S. Fee. 2012. Natural Changes in Brain Temperature Underlie Variations in Song Tempo during a Mating Behavior. *Plos One* **7**.
- Avisoft Bioacoustics e.K., 2014. Avisoft-SASLab Pro version 5.2 [Computer Software]. URL <http://www.avisoft.com/>.
- Bart, J., and J. D. Schoultz. 1984. Reliability of Singing Bird Surveys: Changes in Observer Efficiency with Avian Density. *The Auk* **101**:307–318.
- Bioacoustics Research Program, 2011. Raven Pro: Interactive Sound Analysis Software [Computer Software]. URL <http://www.birds.cornell.edu/raven>.
- Boucher, N. J., and M. Jinnai, 2012. SoundID Documentation: A Revolutionary Sound Recognition System. URL <http://www.soundid.net>.
- Brauer, C., T. M. Donovan, R. M. Mickey, J. Katz, and B. R. Mitchell. In Revision. A Comparison of Acoustic Monitoring Methods for Common Anurans of the Northeastern United States. *Wildlife Society Bulletin* **Submitted 25 Feb. 2014**.

REFERENCES

- Farmer, R. G., M. L. Leonard, and A. G. Horn. 2012. Observer effects and avian-call-count survey quality: rare-species biases and overconfidence. *Auk* **129**:76–86.
- Gibbs, J. P., and D. G. Wenny. 1993. Song Output as a Population Estimator: Effect of Male Pairing Status. *Journal of Field Ornithology* **64**:316–322.
- Gonzalo-Turpin, H., C. Sirami, L. Brotons, L. Gonzalo, and J. L. Martin. 2008. Teasing out biological effects and sampling artifacts when using occupancy rate in monitoring programs. *Journal of Field Ornithology* **79**:159–169.
- Harms, T. M., and S. J. Dinsmore. 2014. Influence of Season and Time of Day on Marsh Bird Detections. *Wilson Journal of Ornithology* **126**:30–38.
- 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.
- Hobson, K. A., R. S. Rempel, H. Greenwood, B. Turnbull, and S. L. V. Wilgenburg. 2002. Acoustic surveys of birds using electronic recordings: new potential from an omnidirectional microphone system. *Wildlife Society Bulletin* **30**:709–720.
- Hoileitner, M., H. Nechtelberger, and J. Dittami. 1993. The Relationship Between Individual-Differences in Male Song Frequency and Parental Care in Blackcaps. *Behaviour* **126**:1–12.
- Katz, J., S. Hafner, and T. Donovan. In prep^a. Automatic detection results in an objective Bayesian format. .
- Katz, J., S. Hafner, and T. Donovan. In prep^b. Avian acoustic monitoring with the R package `monitoR` .
- Kendall, W. L., B. G. Peterjohn, and J. R. Sauer. 1996. First-time observer effects in the North American Breeding Bird Survey. *Auk* **113**:823–829.

REFERENCES

- Kepler, C. B., and J. M. Scott. 1981. Reducing bird count variability by training observers. *Studies in Avian Biology* **6**:366–371.
- Lee, C.-H., C.-C. Han, and C.-C. Chuang. 2008. Automatic Classification of Bird Species From Their Sounds Using Two-Dimensional Cepstral Coefficients. *IEEE Transactions on audio speech and language processing* **16**:1541–1550.
- Legg, C. J., and L. Nagy. 2006. Why most conservation monitoring is, but need not be, a waste of time. *Journal of Environmental Management* **78**:194–199.
- Lehmicke, A. J. J., J. L. Bowman, A. E. Banning, and B. L. Vasilas. 2013. Effect of Tide Level and Time of Day on Detection Rate and Abundance of Clapper Rails (*Rallus longirostris*) in a Mid-Atlantic Tidal Marsh System. *Waterbirds* **36**:364–368.
- MacKenzie, D. I., J. D. Nichols, G. B. Lachman, S. Droege, J. A. Royle, and C. A. Langtimm. 2002. Estimating site occupancy rates when detection probabilities are less than one. *Ecology* **83**:2248–2255.
- Mennill, D. J. 2014. Variation in the Vocal Behavior of Common Loons (*Gavia immer*): Insights from Landscape-level Recordings. *Waterbirds* **37**:26–36.
- Miller, D. A., J. D. Nichols, B. T. McClintock, E. H. C. Grant, L. L. Bailey, and L. A. Weir. 2011. Improving occupancy estimation when two types of observational error occur: non-detection and species misidentification. *Ecology* **92**:422–428.
- Morse, D. H., and A. F. Poole, 2005. Black-throated Green Warbler (*Setophaga virens*).
- Nichols, J. D., and B. K. Williams. 2006. Monitoring for conservation. *Trends in Ecology & Evolution* **21**:668–673.
- Parker, T. A. 1991. On the use of tape-recorders in avifaunal surveys. *Auk* **108**:443–444.

REFERENCES

- R Core Team, 2012. R: A language and environment for statistical computing. URL <http://www.R-project.org/>.
- Ritchison, G. 1995. Characteristics, Use and Possible Functions of the Perch Songs and Chatter Calls of Male Common Yellowthroats. *Condor* **97**:27–38.
- Ross, J. C., and P. E. Allen. 2014. Random Forest for improved analysis efficiency in passive acoustic monitoring. *Ecological Informatics* **21**:34–39.
- Royle, J. A., and W. A. Link. 2006. Generalized site occupancy models allowing for false positive and false negative errors. *Ecology* **87**:835–841.
- 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.
- Sauer, J. R., B. G. Peterjohn, and W. A. Link. 1994. Observer differences in the North-American Breeding Bird Survey. *Auk* **111**:50–62.
- Staicer, C. A., V. Ingalls, and T. W. Sherry. 2006. Singing behavior varies with breeding status of American Redstarts (*Setophaga ruticilla*). *Wilson Journal of Ornithology* **118**:439–451.
- Towsey, M., B. Planitz, A. Nantes, J. Wimmer, and P. Roe. 2012. A toolbox for animal call recognition. *Bioacoustics-the International Journal of Animal Sound and Its Recording* **21**:107–125.
- Weir, L. A., J. A. Royle, P. Nanjappa, and R. E. Jung. 2005. Modeling Anuran Detection and Site Occupancy on North American Amphibian Monitoring Program (NAAMP) Routes in Maryland. *Journal of Herpetology* **39**:627–639.
- Wildlife Acoustics, 2011. Song Scope [Computer Software]. URL <http://www.wildlifeacoustics.com>.

REFERENCES

- Wildlife Acoustics, 2012. Kaleidoscope [Computer Software]. URL <http://www.wildlifeacoustics.com>.
- Yoccoz, N. G., J. D. Nichols, and T. Boulinier. 2001. Monitoring of biological diversity in space and time. *Trends in Ecology & Evolution* **16**:446–453.

FIGURE LEGEND

FIGURE LEGEND

Figure 4.1. Black-throated green warbler templates (blue and orange) used for presence assessment, as plotted over the grayscale spectrograms. Spectrograms visually display the amplitude of various frequencies at many discrete time bins; note the different primary Y-axis scales between the three plots. Secondary X and Y axes display the dimensions of the underlying amplitude matrix.

Figure 4.2. Detection results for a 10 minute survey from MABI02 on June 6, 2010. Panels a, b, and c illustrate results from matching each of the 3 templates against the same survey. Panel d illustrates the results after combining results from the three templates.

Figure 4.3. Receiver Operating Characteristic (a) and total number of true hits (b) and false hits (c) for black-throated green warbler detection by all three templates in training surveys. At score cutoffs less than 10, the number of false positives exceeds the number of true positives in panel b.

Figure 4.4. Distribution of posteriors estimated with an objective prior of 0.5 for three templates. Detection was with a template for song type B 5kHz (a; 11 surveys with hits), song type A (b; 11 surveys with hits), and song type B 5.8kHz (c; 21 surveys with hits). The distribution of all three templates combined are in panel d (37 surveys with hits).

Figure 4.5. Distribution of TP:FP surveys in test group after 500 resampling iterations. The score cutoff only method produced a mean of 1.2:1 TP:FP, while applying a threshold

FIGURE LEGEND

of 0.5 to the posterior probability of presence increased the odds of a TP to 2.6:1.

FIGURE LEGEND

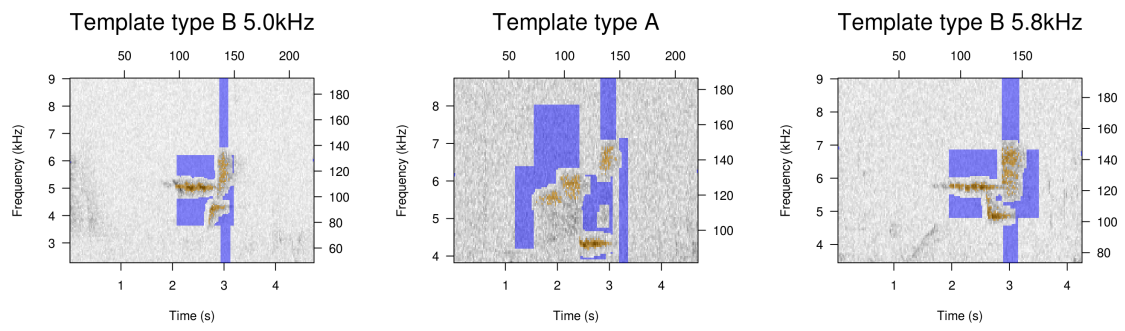


Figure 4.1

FIGURE LEGEND

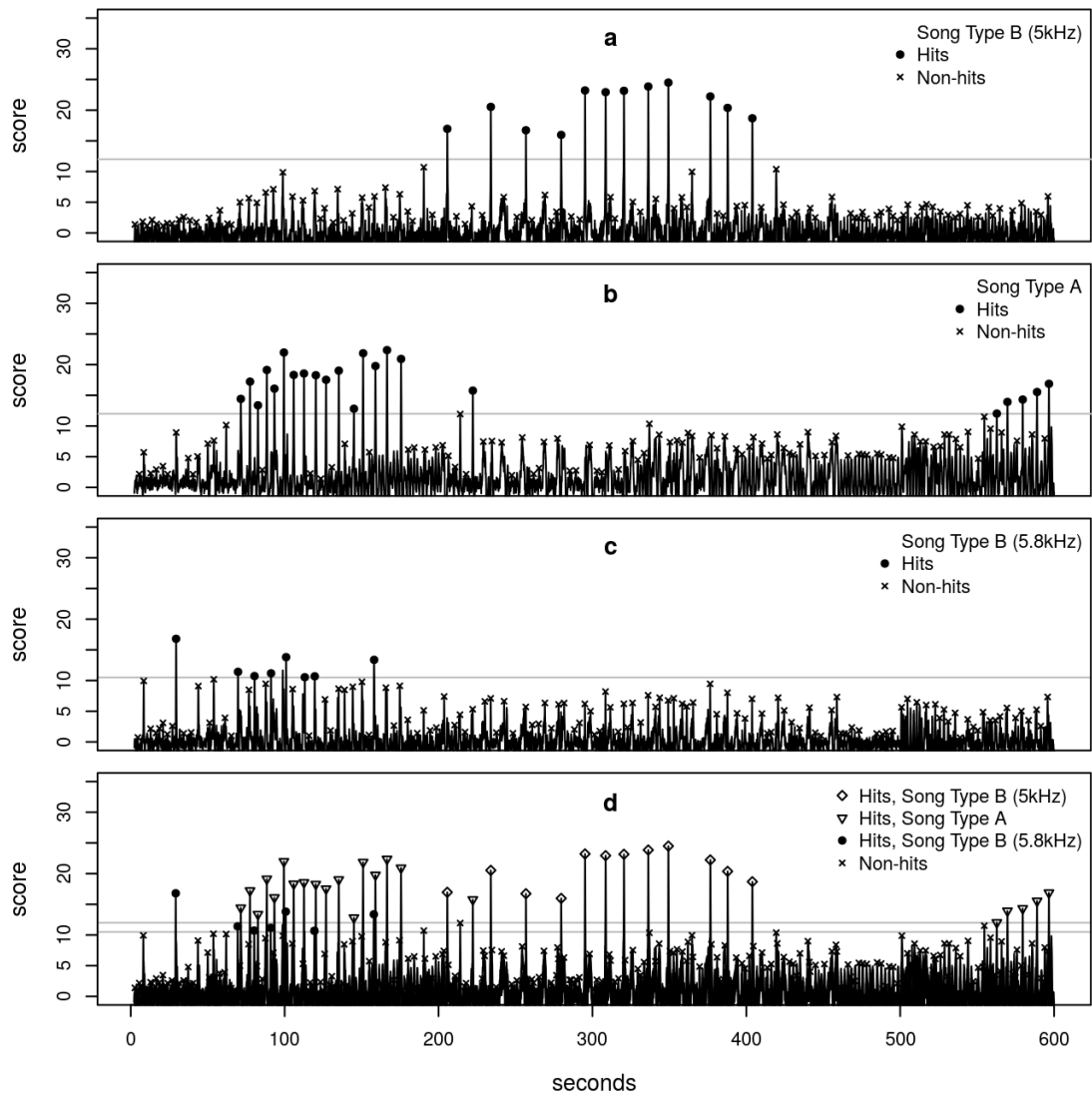


Figure 4.2

FIGURE LEGEND

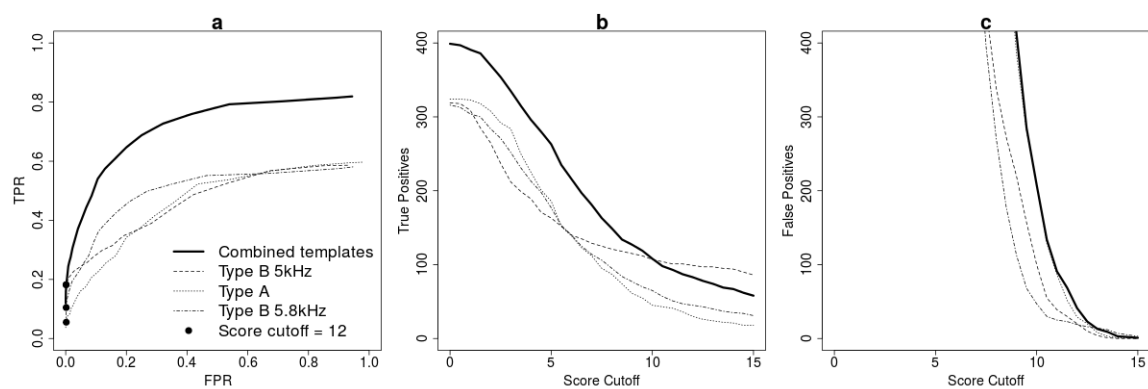


Figure 4.3

FIGURE LEGEND

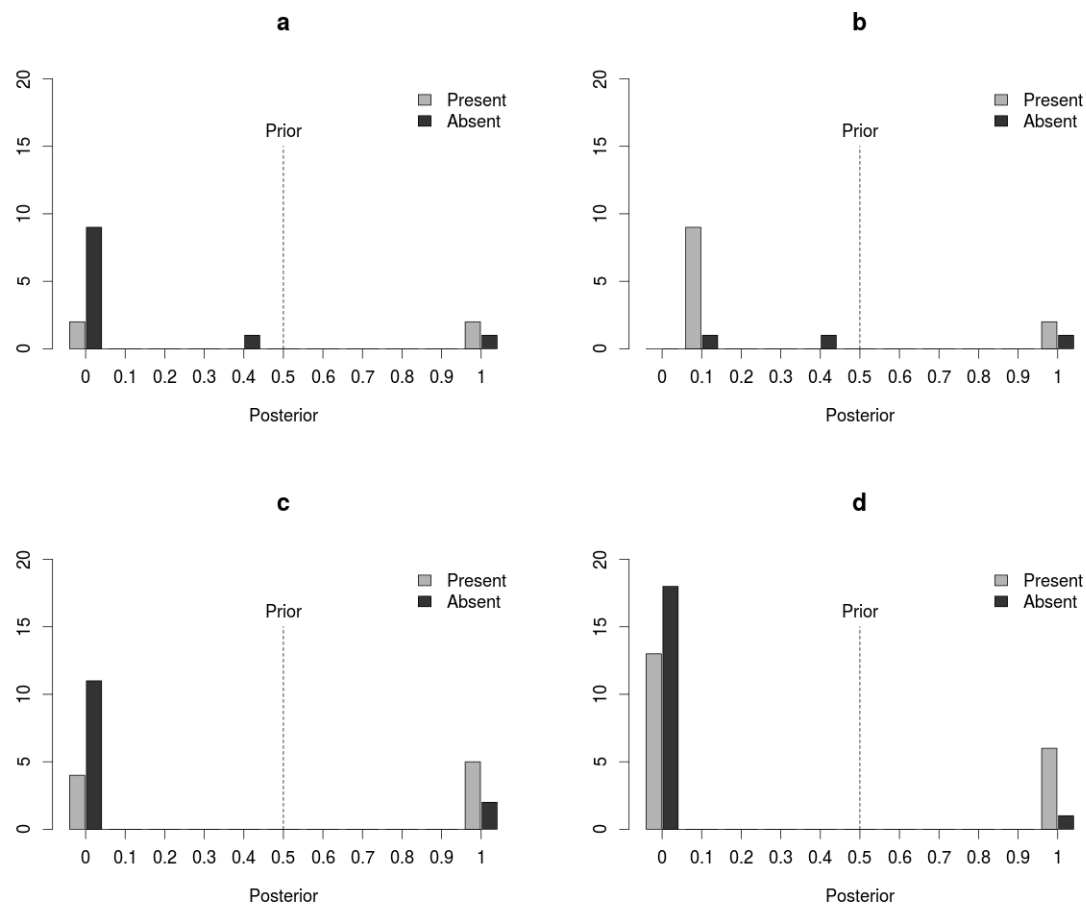


Figure 4.4

FIGURE LEGEND

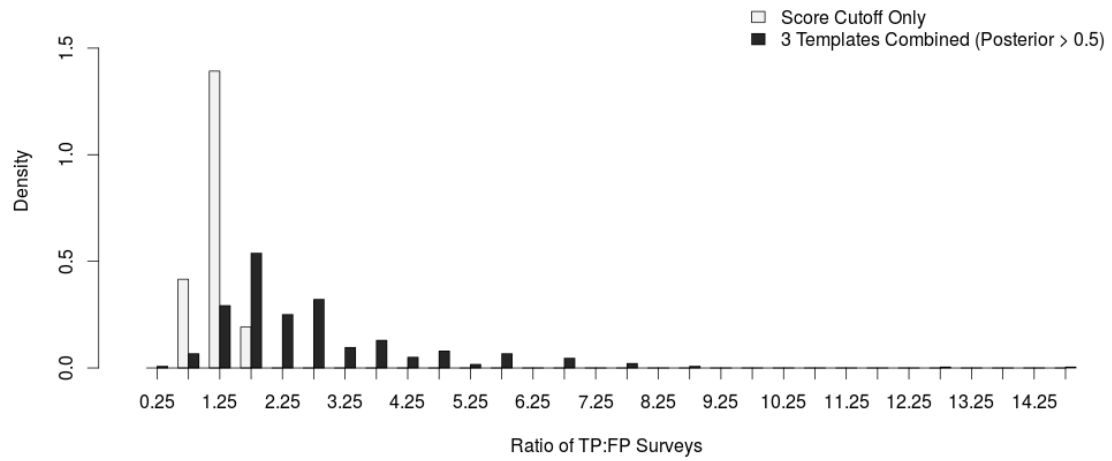


Figure 4.5

CHAPTER 5

OCCUPANCY MODELING WITH AUTOMATED ACOUSTIC DETECTIONS

Occupancy modeling with automated acoustic detections

Jonathan Katz

Vermont Cooperative Fish and Wildlife Research Unit
Rubenstein School of Environment and Natural Resources
University of Vermont, Burlington VT

Sasha Hafner

Hafner Consulting,
University of Odense
Odense, Denmark

Therese Donovan

U.S. Geological Survey
Vermont Cooperative Fish and Wildlife Research Unit
Rubenstein School of Environment and Natural Resources
University of Vermont, Burlington VT

Prepublication Disclaimer:

This draft manuscript is distributed solely for purposes of scientific peer review. Its content is deliberative and predecisional, so it must not be disclosed or released by reviewers. Because the manuscript has not yet been approved for publication by the U.S. Geological Survey (USGS), it does not represent any official USGS finding or policy.

ABSTRACT

Automated detection methods of wildlife populations are increasingly being used for identification and monitoring of landscape-scale responses to climate change and land-use change. Skepticism of automated acoustic monitoring software is largely due to higher false positive and negative error rates than those in traditional human surveys, but the false positive multiple method occupancy model is capable of estimating detection parameters and occupancy state when one method has occasional false positive detections.

We test the accuracy of the model when automated detection of black-throated green warbler is mixed with human detection in 4 recorded surveys at 60 sites. Precision and accuracy are evaluated by simulation, and we use the results to optimize future sampling. In simulation, parameter estimates by the multiple method occupancy model are close to those we computed manually when two surveys are manually analyzed. Our results support the use of the multiple method false positive occupancy model to track detection rates in automated monitoring programs. In addition to its accuracy, the strengths of this tool are that it uses easy to collect data and it simultaneously estimates a range of detection parameters. It can also be used to optimize the survey schedule to the target species by maximizing the detection power of multiple surveys.

INTRODUCTION

Although monitoring has traditionally been used to assess the results of local management activities (Gibbs et al. 1999, Johnson et al. 2006, Runge et al. 2009, Martin et al. 2011), climate change and land use change warrant a shift towards a landscape scale approach to species monitoring. Species presence and the factors that influence patterns of presence (i.e. occupancy) are parameters of interest that have seen widespread adoption by monitoring programs over the past decade (Bailey et al. 2013). Compared to monitoring abundance, monitoring species occupancy is attractive due to the efficiency of data collection (Noon et al. 2012).

The single season occupancy model (MacKenzie et al. 2002) requires a binary detection history comprised of presence/absence data from multiple surveys. For example, for a study in which sites are surveyed three times, a site may have an encounter history of 000, 010, 111, or any other combination of 0's and 1's, where a 1 indicates the species was detected and a 0 indicates non-detection. In its simplest application, the analysis will simultaneously estimate both occupancy probability (the probability that a site is occupied by a target species) and detection probability (the probability that the target species will be detected, given the site is occupied). Separate detection and occupancy covariates can be incorporated in a logistic regression framework, and model selection and model averaging practices such as those described in Anderson (2010) or Burnham and Anderson (2002) can be applied to make inference about the effects of the chosen covariates. With a sufficient number of surveys spread across a landscape and well chosen covariates, spatially explicit occupancy maps can be used to inform wildlife management at any scale from local to landscape (Schwenk and Donovan 2011), and can be used to predict future changes in occupancy in response to management practices or land use trends (Brown et al. 2014).

INTRODUCTION

Several variations of the model have been developed, including the ability to estimate the false positive rate, or the probability that a species will be recorded as present at a site when in fact it is absent (Royle and Link 2006). With this model, a detection (1) during a survey can be a true positive (TP) or false positive (FP), while a non-detection (0) can be a true negative (TN) or a false negative (FN). Additionally, multi-season models extend these variations to estimate the probability of patch extinction, in which occupancy state transitions from occupied to unoccupied, and the probability of patch colonization, in which the transition is from unoccupied to occupied (MacKenzie et al. 2003).

Automated wildlife detection methods are increasingly being used for identification and monitoring of landscape-scale responses to climate change and land-use change. Automated acoustic monitoring uses sound recording devices to capture ambient sounds in environmental surveys. When monitoring changes in a species' distribution, many sites are established across a species' range, and at each site the goal is to identify whether the species is present or not. One approach is to have humans create the encounter history by searching the recordings for identifying songs. However, this is too inefficient to be considered a viable methodology for projects with numerous sites and surveys (Swiston and Mennill 2009), and does not guarantee error-free results (Fritsch and Bruckner 2014).

The second approach is to use a computer algorithm to search for signals, and then to assign signals as "hits" when the signal matches the target sound. Skepticism of automated acoustic monitoring software is largely due to higher false positive and false negative error rates than those in traditional human surveys (Khanna et al. 1997, Swiston and Mennill 2009). However, the greater expense of traditional human surveys renders them a poor

INTRODUCTION

method of tracking changes in species distribution in the face of climate change and land use change. Rather than dismiss automated methods based on the higher error rates, the potential to sample a landscape continuously for species presence should inspire optimistic discussion about how to make these approaches work.

The challenges of automated acoustic monitoring are to 1) search for sounds recorded in a survey (Katz et al. In prep*b*), 2) aggregate the detections (which may be true or false) to determine whether the species was present in the survey (typically 5-10 minutes, although survey length can vary), rendering a 0 or 1 at the survey level (Katz et al. In prep*a*), and 3) aggregate the survey results to determine site occupancy probability, given potential false positive and false negative errors, and to monitor changes in probability over time. In scaling up to an occupancy monitoring framework, errors in detection may be propagated as the data are scaled from records of each detection to survey presence to site occupancy, thus a means of minimizing cascading detection error is necessary.

Previously we introduced a method for evaluating the sensitivity (true positive rate) and specificity (1-false positive rate) of a suite of templates for detecting individuals under the assumption that the detections have some level of correlation (e.g. calls were issued by the same individual and from the same perch). Like a medical test is useful for evaluating risk of disease, this approach is useful for evaluating how effective each single template and group of templates are at detecting a species if they are present, and how many false positives are produced when the species is not present.

From a monitoring perspective, however, we are interested in estimating the probability of occupancy, or prevalence, of a species on the landscape using test results in a series of

INTRODUCTION

surveys. If $P(O)$ is the probability a site is occupied and $P(H)$ is the probability of a hit, Bayes theorem can be used to estimate the probability that a site is occupied given a hit:

$$P(O|H) = \frac{P(H|O) * P(O)}{P(H|O) * P(O) + P(H|\sim O) * P(\sim O)}$$

The probability that a site is occupied, given a hit, is proportional to the likelihood of observing a hit under the hypothesis that the site is occupied ($P(O)$) plus the likelihood of observing a hit under the hypothesis that the site is unoccupied ($P(\sim O)$). Here, $P(H|O)$ is equivalent to sensitivity, and is also referred to as detection probability in monitoring frameworks: the probability of detecting the species, given it is present. The term $P(H|\sim O)$ is the probability of getting a hit, given the site is unoccupied. In monitoring frameworks this is a false positive result.

Similarly, we could use Bayes theorem to estimate the probability of occupancy given non-detection within a survey:

$$P(O|\sim H) = \frac{P(\sim H|O) * P(O)}{P(\sim H|O) * P(O) + P(\sim H|\sim O) * P(\sim O)}$$

where $P(\sim H|O)$ is equivalent to specificity ($TN/(FP + TN)$). The term $P(\sim H|O)$ is the probability of not getting a hit, given the site is occupied. In monitoring frameworks, this is a false negative result. Accurate estimation of occupancy probability at a site is known to depend on accounting for the sensitivity of the detection method (MacKenzie et al. 2002) as well as the specificity (Royle and Link 2006), which are likely to vary by site with temporal or spatial covariates. From an acoustic monitoring perspective, what is needed is an analytical framework that can estimate site occupancy patterns across a suite of sites

INTRODUCTION

simultaneously, while accounting for both false positive and false negative detections that result from automated detection.

Recently, Miller et al. (2011) introduced a single season occupancy model in which these parameters can be estimated with two alternative modeling frameworks. In their multiple method modeling framework, observations (surveys that result in a 0 or 1) are systematically produced by two different methods. This framework appears to be well suited to acoustic monitoring programs, where the majority of observations would be produced by automated methods (method 1, which is capable of false negative and false positive errors), and a minority of detections could be produced by manual evaluation of the recorded surveys (method 2, capable of only false negative errors). The validated surveys can positively contribute to the quality of monitoring, and detection parameters can be estimated simultaneously with occupancy parameters. If such an analytical framework can estimate occupancy with automatic detection data, the framework may assume a more central role in wildlife management despite higher overall error rates relative to traditional human surveys.

In this paper we will 1) use the Miller et al. 2011 multiple method occupancy model with automatic detections for black-throated green warbler (*Setophaga virens*), 2) compare error rates estimated by the multiple method model to manually computed error rates and examine the relationship between per-survey error rates and per-site error rates, and 3) use the survey detection probability to estimate the power of automated acoustic monitoring. All automated detection was performed in the R package *monitoR* (Katz et al. In prep^b).

METHODS

FIELD RECORDINGS

The recorded surveys were gathered at 10 locations across Vermont and eastern New York in the northeast USA between May 17 and July 5, 2010. Sites were located at national parks and a Vermont state park (Saratoga National Historical Park, Stillwater NY; Marsh-Billings-Rockefeller National Historical Park, Woodstock VT; Gifford Woods State Park, Killington VT). We placed autonomous recording units (Wildlife Acoustics SM1) at each location. Field recordings were timed to coincide with those that would have been made by a hypothetical human observer, and were collected within five hours after sunrise. Metadata pertaining to each survey, site, recorder, and digital media card was stored in the default MySQL database provided with the R package *monitoR*. WAC0 files were converted to WAV format using the beta version of Wildlife Acoustics' software *Kaleidoscope* (Wildlife Acoustics 2012) and analyzed manually and with *monitoR* in WAV format.

The single-season multinomial occupancy model proposed by Miller et al. assumes that the sites remain closed to changes in occupancy throughout the entire survey period (MacKenzie et al. 2002, Miller et al. 2011). This assumption would be violated if new breeding territories were established or territories were vacated during survey collection, an assumption that is likely to be violated over the 50 day sampling period (Rota et al. 2009). While a multi-season model is appropriate to estimate occupancy and changes of occupancy during our two month sampling period, our 10 sites did not have the power to estimate all parameters of a multi-season model. Because our primary goal in this paper was to assess the performance of the Miller single season occupancy model utilizing acoustic data (rather than to

METHODS

study changes in site occupancy across 10 sites), we controlled for changing occupancy over time by analyzing each location-date combination as a site in the encounter history and analyzing four surveys from each location-date.

With a 50 day sampling period and 10 locations, 500 location-dates were available for analysis. We systematically selected a total of 60 location-date sites from recordings on non-consecutive days to resemble scheduled surveys. Due to high human visitation rates to the sites on weekends, weekends were excluded from the sampling frame to minimize the effects human presence may have on natural song rates of breeding birds. For each location-date, a 40-minute recording from within 5 hours after sunrise was split into four back-to-back ten-minute surveys. Using back-to back surveys is a common strategy to avoid violating the closed-occupancy assumption, but we recognize that it is more likely to violate the assumption of random detection as presence in the later surveys probably correlates with presence in earlier surveys. We accepted this risk because we are more interested in the detection parameters than the state parameters.

STUDY SPECIES

The black-throated green warbler was selected as a common and vocal species that was likely to occur in a sufficient number of surveys for parameter estimation. The black-throated green warbler is a forest-nesting Neotropical migrant that breeds in mixed deciduous/coniferous forest cover and typically arrives on breeding territories in the northeastern United States during early May. Singing peaks several weeks after arrival and then declines during nest incubation and provisioning. It sings two recognized song types: type “A” is more descriptively known as the unaccented ending song (phonetically: zee zee zee zoo zee) and type

METHODS

“B” is known as the accented ending song (phonetically: zoo zee zoo zoo zee). Both songs consists of five or six notes and have a duration of approximately 1.5 - 2.0 seconds. Loose associations have been made between daily activities and song types, with type A observed being used more for territorial activities and type B observed more often in the presence of a female. The song rate for song type A is approximately 6-8 songs per minute, and the rate for song type B is approximately 2-4 songs per minute (all data from Morse and Poole 2005).

MANUAL DETECTION

Black-throated green warbler presence was manually identified in the spectrograms of all surveys via a combination of visual and aural inspection using the function `viewSpec` in the R package `monitoR` (Katz et al. In prep^b). Detection of one or more songs in a survey was simplified to presence and coded with a 1 in cases where identification was unambiguous, regardless of the number of songs in the surveys or the song types present. We assumed that false positives with manual detection was not possible. A 0 was assigned to surveys where the species was not detected.

We then aggregated the 4 surveys at each site into an encounter history, such as 0100. A site with this encounter history would be considered occupied because black-throated green warbler was observed during the second survey, and 3 of the surveys would thus have failed to detect the species. It is important to note that a survey 0 is conditional on site occupancy state; if the site is unoccupied a 0 occurs because the species is truly absent, and if the site is occupied a 0 occurs because the species was not vocalizing during the survey. Likewise, a 0000 history can indicate that the site was truly unoccupied, or that it was occupied but that the species never vocalized and/or was not recorded by our system in any of our surveys.

METHODS

AUTOMATIC SONG DETECTION

We used `monitoR` to detect black-throated green warbler in the recorded surveys. This process required creating templates of black-throated green warbler songs and then scanning the recordings to match unknown sounds against each template. We used the `monitoR` function `makeBinTemplate` to create three binary point templates that could be used to automatically locate black throated green warbler songs within a recording. We made a single binary point template for song type A, and for song type B we made templates for songs issued at 5.0 kHz and 5.8 kHz (Figure 5.1).

We then scanned each survey with one of `monitoR`'s template matching systems. We previously demonstrated that the results of spectrogram cross correlation and binary point matching are similar (Katz et al. In prep), so we used only binary point matching in this paper. The primary advantage of binary point matching over spectrogram cross correlation is that the score for binary point matching has units that are easily interpreted (dB difference between the on points and the off points), whereas spectrogram cross correlation uses a more abstract correlation coefficient. For example, a score of 10 indicates that mean signal (on points) amplitude is 10dB greater than mean background noise (off points) amplitude.

A score cutoff was chosen for each template, and all surveys were scored with the `binMatch` function. The scores for each survey were reduced to “peaks”, or local maxima within a rolling time window, with the `findPeaks` function (Figure 5.2). This function also applied the score cutoff to the peaks: all peaks with a score greater than the score cutoff were considered positive detections (“hits”), and the remainder of peaks were considered nega-

METHODS

tive detections. Each positive detection may be either a true positive or a false positive; likewise, each negative detection may be either a true negative or a false negative detection. For additional details about the detection process in `monitoR` see Katz et al. (In prep^b) and Katz et al. (In prep^c).

The next step is to aggregate detections to the survey level using results from all three templates. For a perfect template detector, a single detection by any template should provide sufficient evidence to declare the species present at the site during the survey. However, for a variety of reasons, templates will never match an unknown signal perfectly. During a 10 minute survey containing nothing but white noise at a constant volume, the `findPeaks` function in `monitoR` with default arguments will identify a peak once for every window of identical duration as the template. For example, a template with a duration of 4 seconds will produce roughly 150 peaks throughout the 10-minute survey, all of which will all be evaluated against the score cutoff and presumably represent true negative detections. Traditional hypothesis testing requires choosing an acceptable rate at which false positive results may occur by pure chance; this value (*alpha*) is typically chosen *a priori*. If a hypothetical survey consists of 10 minutes of ambient noise from a northeast hardwood forest containing other bird songs, wind noise, tree branches moving, and microphone clicks, and we are willing to accept that roughly 1% of detections of random noises may match the template well enough to surpass the score cutoff by pure chance ($\alpha = 0.01$), then this false positive rate of 1% would produce an expected $150 * 0.01 = 1.5$ false positive detections per template per 10 minute survey, which we rounded to the nearest integer. In the past, we have observed a false positive rate per event for binary point matching of approximately 1% at the score cutoffs we chose for detection, which are between 10.5 and 12 dB difference depending on the template (Katz et al. In prep^c), so we used this rate for all three templates (which were of similar length). Thus, we required 3 or more detections by any single template before

METHODS

the species was declared present in a given survey. This approach should eliminate random false positives, although we did not expect it to eliminate false positives due to systemic mis-classification of another species. By using this approach, we increased the predictive value of our three-template system for detecting survey presence. If the target species was present and singing, we expected to detect more than three songs per survey because the black-throated green warbler song rate is 2-8 songs per minute, potentially yielding as many as 20 songs in each 10 minute survey. Survey results were then aggregated into an encounter history for each site-day.

OBJECTIVE 1. ESTIMATING SINGLE-SEASON MULTIPLE METHOD PARAMETERS AND PARAMETER CONFIDENCE

We used the Miller et al. (2011) single-season, multiple method occupancy model to analyze the detection histories across location-days. This model allows both true and false positives as well as true and false negatives, under the assumption that the true occupancy state is static across surveys. We generated an encounter history y that was composed of observations produced by binary point matching in monitoR from $T = 2$ surveys at all $R = 60$ sites, and we combined it with encounter history w composed of observations made by manual detection in $S = 2$ surveys at all $R = 60$ sites. Given the encounter histories y and w , the model's parameters can be estimated by maximizing the function

$$L(\pi, \tau, \psi | y, w) \propto \prod_{i=1}^R \left(\sum_{k=0}^{K-1} \left[\left\{ \prod_{t=1}^T \pi_{y_{it}k} \right\} \times \left\{ \prod_{s=1}^S \tau_{w_{is}k} \right\} \psi_{ik} \right] \right)$$

where 1) π_{yk} is the probability of recording an observation from y at the i^{th} site on the t^{th} survey with monitoR conditional on occupancy state k , where the number of states is

METHODS

$K = 2$ (occupied and unoccupied), 2) τ_{wk} is the probability of observing an observation from w at the i^{th} site on the s^{th} survey with manual detection, and 3) ψ is the probability that the site i is in occupancy state k (Miller et al. 2011).

We randomly assigned two of the four surveys at each site to be manually analyzed, and analyzed the remaining two surveys with `monitoR`. The single-season, multiple method model was estimated in the R package `unmarked` (Fiske and Chandler 2011) using the `occuFP` fitting function with two surveys of type 1 detections (manual) and two surveys of type 2 detections (automated). Note that the likelihood above can potentially resolve parameters to the site and survey level with inclusion of site and survey covariates. In `unmarked`, the `occuFP` function requires a binary method detection covariate be included to distinguish π_{11} from τ_{11} (D. Miller, personal communication, October 8, 2014). We included a method detection covariate and estimated only the intercept for occupancy for our model, which generated the following estimates: mean site occupancy (ψ), mean true positive rate (sensitivity, detection probability) for `monitoR` (π_{11}), mean true positive rate (sensitivity, detection probability) for a human analyst (τ_{11}), and mean false positive rate (π_{10}) for `monitoR` (1-specificity). The model assumes the false positive rate for manual detection was 0.

OBJECTIVE 2. COMPARISON BETWEEN AUTOMATED AND MANUAL DETECTION RATES.

We used a bootstrap analysis to compare the Miller et al. true positive rate and false positive rate with manual and automated detection methods. We created an encounter history for each of 4 surveys at the 60 sites based on human detection. Sites were deemed occupied if a detection was made in any of the four surveys, and unoccupied if no detections were

METHODS

made. We then randomly sampled the 60 sites, with replacement, to generate a bootstrap sample. We computed the sensitivity in each bootstrap sample as

$$\tau_{11} = \frac{TP}{TP + FN}$$

We assumed that human detection had a false positive rate of 0.

For automated surveys, we created an encounter history for each of 4 surveys at the 60 sites based on automated detection. The same 60 randomly sampled sites selected for the manual analysis were chosen to generate a bootstrap sample. For each bootstrap sample, we computed sensitivity as

$$\pi_{11} = \frac{TP}{TP + FN}$$

And the false positive rate as

$$\pi_{10} = \frac{FP}{FP + TN}$$

Each survey was identified as a TP, FN, FP, or TN by comparing the automated detection result with the human verified site occupancy. For both manual and automated analyses, 1000 bootstrap samples were run to generate mean estimates and their precision.

METHODS

OBJECTIVE 3. OPTIMIZING AUTOMATED ACOUSTIC MONITORING FOR THE TARGET SPECIES

An optimal acoustic monitoring project is defined by balancing the number of surveys performed with the number required to confidently detect the target species. To optimize a project monitoring black-throated green warbler occupancy with binary point matching in `monitoR` we estimated the mean true positive rate for `monitoR` (π_{11}) in unmarked. The complement of π_{11} is the false negative rate (π_{01}), which is the probability of not detecting black-throated green warbler in a single survey at an occupied site. The false negative rate (also known as the type II error rate) was raised to n , the number of survey attempts, to compute the probability of n false negatives after n surveys. The complement to this probability is $1 - \beta$, or power, the probability of detecting black-throated green warbler at least once in n surveys:

$$power = 1 - (1 - \pi_{11})^n$$

Uncertainty around this estimate was calculated via simulation, in which the estimate for π_{11} was a normal distribution with a standard error identical to the standard error of our 1000 bootstrapped estimates of π_{11} after $n = 4$ surveys (Shapiro-Wilk for normality $W=0.99$, $n=1000$). The power calculation above was repeated 1000 times for each value of n from 2 to 25, drawing a random value of π_{11} from the distribution for each iteration. The mean and standard error of the 1000 power calculations were assumed to be proportional to the mean and standard error of the actual estimate of sampling power after n surveys.

RESULTS

The distribution of negative and positive detections per site is presented in Figure 5.3. There were a total of 4 surveys per each of the 60 sites. With perfect detection, each site would have had either 0 positive surveys (and hence be unoccupied), or have had 4 positive surveys (and hence been occupied). However, with imperfect detection both 0000 and 1111 histories could be generated as false negatives or false positives, respectively. For both automated and manual detection methods, the majority of surveys resulted in a 0000 encounter history (Figure 5.3).

OBJECTIVE 1. ESTIMATING SINGLE-SEASON MULTIPLE METHOD PARAMETERS

The parameter estimates from the simulation of the multiple method occupancy model are illustrated in the first row of Figure 5.4. As expected, the mean estimate of τ_{11} (the true positive rate for manual detection) was much higher than the mean estimate of π_{11} (the true positive rate for automated detection; 0.67 vs 0.21). Thus, given a site is occupied, the probability of detecting the target species in one survey was 3.2 times greater for manual detection than automated detection. The mean estimate for π_{10} (false positive rate) was 0.01.

RESULTS

OBJECTIVE 2. COMPARISON OF DETECTION RATES

With human detection in all 4 surveys at the 60 sites (240 total surveys), presence was identified in 45 surveys across 23 sites (true positive rate of 0.49; Table 5.1). With automated detection in all 240 surveys, presence was detected at 18 of the same surveys across 12 of the same sites (true positive rate of 0.20), as well as in 10 surveys across 6 likely unoccupied sites (false positive rate of 0.07).

The top row of Figure 5.4 shows the τ_{11} , π_{11} and π_{10} parameter estimates produced by unmarked, with the distribution arising as a result of randomly varying which 2 of the 4 surveys in the encounter history were analyzed manually. The mean estimate and standard error from a single maximum likelihood iteration of unmarked are superimposed on the histograms, illustrating that our estimate of standard error achieved through encounter history simulation closely matches that from unmarked in a single iteration. The bottom row of Figure 5.4 shows our manually calculated estimates for the same parameters, with the distribution arising as a result of bootstrap sampling with replacement which sites were included in the parameter estimation.

The unmarked estimate of τ_{11} (top row Figure 5.4a) overshoot our calculated estimate (bottom row Figure 5.4a). Estimates of π_{11} from both unmarked (top row Figure 5.4b) and our calculations (bottom row Figure 5.4b) agreed almost perfectly, while π_{10} appeared to be underestimated by unmarked (top row Figure 5.4c) relative to our calculation (bottom row Figure 5.4c). Mean estimates of τ_{11} , π_{11} and π_{10} obtained by bootstrap resampling agreed closely with the values observed in the complete set of encounter histories (Table 5.1).

DISCUSSION

OBJECTIVE 3. ESTIMATING MONITORING POWER

The distribution of true positive rate estimates produced by unmarked yielded power estimates that surpass 95% after 13 surveys. As expected, for human detection in recorded surveys the 95% power is surpassed much earlier at 6 surveys (Figure 5.5a). To model which method represents a better investment it would be necessary to factor in the amount of setup and implementation time required by each method and any storage or computational burden evaluating an additional 7 surveys with *monitoR* would entail, all of which are beyond the scope of the current project. We can, however, extend this estimate of power from black-throated green with its fixed sensitivity value (where power is only a function of the number of surveys performed) to a more general case where power is a function of both the number of surveys and the detection probability (Figure 5.5b). A detector with high sensitivity will be high on the y-axis and will require few surveys to achieve 95% power (e.g. human detection at about 0.68), whereas one with low sensitivity will be low on the y-axis and will require many more surveys to approach the 95% certainty of detection.

DISCUSSION

Although estimates of detection parameters are rarely primary monitoring goals, they are still a critical part of evaluating monitoring data for decision making purposes. Our results suggest the Miller et al. (2011) false positive occupancy model can be used to estimate detection rates in automatic detection data, without the need to implement simulated callback studies (such as Miller et al. 2012) and without the need to manually analyze the complete survey history. Although we did not evaluate it here, the multi-season extension to the multiple method occupancy model with false negatives and false positives provides

DISCUSSION

a compelling framework for long-term monitoring. For any monitoring program, careful project planning and implementation is paramount, as are the tools to plan survey collection, perform automated detection, and estimate occupancy from the results.

Our study suggests a number of factors should be considered in establishing an acoustic monitoring study. First, the Miller et al. (2011) model requires a portion of the surveys be manually identified. In our study, redundant encounter histories were gathered in two ways: 100% with monitoR and 100% with human detection. This allowed us to randomly select 2 surveys from each method to test the Miller model. In practice, a large monitoring program will not be able to afford 100% human detection, and may choose to either randomize manual verification or systematize it so that certain dates are designated for manual analysis. Although Miller et al. (2011) explicitly state that not every site must have an encounter history from both methods to estimate occupancy, the detection parameters (and by extension the state parameter) will best represent all sites when an effort is made to distribute the verification evenly across sites. The level of verification will likely vary from study to study. When we attempted the multiple method analysis with only a single survey of human detections, we found the model was unable to estimate the false positive rate. Two surveys represented 50% of our surveys at each location-date; while it is unlikely that wildlife monitoring program staff will check 50% of the surveys manually, it appears that manually checking more than one survey per site will produce more accurate estimates. Ultimately it must be evaluated within the program to determine what fraction of surveys is practical to manually verify.

Second, given the low sensitivity of automated monitoring compared to human monitoring, the number of surveys to evaluate must be considered. The low detection probabilities we observed are tolerable for monitoring relatively common and vocal species such as the black-throated green warbler, but the use of automated monitoring tools for monitoring

DISCUSSION

rare species present unique challenges—the decreased detection probability effectively makes all species, especially rare ones, even more difficult to detect than they already are with manual methods. Testing and modeling detection probabilities during the startup phase of a monitoring program will be time consuming but rewarding, as extrapolating error rates from the per-survey level to the per-site level can spur insights on whether the strengths of automated tools (e.g. the ability to approach sampling “saturation”, when additional surveys no longer yield additional information) will justify the magnitude or volume of errors scaled to the site level, or whether it is feasible to survey for a species when the $1 - \beta$ value after the chosen number of surveys is known. With human-based data collection, a major issue is dealing with the tradeoff between the number of sites versus the number of surveys per site. Mackenzie and Royle (2005) show that the optimal sample design will not be to always survey more sites or to always survey each site more often. When detection rates are low and the number of sites are fixed, more surveys will always be better than fewer surveys. With the ability to collect data 24-7 with automated systems, the question becomes what is the minimum number of surveys to evaluate to accurately estimate occupancy. For black-throated green warblers in Vermont, 13 surveys per site will yield a study design with a power of 0.95; thus, if the species is present, we will have a 95% chance of detection at each site. Again, this result is species and location specific.

Third, the timing of surveys within a site is an important additional consideration to meet the assumptions of closure. An important assumption of the multiple method model, and many other single season occupancy models, is that the site does not change occupancy state across surveys. Recently, Rota et al. (2009) called attention to the fact that this assumption may be violated frequently. We used back-to-back surveys of 10 minute duration, but the criticism with this approach is that the surveys will not be independent. We examined this issue for our data and found between 0.22 and 0.55 correlation (mean square con-

DISCUSSION

tingency) coefficients for temporally adjacent surveys. Correlations between non-adjacent surveys ranged between 0.11 and 0.31. However, since most of the sites were unoccupied, we expected high correlations as many of the surveys would have 0 detections with a low false positive rate. When considering only sites that are occupied, the correlations range between 0.29 and -0.29, suggesting that detections between consecutive surveys during the dawn chorus may not be as dependent as thought. The disparity between correlation values for all sites and occupied sites suggests that a negative detection in a survey is a fairly good predictor of a negative detection in the previous or next survey, while a positive detection is only a moderately good predictor of a positive detection in the next survey but is a weak-to-moderate predictor of non-detection in surveys that are separated by 10 minutes or more. The negative correlation at 10 minutes or more separation is possibly due to the birds systematically moving within their territory and thus singing from outside the recorded area.

Finally, although the true positive rates and false positive rates calculated here are not dependent on occupancy rates (but are instead conditional on them), ultimately the template matching system will be most useful for species with moderate to high occupancy across surveyed sites. Restricting automated monitoring to species that are not excessively rare will avoid the false positive paradox, in which even though the true positive rate is much higher than the false positive rate, more false positive than true positive detections will be logged due to the low occupancy in the study area. Situations such as this are bound to be more common with automated monitoring, as true positive rates may be depressed and false positive rates elevated due to the detection method. The detectability and distinctiveness of the target species vocalizations are additional factors that will influence true positive and false positive rates. Careful selection of the target species should allow monitoring of important trends in ecosystem health and function with detection rates that produce consistent and accurate results; however, the method may not be suitable for all species, nor

ACKNOWLEDGEMENTS

would it be practical to monitor hundreds of species.

ACKNOWLEDGEMENTS

We thank David Miller for his assistance and input. We also thank _____ for their critical review of this manuscript. The Vermont Cooperative Fish and Wildlife Research Unit is jointly supported by the U.S. Geological Survey, the University of Vermont, the Vermont Department of Fish and Wildlife, and the Wildlife Management Institute. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

REFERENCES

- Anderson, D. R. 2010. Model Based Inference in the Life Sciences: A Primer on Evidence. Springer. New York.
- Bailey, L. L., D. I. MacKenzie, and J. D. Nichols. 2013. Advances and applications of occupancy models. *Methods in Ecology and Evolution* pages n/a–n/a.
- Brown, M. L., T. M. Donovan, W. S. Schwenk, and D. M. Theobald. 2014. Predicting impacts of future human population growth and development on occupancy rates of forest-dependent birds. *Biological Conservation* **170**:311–320.
- Burnham, K. P., and D. R. Anderson. 2002. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. Springer. New York.
- Fiske, I., and R. Chandler. 2011. unmarked: An R Package for Fitting Hierarchical Models of Wildlife Occurrence and Abundance. *Journal of Statistical Software* **43**:1–23. URL <http://www.jstatsoft.org/v43/i10/>.
- Fritsch, G., and A. Bruckner. 2014. Operator bias in software-aided bat call identification. *Ecology and Evolution* **4**:2703–2713.
- Gibbs, J. P., H. L. Snell, and C. E. Causton. 1999. Effective monitoring for adaptive wildlife management: Lessons from the Galapagos Islands. *Journal of Wildlife Management* **63**:1055–1065.

REFERENCES

- Johnson, L., R. J. Camp, K. W. Brinck, and P. C. Banko. 2006. Long-Term Population Monitoring: Lessons Learned From an Endangered Passerine in Hawai'i. *Wildlife Society Bulletin* **34**:1055–1063.
- Katz, J., S. Hafner, and T. Donovan. In prep^a. Assessing survey presence with automated detection. .
- Katz, J., S. Hafner, and T. Donovan. In prep^b. Avian acoustic monitoring with the R package *monitoR* .
- Katz, J., S. Hafner, and T. Donovan. In prep^c. Evaluating error rates of the R package *monitoR* .
- Khanna, H., S. L. L. Gaunt, and D. A. McCallum. 1997. Digital spectrographic cross-correlation: tests of sensitivity. *Bioacoustics* **7**:209–234.
- Mackenzie, D., and J. Royle. 2005. Designing occupancy studies: general advice and allocating survey effort. *Journal of Applied Ecology* **42**:1105–1114.
- MacKenzie, D. I., J. D. Nichols, J. E. Hines, M. G. Knutson, and A. B. Franklin. 2003. Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly. *Ecology* **84**:2200–2207.
- MacKenzie, D. I., J. D. Nichols, G. B. Lachman, S. Droege, J. A. Royle, and C. A. Langtimm. 2002. Estimating site occupancy rates when detection probabilities are less than one. *Ecology* **83**:2248–2255.
- Martin, J., P. L. Fackler, J. D. Nichols, B. C. Lubow, M. J. Eaton, M. C. Runge, B. M. Stith, and C. A. Langtimm. 2011. Structured decision making as a proactive approach to dealing with sea level rise in Florida. *Climatic Change* **107**:185–202.

REFERENCES

- Miller, D. A., J. D. Nichols, B. T. McClintock, E. H. C. Grant, L. L. Bailey, and L. A. Weir. 2011. Improving occupancy estimation when two types of observational error occur: non-detection and species misidentification. *Ecology* **92**:422–428.
- Miller, D. A. W., L. A. Weir, B. T. McClintock, E. H. C. Grant, L. L. Bailey, and T. R. Simons. 2012. Experimental investigation of false positive errors in auditory species occurrence surveys. *Ecological Applications* **22**:1665–1674.
- Morse, D. H., and A. F. Poole, 2005. Black-throated Green Warbler (*Setophaga virens*).
- Noon, B. R., L. L. Bailey, T. D. Sisk, and K. S. McKelvey. 2012. Efficient Species-Level Monitoring at the Landscape Scale. *Conservation Biology* **26**:432–441.
- Rota, C. T., R. J. Fletcher Jr, R. M. Dorazio, and M. G. Betts. 2009. Occupancy estimation and the closure assumption. *Journal of Applied Ecology* **46**:1173–1181.
- Royle, J. A., and W. A. Link. 2006. Generalized site occupancy models allowing for false positive and false negative errors. *Ecology* **87**:835–841.
- Runge, M. C., J. R. Sauer, M. L. Avery, B. F. Blackwell, and M. D. Koneff. 2009. Assessing Allowable Take of Migratory Birds. *Journal of Wildlife Management* **73**:556–565.
- Schwenk, W. S., and T. M. Donovan. 2011. A Multispecies Framework for Landscape Conservation Planning. *Conservation Biology* **25**:1010–1021.
- Swiston, K. A., and D. J. Mennill. 2009. Comparison of manual and automated methods for identifying target sounds in audio recordings of Pileated, Pale-billed, and putative Ivory-billed woodpeckers. *Journal of Field Ornithology* **80**:42–50.
- Wildlife Acoustics, 2012. Kaleidoscope [Computer Software]. URL <http://www.wildlifeacoustics.com>.

REFERENCES

Table 5.1: Raw numbers of detections and overall detection rates for all 240 surveys and 60 sites.

	Survey Detections			Site Detections		
	Manual	Detected	Not Detected	Manual	Detected	Not Detected
Present	45	18	27	23	12	11
Absent	195	10	185	37	6	31
	TPR	TPR	FNR	TPR	TPR	FNR
Present	0.49	0.20	0.80	0.38	0.52	0.48
	TNR	FPR	TNR	TNR	FPR	TNR
Absent	0.51	0.07	0.93	0.62	0.16	0.84

FIGURE LEGEND

FIGURE LEGEND

Figure 5.1. Binary point templates for black-throated green warbler detection.

Figure 5.2. Spectrogram of a survey with four detected black-throated green warbler songs (top, in boxes) and automated detection scores (bottom, solid black line) and “peaks” (points); dashed line indicates score cutoff, above which peaks become “hits”.

Figure 5.3. Count of positive surveys per site as detected by monitoR (a) and human detection (b); $n = 60$.

Figure 5.4. Automated detection parameter estimates using the multiple method occupancy model by Miller et al. (2011), 2 of 4 surveys manually evaluated. (a) Detection probability for a human analyst (τ_{11}); (b) detection probability for monitoR (π_{11}), and (c) false positive probability for monitoR (π_{10}). The solid line is the normal estimate from a representative iteration in unmarked with its standard error value.

Figure 5.5. Power of manual and automated detection by number of surveys. (a) For black-throated green warbler - points indicate $1 - \beta$ value for detection, SE bars were estimated by simulation from the distributions on calculated values of τ_{11} and π_{11} . (b) The general case for all π_{11} values and two power thresholds. The general case is estimated using the mean calculated values for τ_{11} and π_{11} .

FIGURE LEGEND

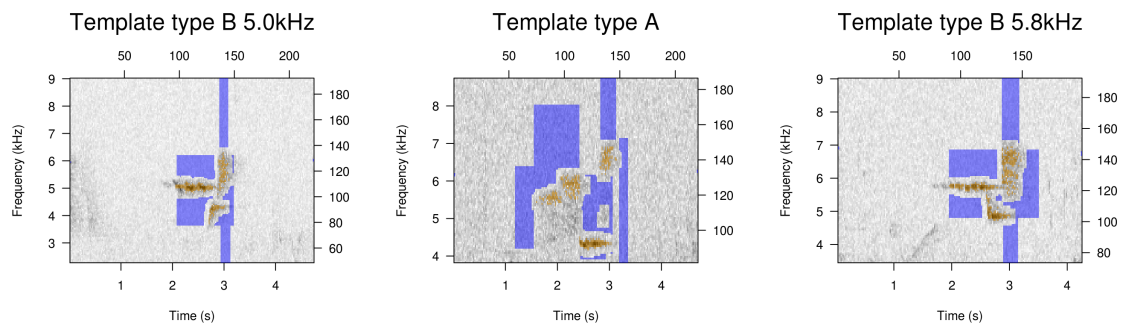


Figure 5.1

FIGURE LEGEND

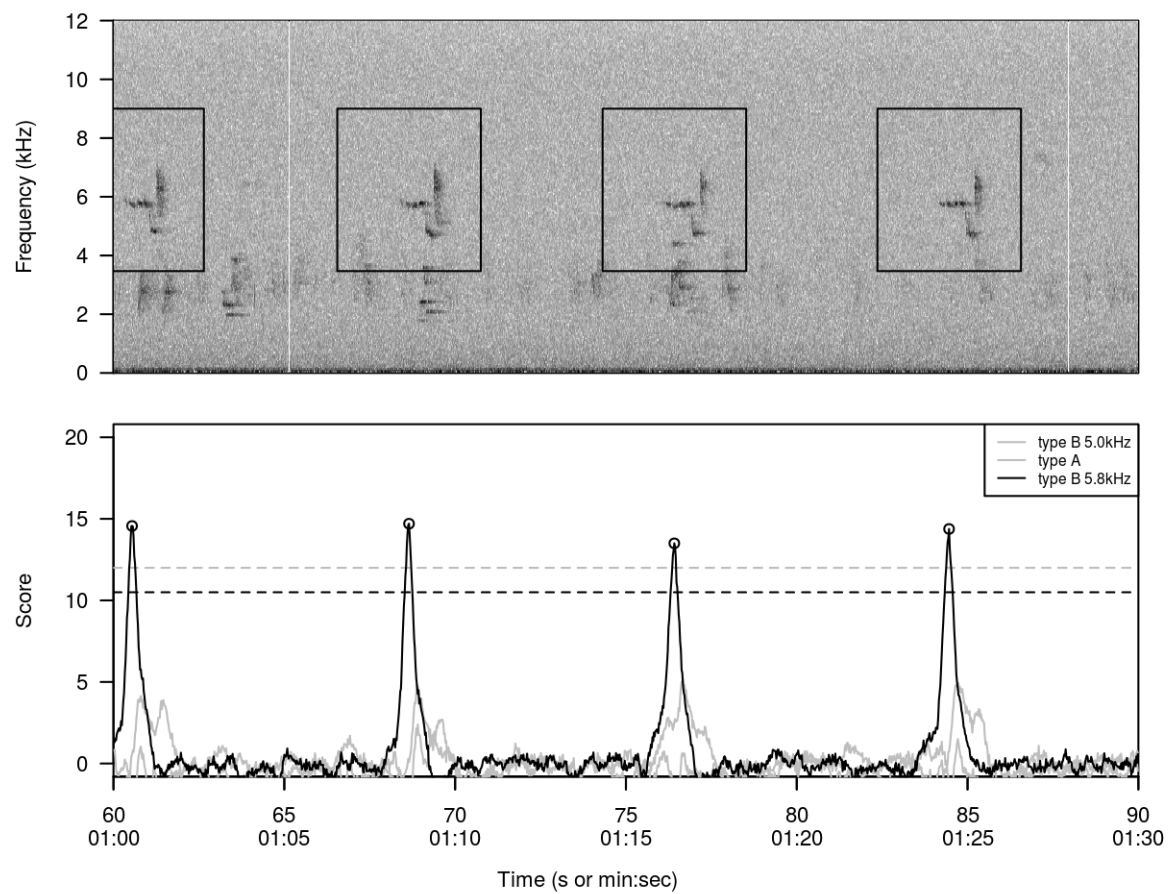


Figure 5.2

FIGURE LEGEND

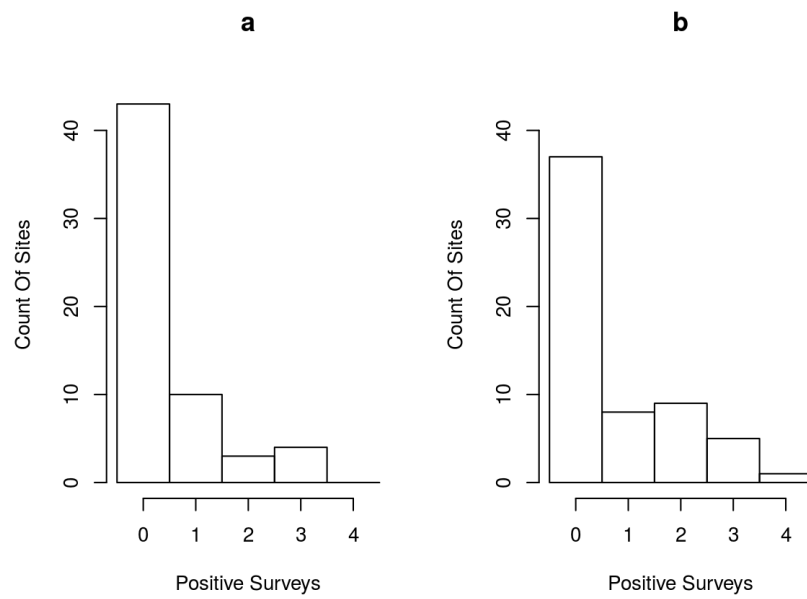


Figure 5.3

FIGURE LEGEND

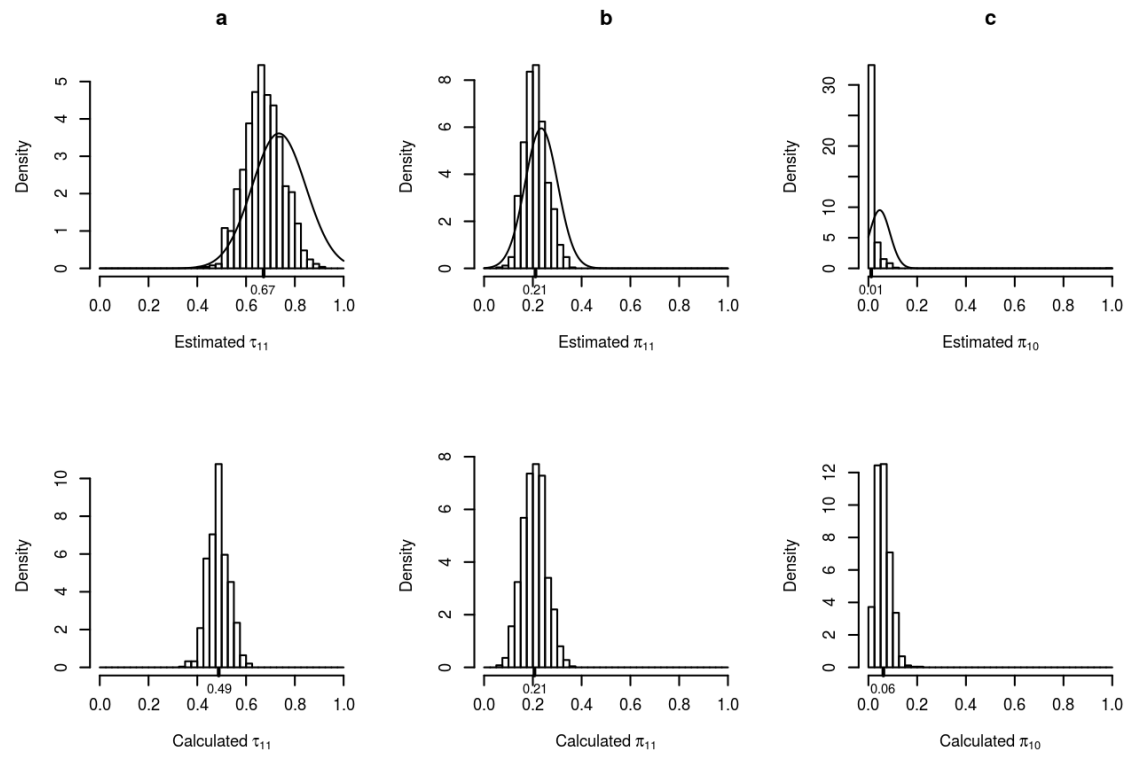


Figure 5.4

FIGURE LEGEND

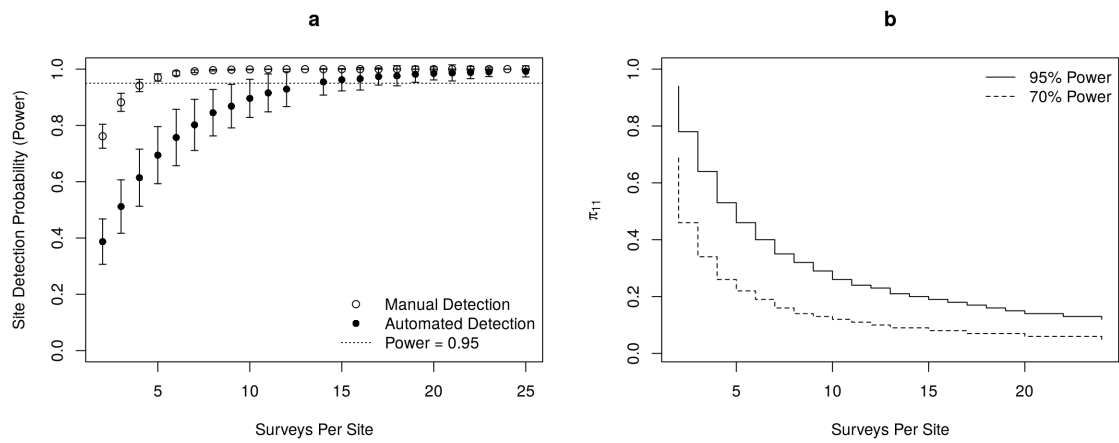


Figure 5.5

APPENDIX A

APPENDIX A. BINARY POINT MATCHING AND SPECTROGRAM CROSS-CORRELATION

Binary point matching and spectrogram cross-correlation

Jonathan Katz

BINARY POINT MATCHING

THE TEMPLATE

The business component of a Binary point matching template consists of two matrices: one specifies the time (x) and frequency (y) coordinates of “on” points, and the other specifies the x and y coordinates of “off” points. The on points are locations where signal is expected; the off points are locations where no signal is expected. When the on and off points are co-plotted (Figure 1a), the gaps around the ragged matrix of the template and a regular 3x4 matrix are not assigned values; the analysis simply ignores points that are neither on nor off. The on matrix for the simplified template in Figure 1a is:

$$\begin{bmatrix} t & frq \\ 2 & 4 \\ 3 & 4 \end{bmatrix}$$

and it is plotted in orange. The off matrix is:

$$\begin{bmatrix} t & frq \\ 1 & 4 \\ 2 & 2 \\ 2 & 6 \\ 4 & 4 \end{bmatrix}$$

BINARY POINT MATCHING

and it is plotted in blue. In practice, selection of on and off points for the template would be guided by a spectrogram of a song (in grayscale behind colors), but notice that no amplitude information is stored from the song that is used as a guide—only the locations where we expect high and low amplitude to be in the survey matrix.

The template also contains additional attributes such as the frequency bandpass parameters, the file path to the original sound clip, the Fourier transform parameters, and the score cutoff.

SCORING

When scoring, the points in the survey spectrogram corresponding to the (x,y) locations in the template are extracted. Note that the survey spectrogram usually contains more columns (time bins) than the template (Figure 1b). The example survey spectrogram below will compute two survey on and off matrices; the first spans time bins 1:4, and the second is shifted one time bin to the right and spans time bins 2:5. The average amplitude of the on and off points in the survey matrices are compared to generate a score for each:

$$score = \frac{\sum amp_{on}}{N_{on}} - \frac{\sum amp_{off}}{N_{off}}$$

The primary difference between the implementation in *monitoR* and the method described by Towsey et al. (2012) is that our *on* and *off* points do not cover the entire template matrix, as described above (an implicit third category of points are those which are ignored

BINARY POINT MATCHING

by the analysis).

In our example, the first set of on points fall directly on the regions of high signal and the off points fall on random noise; the average on amplitude is -21, the average off amplitude is -59.75, and the score is the difference between those values, or 38.75. For the second set of survey vectors one of the cells of high signal in the survey matrix is under an off point and only one on point is on a cell of high signal. The average on amplitude is 40, the average off amplitude is 53.25, and the score is 13.25.

IN MONITOR

If these were actual wave files rather than matrices we would make the template with `makeBinTemplate`, run `binMatch`, and then run `findPeaks`. Without synthesizing the sounds we'll have to just pretend, so the following is not run but illustrates a typical sequence of calls:

```
templates <- makeBinTemplate(clip = "song.wav", frq.lim = c(2, 6),
  select = "click", binary = TRUE, name = "example", score.cutoff = 15)
scores <- binMatch(survey = survey, templates = templates,
  time.source = "fileinfo")
pks <- findPeaks(scores)
detections <- getDetections(pks)
```

BINARY POINT MATCHING

```
detections

##      template      date.time time score  hit
## 1  example 2014-08-08 00:46:06  2.5 38.75 TRUE
```

The function `findPeaks` doesn't have a lot to work with here, but it is designed to separate the peak score from the “attack” and “decay” produced by partial overlap of the template with the signal, and it should return only the higher score in the pair. If we checked the scores directly (`scores@scores`) we would see both scores reported.

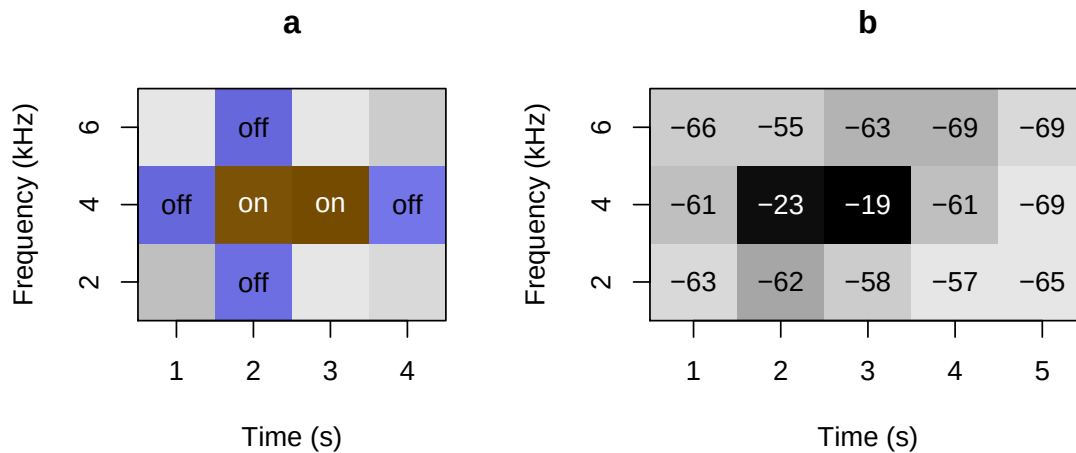


Figure A.1: A binary point matching template (a) and a short survey spectrogram (b). Note that the survey spectrogram must have at least as many time bins as the template, and in most cases will have many more.

SPECTROGRAM CROSS-CORRELATION

THE TEMPLATE

The core of a Spectrogram cross-correlation template consists of one matrix containing a time (x) component, a frequency (y) component, and an amplitude (z) component. The x and y values are selected from a spectrogram of an example song that the detection process will target; the z values are the amplitudes from the spectrogram of the example clip. In the example template depicted in Figure 2a the matrix is:

$$\begin{bmatrix} t & frq & amp \\ 1 & 4 & -58 \\ 2 & 2 & -59 \\ 2 & 4 & -18 \\ 2 & 6 & -62 \\ 3 & 4 & -14 \\ 4 & 4 & -64 \end{bmatrix}$$

It is plotted in orange.

The template also contains additional attributes such as the frequency bandpass parameters, the file path to the original sound clip, the Fourier transform parameters, and the score cutoff.

SCORING

When scoring begins, the points in the survey spectrogram corresponding to the (x,y) locations in the template are extracted, and the amplitudes of those points are stored in a survey amplitude vector. The Pearson's correlation coefficient between the template amplitude component and the survey amplitude vector is computed and becomes the score for that set of time bins.

Note that the dimensions of the survey spectrogram usually contains more columns (time bins) than the template (Figure 2b). The example survey spectrogram below computes two sets of survey matrices; the first spans time bins 1:4, and the second is shifted one time bin to the right and spans time bins 2:5. Spectrogram cross-correlation is defined by Mellinger and Clark (1997) in more general terms than the Pearson's correlation coefficient, but by assuming Fourier transforms of the template and survey and then reducing the template and survey to vectors of selected points rather than entire matrices it becomes possible to be very specific with the definition:

$$r_{T,S} = \frac{n \sum T S_i - \sum T \sum S_i}{\sqrt{n \sum T^2 - (\sum T)^2} \sqrt{n \sum S_i^2 - (\sum S_i)^2}}$$

Where the index i is the time in the survey corresponding to the center time of the survey matrix. Note that the template matrix remains static and does not need to be indexed.

The first set of survey amplitudes in our example is (-61,-62,-23,-55,-19,-61) and the second set is (-23,-58,-19,-63,-61,-69). The corresponding scores for these amplitude vectors are 0.98 and 0.33 when the cross-correlation of each vector is computed with the template amp

SPECTROGRAM CROSS-CORRELATION

vector.

IN MONITOR

Again, if these were actual wave files rather than matrices, we would make the template with `makeCorTemplate`, run `corMatch`, and then run `findPeaks`. The following is not run:

```
templates <- makeCorTemplate(clip = "song.wav", frq.lim = c(2, 6),
  select = "click", name = "example", score.cutoff = 0.4)
scores <- corMatch(survey = survey, templates = templates,
  time.source = "fileinfo")
pks <- findPeaks(scores)
detections <- getDetections(pks)
```

detections

```
##   template          date.time time score  hit
## 1  example 2014-08-08 00:46:06  2.5  0.98 TRUE
```

SPECTROGRAM CROSS-CORRELATION

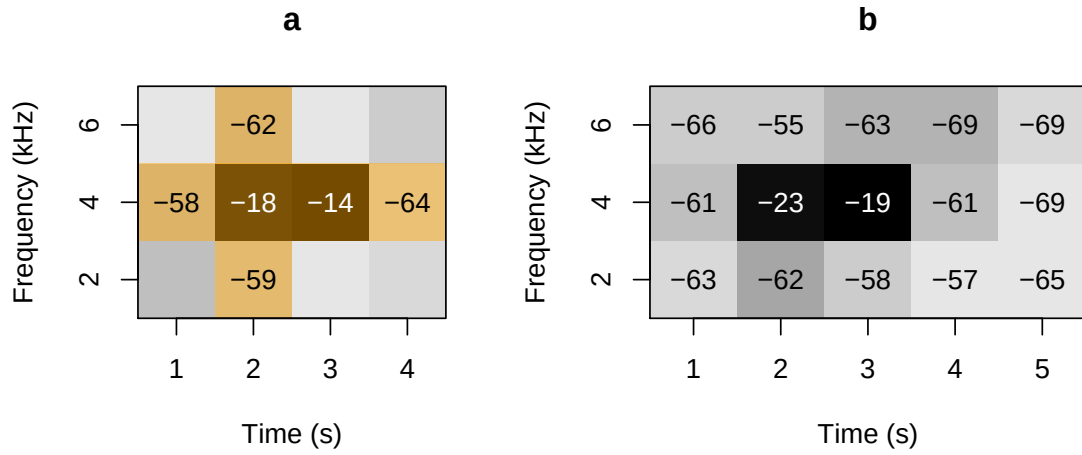


Figure A.2: A spectrogram cross-correlation template (a) and the same short survey spectrogram as in Figure 1 (b). Again, the survey spectrogram must have at least as many time bins as the template, and in most cases will have many more.

REFERENCES

- Mellinger, D. K. and C. W. Clark. 1997. Methods for automatic detection of mysticete sounds. *Marine and Freshwater Behaviour and Physiology*. 29:163-181.
- Towsey, M., B. Planitz, A. Nantes, J. Wimmer, and P. Roe. 2012. A toolbox for animal call recognition. *Bioacoustics-the International Journal of Animal Sound and Its Recording*. 21:107-125.

BIBLIOGRAPHY

- Acevedo, M. A., and L. J. Villanueva-Rivera. 2006. Using Automated Digital Recording Systems as Effective Tools for the Monitoring of Birds and Amphibians. *Wildlife Society Bulletin* **34**:211–214.
- Amar, A., C. M. Hewson, R. M. Thewlis, K. W. Smith, R. J. Fuller, J. A. Lindsell, G. Conway, S. Butler, and M. A. MacDonald. 2006. What’s happening to our woodland birds? Long-term changes in the populations of woodland birds. A re-survey of breeding bird populations of broadleaved and mixed woodlands in 2003-04 to examine changes since the 1960s, 1970s and 1980s and test these changes against a range of hypotheses for their causes. *RSPB Research Report* **19**:i–vi, 1–204.
- Anderson, D. R. 2010. *Model Based Inference in the Life Sciences: A Primer on Evidence*. Springer. New York.
- Aronov, D., and M. S. Fee. 2012. Natural Changes in Brain Temperature Underlie Variations in Song Tempo during a Mating Behavior. *Plos One* **7**.
- Avisoft Bioacoustics e.K., 2014. Avisoft-SASLab Pro version 5.2 [Computer Software]. URL <http://www.avisoft.com/>.
- Bagwell, C., R. Sykes, and P. Giard, 2013. SoX: Sound eXchange. Version 14.4.1. URL <http://sox.sourceforge.net/Main/HomePage>.

BIBLIOGRAPHY

- Bailey, L. L., D. I. MacKenzie, and J. D. Nichols. 2013. Advances and applications of occupancy models. *Methods in Ecology and Evolution* pages n/a–n/a.
- Barbraud, C., and H. Weimerskirch. 2006. Antarctic birds breed later in response to climate change. *Proceedings of the National Academy of Sciences of the United States of America* **103**:6248–6251.
- Bart, J., and J. D. Schoultz. 1984. Reliability of Singing Bird Surveys: Changes in Observer Efficiency with Avian Density. *The Auk* **101**:307–318.
- Beecher, M., P. Stoddard, S. Campbell, and C. Horning. 1996. Repertoire matching between neighbouring song sparrows. *Animal Behaviour* **51**:917–923.
- Bioacoustics Research Program, 2011. Raven Pro: Interactive Sound Analysis Software [Computer Software]. URL <http://www.birds.cornell.edu/raven>.
- Blickley, J. L., D. Blackwood, and G. L. Patricelli. 2012. Experimental Evidence for the Effects of Chronic Anthropogenic Noise on Abundance of Greater Sage-Grouse at Leks. *Conservation Biology* **26**:461–471.
- Bolsinger, J. S. 2000. Use of two song categories by Golden-cheeked Warblers. *Condor* **102**:539–552.
- Both, C., S. Bouwhuis, C. M. Lessells, and M. E. Visser. 2006. Climate change and population declines in a long-distance migratory bird. *Nature* **441**:81–83.
- Boucher, N. J., and M. Jinnai, 2012. SoundID Documentation: A Revolutionary Sound Recognition System. URL <http://www.soundid.net>.
- Boulanger, J., S. Gordon, and R. Munro. 2004. Sources of Heterogeneity Bias When DNA Mark-Recapture Sampling Methods Are Applied to Grizzly Bear (*Ursus arctos*) Populations. *Journal of Mammalogy* **85**:618–624. URL <http://www.jstor.org/stable/1383780>.

BIBLIOGRAPHY

- Bower, J., and C. Clark. 2005. A field test of the accuracy of a passive acoustic location system. *Bioacoustics-The International Journal of Animal Sound and Its Recording* **15**:1–14.
- Brauer, C., T. M. Donovan, R. M. Mickey, J. Katz, and B. R. Mitchell. In Revision. A Comparison of Acoustic Monitoring Methods for Common Anurans of the Northeastern United States. *Wildlife Society Bulletin* **Submitted 25 Feb. 2014**.
- Brewster, J. P., and T. R. Simons. 2009. Testing the importance of auditory detections in avian point counts. *Journal of Field Ornithology* **80**:178–182.
- Briggs, F., B. Lakshminarayanan, L. Neal, X. Z. Fern, R. Raich, S. J. K. Hadley, A. S. Hadley, and M. G. Betts. 2012. Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach. *Journal of the Acoustical Society of America* **131**:4640–4650.
- Brown, M. L., T. M. Donovan, W. S. Schwenk, and D. M. Theobald. 2014. Predicting impacts of future human population growth and development on occupancy rates of forest-dependent birds. *Biological Conservation* **170**:311–320.
- Burnham, K. P., and D. R. Anderson. 2002. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer. New York.
- Burt, J., 2006. Syrinx-PC. URL <http://www.syrinxpc.com/>.
- Butler, C. J. 2003. The disproportionate effect of global warming on the arrival dates of short-distance migratory birds in North America. *Ibis* **145**:484–495.
- Campbell, M., and C. M. Francis. 2011. Using stereo-microphones to evaluate observer variation in North American Breeding Bird Survey point counts. *Auk* **128**:303–312.
- Carson, R. 1987. *Silent Spring*. 25th anniversary edition edition. Houghton Mifflin, Boston.

BIBLIOGRAPHY

- 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.
- Clark, C. W. 1980. A real-time direction finding device for determining the bearing to the underwater sounds of Southern Right Whales, *Eubalaena australis*. *The Journal of the Acoustical Society of America* **68**:508–511.
- Clark, C. W., and D. B. Peters. 2009. Acoustic System Monitors and Mitigates Harm to Marine Mammals in Real Time. *Sea Technology* **50**:10–14.
- Cortopassi, K. A., and J. W. Bradbury. 2000. The comparison of harmonically related rich sounds using spectrographic cross-correlation and principal components analysis. *Bioacoustics* **11**:89–127.
- Dawson, D. K., and M. G. Efford. 2009. Bird population density estimated from acoustic signals. *Journal of Applied Ecology* **46**:1201–1209.
- Devictor, V., R. Julliard, and F. Jiguet. 2008. Distribution of specialist and generalist species along spatial gradients of habitat disturbance and fragmentation. *Oikos* **117**:507–514.
- Digby, A., M. Towsey, B. D. Bell, and P. D. Teal. 2013. A practical comparison of manual and autonomous methods for acoustic monitoring. *Methods in Ecology and Evolution* **4**:675–683.
- Donovan, T. M., and C. H. Flather. 2002. Relationships among north American songbird trends, habitat fragmentation, and landscape occupancy. *Ecological Applications* **12**:364–374.
- Efford, M. G., D. K. Dawson, and D. L. Borchers. 2009. Population density estimated from locations of individuals on a passive detector array. *Ecology* **90**:2676–2682.

BIBLIOGRAPHY

- Evans, W. R., and D. K. Mellinger. 1999. Monitoring grassland birds in nocturnal migration. *Studies in Avian Biology* **19**:219–229,.
- Farmer, R. G., M. L. Leonard, J. E. M. Flemming, and S. C. Anderson. 2014. Observer aging and long-term avian survey data quality. *Ecology and Evolution* **4**:2563–2576.
- Farmer, R. G., M. L. Leonard, and A. G. Horn. 2012. Observer effects and avian-call-count survey quality: rare-species biases and overconfidence. *Auk* **129**:76–86.
- Farnsworth, G. L., J. D. Nichols, J. R. Sauer, S. G. Fancy, K. H. Pollock, S. A. Shriner, and T. R. Simons, 2005. Statistical Approaches to the Analysis of Point Count Data: A Little Extra Information Can Go a Long Way. USDA Forest Service General Technical Report PSW-GTR-191, 736–743. URL http://www.fs.fed.us/psw/publications/documents/psw_gtr191/Asilomar/pdfs/736-743.pdf.
- Figueroa, H., 2011. XBAT [Computer Software]. Bioacoustics Research Program. URL [Notavailable](#).
- Fiske, I., and R. Chandler. 2011. unmarked: An R Package for Fitting Hierarchical Models of Wildlife Occurrence and Abundance. *Journal of Statistical Software* **43**:1–23. URL <http://www.jstatsoft.org/v43/i10/>.
- Fitzpatrick, J. W., M. Lammertink, M. D. Luneau, T. W. Gallagher, B. R. Harrison, G. M. Sparling, K. V. Rosenberg, R. W. Rohrbaugh, E. C. H. Swarthout, P. H. Wrege, S. B. Swarthout, M. S. Dantzker, R. A. Charif, T. R. Barksdale, J. V. Remsen, S. D. Simon, and D. Zollner. 2005. Ivory-billed woodpecker (*Campephilus principalis*) persists in continental North America. *Science* **308**:1460–1462.
- Forcey, G. 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.

BIBLIOGRAPHY

- Forero-Medina, G., J. Terborgh, S. J. Socolar, and S. L. Pimm. 2011. Elevational Ranges of Birds on a Tropical Montane Gradient Lag behind Warming Temperatures. *Plos One* **6**:e28535.
- Fritsch, G., and A. Bruckner. 2014. Operator bias in software-aided bat call identification. *Ecology and Evolution* **4**:2703–2713.
- Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 1995. *Bayesian Data Analysis*. Chapman and Hall. New York, NY.
- Gibbs, J. P., H. L. Snell, and C. E. Causton. 1999. Effective monitoring for adaptive wildlife management: Lessons from the Galapagos Islands. *Journal of Wildlife Management* **63**:1055–1065.
- Gibbs, J. P., and D. G. Wenny. 1993. Song Output as a Population Estimator: Effect of Male Pairing Status. *Journal of Field Ornithology* **64**:316–322.
- Gonzalo-Turpin, H., C. Sirami, L. Brotons, L. Gonzalo, and J. L. Martin. 2008. Teasing out biological effects and sampling artifacts when using occupancy rate in monitoring programs. *Journal of Field Ornithology* **79**:159–169.
- 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.
- Graves, T. A., J. A. Royle, K. C. Kendall, P. Beier, J. B. Stetz, and A. C. Macleod. 2012. Balancing Precision and Risk: Should Multiple Detection Methods Be Analyzed Separately in N-Mixture Models? *Plos One* **7**. URL [GotoISI://WOS:000313236200015](https://doi.org/10.1371/journal.pone.0003132).
- Grieffmann, B., and M. Naguib. 2002. Song Sharing in Neighboring and Non-Neighboring Thrush Nightingales (*Luscinia luscinia*) and its Implications for Communication. *Ethology* **108**:377–387.

BIBLIOGRAPHY

- Habib, L., E. M. Bayne, and S. Boutin. 2007. Chronic industrial noise affects pairing success and age structure of ovenbirds *Seiurus aurocapilla*. *Journal of Applied Ecology* **44**:176–184.
- Harms, T. M., and S. J. Dinsmore. 2014. Influence of Season and Time of Day on Marsh Bird Detections. *Wilson Journal of Ornithology* **126**:30–38.
- Harris, J. B. C., D. D. Putra, S. D. Gregory, B. W. Brook, D. M. Prawiradilaga, N. S. Sodhi, D. Wei, and D. A. Fordham. 2014. Rapid deforestation threatens mid-elevational endemic birds but climate change is most important at higher elevations. *Diversity and Distributions* **20**:773–785.
- 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.
- Heldbjerg, H., and T. Fox. 2008. Long-term population declines in Danish trans-Saharan migrant birds. *Bird Study* **55**:267–279.
- Hobson, K. A., R. S. Rempel, H. Greenwood, B. Turnbull, and S. L. V. Wilgenburg. 2002. Acoustic surveys of birds using electronic recordings: new potential from an omnidirectional microphone system. *Wildlife Society Bulletin* **30**:709–720.
- Hoileitner, M., H. Nechtelberger, and J. Dittami. 1993. The Relationship Between Individual-Differences in Male Song Frequency and Parental Care in Blackcaps. *Behaviour* **126**:1–12.
- 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.
- Johnson, L., R. J. Camp, K. W. Brinck, and P. C. Banko. 2006. Long-Term Population Monitoring: Lessons Learned From an Endangered Passerine in Hawai'i. *Wildlife Society Bulletin* **34**:1055–1063.

BIBLIOGRAPHY

- Johnson, R. R., B. T. Brown, L. T. Haight, and J. M. Simpson. 1981. Playback recordings as a special avian censusing technique, volume 6 of *Studies in Avian Biology*, Book section estimating relative abundance - part ii, pages 68–75 . Cooper Ornithological Society.
- Katz, J., S. Hafner, and T. Donovan. In prep*a*. Assessing survey presence with automated detection. .
- Katz, J., S. Hafner, and T. Donovan. In prep*b*. Automatic detection results in an objective Bayesian format. .
- Katz, J., S. Hafner, and T. Donovan. In prep*c*. Avian acoustic monitoring with the R package `monitoR` .
- Katz, J., S. Hafner, and T. Donovan. In prep*d*. Evaluating error rates of the R package `monitoR` .
- Kendall, K. C., J. B. Stetz, D. A. Roon, L. P. Waits, J. B. Boulanger, and D. Paetkau. 2008. Grizzly Bear Density in Glacier National Park, Montana. *The Journal of Wildlife Management* **72**:1693–1705. URL <http://www.jstor.org/stable/40208452>.
- Kendall, W. L., B. G. Peterjohn, and J. R. Sauer. 1996. First-time observer effects in the North American Breeding Bird Survey. *Auk* **113**:823–829.
- Kepler, C. B., and J. M. Scott. 1981. Reducing bird count variability by training observers. *Studies in Avian Biology* **6**:366–371.
- Khanna, H., S. L. L. Gaunt, and D. A. McCallum. 1997. Digital spectrographic cross-correlation: tests of sensitivity. *Bioacoustics* **7**:209–234.
- 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. Estimaciones de la Densidad y de las Probabilidades de Detección a Partir de Muestreos

BIBLIOGRAPHY

- Utilizando Conteos en Puntos: Una Combinación de Muestreos de Distancia y de Doble Observador. **123**:735–752.
- Lee, C.-H., C.-C. Han, and C.-C. Chuang. 2008. Automatic Classification of Bird Species From Their Sounds Using Two-Dimensional Cepstral Coefficients. *IEEE Transactions on audio speech and language processing* **16**:1541–1550.
- Legg, C. J., and L. Nagy. 2006. Why most conservation monitoring is, but need not be, a waste of time. *Journal of Environmental Management* **78**:194–199.
- Lehmicke, A. J. J., J. L. Bowman, A. E. Banning, and B. L. Vasilas. 2013. Effect of Tide Level and Time of Day on Detection Rate and Abundance of Clapper Rails (*Rallus longirostris*) in a Mid-Atlantic Tidal Marsh System. *Waterbirds* **36**:364–368.
- Lemoine, N., H.-C. Schaefer, and K. Boehning-Gaese. 2007. Species richness of migratory birds is influenced by global climate change. *Global Ecology and Biogeography* **16**:55–64.
- Ligges, U., 2011. tuneR: Analysis of music. URL <http://r-forge.r-project.org/projects/tuner/>.
- Ligges, U., S. Krey, O. Mersmann, and S. Schnackenberg, 2014. tuneR: Analysis of music. URL <http://r-forge.r-project.org/projects/tuner/>.
- Luther, D., and L. Baptista. 2010. Urban noise and the cultural evolution of bird songs. *Proceedings of the Royal Society B: Biological Sciences* **277**:469–473.
- MacKenzie, D., and J. Hines, 2014. PRESENCE [Computer Software]. URL <http://www.mbr-pwrc.usgs.gov/software/presence.html>.
- Mackenzie, D., and J. Royle. 2005. Designing occupancy studies: general advice and allocating survey effort. *Journal of Applied Ecology* **42**:1105–1114.

BIBLIOGRAPHY

- MacKenzie, D. I., J. D. Nichols, J. E. Hines, M. G. Knutson, and A. B. Franklin. 2003. Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly. *Ecology* **84**:2200–2207.
- MacKenzie, D. I., J. D. Nichols, G. B. Lachman, S. Droege, J. A. Royle, and C. A. Langtimm. 2002. Estimating site occupancy rates when detection probabilities are less than one. *Ecology* **83**:2248–2255.
- Martin, J., P. L. Fackler, J. D. Nichols, B. C. Lubow, M. J. Eaton, M. C. Runge, B. M. Stith, and C. A. Langtimm. 2011*a*. Structured decision making as a proactive approach to dealing with sea level rise in Florida. *Climatic Change* **107**:185–202.
- Martin, J., J. A. Royle, D. I. Mackenzie, H. H. Edwards, M. Kery, and B. Gardner. 2011*b*. Accounting for non-independent detection when estimating abundance of organisms with a Bayesian approach. *Methods in Ecology and Evolution* **2**:595–601.
- Martin, P., and T. Martin. 2001. Behavioral interactions between coexisting species: Song playback experiments with wood warblers. *Ecology* **82**:207–218.
- Martindale, S. 1980. A Numerical Approach to the Analysis of Solitary Vireo Songs. *The Condor* **82**:199–211.
- McCallum, D. A., 2005. A Conceptual Guide to Detection Probability for Point Counts and Other Count-based Survey Methods. USDA Forest Service General Technical Report PSW-GTR-191, 754–761. URL http://www.fs.fed.us/psw/publications/documents/psw_gtr191/Asilomar/pdfs/754-761.pdf.
- McGregor, P. K., T. Dabelsteen, C. W. Clark, J. L. Bower, J. P. Tavares, and J. Holland. 1997. Accuracy of a passive acoustic location system: empirical studies in terrestrial habitats. *Ethology Ecology & Evolution* **9**:269–286.

BIBLIOGRAPHY

- Mellinger, D. K., and C. W. Clark. 1997. Methods for automatic detection of mysticete sounds. *Marine and Freshwater Behaviour and Physiology* **29**:163–181.
- Mennill, D. J. 2014. Variation in the Vocal Behavior of Common Loons (*Gavia immer*): Insights from Landscape-level Recordings. *Waterbirds* **37**:26–36.
- Mennill, D. J., J. M. Burt, K. M. Fristrup, and S. L. Vehrencamp. 2006. Accuracy of an acoustic location system for monitoring the position of duetting songbirds in tropical forest. *Journal of the Acoustical Society of America* **119**:2832–2839.
- Miller, D. A., J. D. Nichols, B. T. McClintock, E. H. C. Grant, L. L. Bailey, and L. A. Weir. 2011. Improving occupancy estimation when two types of observational error occur: non-detection and species misidentification. *Ecology* **92**:422–428.
- Miller, D. A. W., L. A. Weir, B. T. McClintock, E. H. C. Grant, L. L. Bailey, and T. R. Simons. 2012. Experimental investigation of false positive errors in auditory species occurrence surveys. *Ecological Applications* **22**:1665–1674.
- Moller, A. P. 1992. Interspecific Response To Playback of Bird Song. *Ethology* **90**:315–320.
- Moore, J. E., D. M. Scheiman, and R. K. Swihart. 2004. Field comparison of removal and modified double-observer modeling for estimating detectability and abundance of birds. *Auk (American Ornithologists Union)* **121**:865–876.
- Morse, D. H., and A. F. Poole, 2005. Black-throated Green Warbler (*Setophaga virens*).
- Mossman, D., and J. O. Berger. 2001. Intervals for Post-Test Probabilities: A Comparison of Five Methods. *Medical Decision Making* **21**:498–507.
- Nichols, J. D., and B. K. Williams. 2006. Monitoring for conservation. *Trends in Ecology & Evolution* **21**:668–673.

BIBLIOGRAPHY

- Noon, B. R., L. L. Bailey, T. D. Sisk, and K. S. McKelvey. 2012. Efficient Species-Level Monitoring at the Landscape Scale. *Conservation Biology* **26**:432–441.
- Parker, T. A. 1991. On the use of tape-recorders in avifaunal surveys. *Auk* **108**:443–444.
- Payne, K. B., M. Thompson, and L. Kramer. 2003. Elephant calling patterns as indicators of group size and composition: the basis for an acoustic monitoring system. *African Journal of Ecology* **41**:99–107.
- Pieretti, N., A. Farina, and D. Morri. 2011. A new methodology to infer the singing activity of an avian community: The Acoustic Complexity Index (ACI). *Ecological Indicators* **11**:868–873.
- Pijanowski, B. C., A. Farina, S. H. Gage, S. L. Dumyahn, and B. L. Krause. 2011. What is soundscape ecology? An introduction and overview of an emerging new science. *Landscape Ecology* **26**:1213–1232.
- R Core Team, 2012. R: A language and environment for statistical computing. URL <http://www.R-project.org/>.
- Rappole, J. H., and M. V. McDonald. 1994. Cause and effect in population declines of migratory birds. *Auk* **111**:652–660.
- 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.
- Ritchison, G. 1995. Characteristics, Use and Possible Functions of the Perch Songs and Chatter Calls of Male Common Yellowthroats. *Condor* **97**:27–38.
- Rodenhouse, N. L., S. N. Matthews, K. P. McFarland, J. D. Lambert, L. R. Iverson, A. Prasad, T. S. Sillett, and R. T. Holmes. 2008. Potential effects of climate change

BIBLIOGRAPHY

- on birds of the Northeast. *Mitigation and Adaptation Strategies for Global Change* **13**:517–540.
- Rosenstock, S. S., D. R. Anderson, K. M. Giesen, T. Leukering, and M. F. Carter. 2002. Landbird counting techniques: Current practices and an alternative. *AUK* **119**:46–53.
- Ross, J. C., and P. E. Allen. 2014. Random Forest for improved analysis efficiency in passive acoustic monitoring. *Ecological Informatics* **21**:34–39.
- Rossell, C. R. 2001. Song perch characteristics of Golden-winged Warblers in a mountain wetland. *Wilson Bulletin* **113**:246–248.
- Rota, C. T., R. J. Fletcher Jr, R. M. Dorazio, and M. G. Betts. 2009. Occupancy estimation and the closure assumption. *Journal of Applied Ecology* **46**:1173–1181.
- 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., and J. D. Nichols. 2003. Estimating abundance from repeated presence-absence data or point counts. *Ecology* **84**:770–790.
- Runge, M. C., J. R. Sauer, M. L. Avery, B. F. Blackwell, and M. D. Koneff. 2009. Assessing Allowable Take of Migratory Birds. *Journal of Wildlife Management* **73**:556–565.
- Sanderson, F. J., P. F. Donald, D. J. Pain, I. J. Burfield, and F. P. J. van Bommel. 2006. Long-term population declines in Afro-Palearctic migrant birds. *Biological Conservation* **131**:93–105.
- 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.

BIBLIOGRAPHY

- Sauer, J. R., B. G. Peterjohn, and W. A. Link. 1994. Observer differences in the North-American Breeding Bird Survey. *Auk* **111**:50–62.
- Schwenk, W. S., and T. M. Donovan. 2011. A Multispecies Framework for Landscape Conservation Planning. *Conservation Biology* **25**:1010–1021.
- Sedlacek, O., B. Cikanova, and R. Fuchs. 2006. Heterospecific rival recognition in the Black Redstart (*Phoenicurus ochruros*). *Ornis Fennica* **83**:153–161.
- Servick, K. 2014. Eavesdropping on Ecosystems. *Science* **343**:834–837.
- Slabbekoorn, H., and M. Peet. 2003. Ecology: Birds sing at a higher pitch in urban noise. *Nature* **424**:267–267.
- Staicer, C. A., V. Ingalls, and T. W. Sherry. 2006. Singing behavior varies with breeding status of American Redstarts (*Setophaga ruticilla*). *Wilson Journal of Ornithology* **118**:439–451.
- Sueur, J., T. Aubin, and C. Simonis. 2008. Seewave: a free modular tool for sound analysis and synthesis. *Bioacoustics* **18**:213–226.
- Swiston, K. A., and D. J. Mennill. 2009. Comparison of manual and automated methods for identifying target sounds in audio recordings of Pileated, Pale-billed, and putative Ivory-billed woodpeckers. *Journal of Field Ornithology* **80**:42–50.
- Szabo, A., 2013. CorrBin: Nonparametrics with clustered binary and multinomial data. R package version 1.4. URL <http://CRAN.R-project.org/package=CorrBin>.
- Thompson, W. L. 2002. Towards Reliable Bird Surveys: Accounting For Individuals Present But Not Observed. *The Auk* **119**:18–25.
- Tierney, G., B. Mitchell, A. Miller-Rushing, J. Katz, E. Denny, C. Brauer, T. Donovan, A. Richardson, M. Toomey, A. Kozlowski, J. Weltzin, K. Gerst, E. Sharron, O. Sonnen-

BIBLIOGRAPHY

- tag, C. Teshera-Sterne, and F. Dieffenbach, 2013. Phenology monitoring protocol: Northeast Temperate Network. Report Natural Resource Report NPS/NETN/NRR-2013/681. Published Report-2197242.
- Towsey, M., B. Planitz, A. Nantes, J. Wimmer, and P. Roe. 2012. A toolbox for animal call recognition. *Bioacoustics-the International Journal of Animal Sound and Its Recording* **21**:107–125.
- Urazghildiiev, I. R., and C. W. Clark. 2007. Detection performances of experienced human operators compared to a likelihood ratio based detector. *Journal of the Acoustical Society of America* **122**:200–204.
- Villanueva-Rivera, L. J., and B. C. Pijanowski, 2014. soundecology: Soundscape ecology. R package version 1.1.1. URL <http://CRAN.R-project.org/package=soundecology>.
- Villanueva-Rivera, L. J., B. C. Pijanowski, J. Doucette, and B. Pekin. 2011. A primer of acoustic analysis for landscape ecologists. *Landscape Ecology* **26**:1233–1246.
- Weir, L. A., J. A. Royle, P. Nanjappa, and R. E. Jung. 2005. Modeling Anuran Detection and Site Occupancy on North American Amphibian Monitoring Program (NAAMP) Routes in Maryland. *Journal of Herpetology* **39**:627–639.
- Wildlife Acoustics, 2011. Song Scope [Computer Software]. URL <http://www.wildlifeacoustics.com>.
- Wildlife Acoustics, 2012. Kaleidoscope [Computer Software]. URL <http://www.wildlifeacoustics.com>.
- Wimmer, J., M. Towsey, P. Roe, and I. Williamson. 2013. Sampling environmental acoustic recordings to determine bird species richness. *Ecological Applications* **23**:1419–1428.
- Wintle, B. A., M. C. Runge, and S. A. Bekessy. 2010. Allocating Monitoring Effort In The Face Of Unknown Unknowns. *Ecology Letters* **13**:1325–1337.

BIBLIOGRAPHY

- Witt, G. 2004. Moody's Correlated Binomial Default Distribution. Working Paper (Moody's Investors Service: Structured Finance Rating Methodology). **URL:** *http://fcic-static.law.stanford.edu/cdn__media/fcic-testimony/2010-0602-exhibit-binomial.pdf*.
- Wood, P. B., and K. A. Perkins. 2012. Behavioral Activities of Male Cerulean Warblers in Relation to Habitat Characteristics. *Wilson Journal of Ornithology* **124**:497–505.
- Yoccoz, N. G., J. D. Nichols, and T. Boulinier. 2001. Monitoring of biological diversity in space and time. *Trends in Ecology & Evolution* **16**:446–453.