AVIAN CONSERVATION & ECOLOGY

VOLUME 12, ISSUE 2, ARTICLE 2

Venier, L. A., M. J. Mazerolle, A. Rodgers, K. A. McIlwrick, S. Holmes, and D. Thompson. 2017. Comparison of semiautomated bird song recognition with manual detection of recorded bird song samples. *Avian Conservation and Ecology* 12(2):2. https://doi.org/10.5751/ACE-01029-120202

Research Paper, part of a Special Feature on Advancing bird population monitoring with acoustic recording technologies

Comparison of semiautomated bird song recognition with manual detection of recorded bird song samples

Lisa A. Venier [1], *Marc J. Mazerolle* [2], *Anna Rodgers* [1], *Ken A. McIlwrick* [1], *Stephen Holmes* [1] *and Dean Thompson* [1]
[1]Canadian Forest Service, Natural Resources Canada, Sault Ste. Marie, Ontario, Canada, [2]Centre d'étude de la forêt, Département des sciences du bois et de la forêt, Université Laval, Québec, Québec, Canada

ABSTRACT. Automated recording units are increasingly being used to sample wildlife populations. These devices can produce large amounts of data that are difficult to process manually. However, the information in the recordings can be summarized with semiautomated sound recognition software. Our objective was to assess the utility of the semiautomated bird song recognizers to produce data useful for conservation and sustainable forest management applications. We compared detection data generated from expert-interpreted recordings of bird songs collected with automated recording units and data derived from a semiautomated recognition process. We recorded bird songs at 109 sites in boreal forest in 2013 and 2014 using automated recording units. We developed bird-song recognizers for 10 species using Song Scope software (Wildlife Acoustics) and each recognizer was used to scan a set of recordings that was also interpreted manually by an expert in birdsong identification. We used occupancy models to estimate the detection probability associated with each method. Based on these detection probability estimates we produced cumulative detection probability curves. In a second analysis we estimated detection probability of bird song recognizers using multiple 10-minute recordings for a single station and visit (35–63, 10-minute recordings in each of four one-week periods). Results show that the detection probability of most species from single 10-min recordings is substantially higher using expert-interpreted bird song recordings than using the song recognizer software. However, our results also indicate that detection probabilities for song recognizers can be significantly improved by using more than a single 10-minute recording, which can be easily done with little additional cost with the automate procedure. Based on these results we suggest that automated recording units and song recognizer software can be valuable tools to estimate detection probability and occupancy of boreal forest birds, when sampling for sufficiently long periods.

Comparaison de la reconnaissance semi-automatisée de chants d'oiseaux avec des détections manuelles d'échantillons de chants d'oiseaux enregistrés

RÉSUMÉ. Les unités d'enregistrement automatisé sont de plus en plus utilisées pour échantillonner les populations fauniques. Ces instruments peuvent produire une grande quantité de données qui s'avèrent difficiles à traiter manuellement. Toutefois, les informations contenues sur les enregistrements peuvent être résumées à l'aide de logiciels de reconnaissance vocale semi-automatisée. L'objectif de notre étude était d'évaluer l'utilité des reconnaisseurs de chants d'oiseaux semi-automatisés pour produire des données utiles à la conservation et à l'application de mesures d'aménagement forestier durable. Nous avons comparé les données de détection générées par les experts ayant écouté les enregistrements de chants d'oiseaux collectés au moyen d'unités d'enregistrement automatisé avec les données obtenues au moyen d'un processus de reconnaissance semi-automatisée. Nous avons enregistré des chants d'oiseaux à 109 sites en forêt boréale en 2013 et 2014 à l'aide d'unités d'enregistrement automatisé. Nous avons élaboré des reconnaisseurs de chants pour 10 espèces grâce au logiciel Song Scope (Wildlife Acoustics) et chaque reconnaisseur a été utilisé pour balayer un jeu d'enregistrements qui avait aussi été écouté par un expert en identification de chants d'oiseaux. Nous avons utilisé des modèles d'occupation pour estimer la probabilité de détection associée avec chaque méthode. À partir de ces estimations de probabilité de détection, nous avons produit des courbes de probabilité cumulée de détection. Ensuite, nous avons estimé la probabilité de détection des reconnaisseurs de chants au moyen d'enregistrements multiples de 10 minutes pour une unique station et visite (35 à 63 enregistrements de 10 minutes dans chacune de quatre périodes d'une semaine). Nos résultats indiquent que la probabilité de détection de la plupart des espèces dans les enregistrements de 10 minutes est beaucoup plus élevée lorsque les enregistrements sont écoutés par un expert comparativement à l'utilisation d'un logiciel de reconnaissance de chants. Cependant, nos résultats montrent aussi que la probabilité de détection par les reconnaisseurs de chants peut être améliorée si on utilise davantage qu'un seul enregistrement de 10 minutes, ce qui peut aisément être fait, à faible coût, grâce au processus automatisé. À la lumière de ces résultats, nous pensons que les unités d'enregistrement automatisé et les logiciels de reconnaissance de chants peuvent être des outils utiles afin d'estimer la probabilité de détection et l'occurrence des oiseaux forestiers boréaux si l'on échantillonne durant des périodes suffisamment longues.

Key Words: *automated recording units; boreal forest birds; detection probability; point counts; song recognition; song recognizer software*

Address of Correspondent: Lisa A. Venier, 1219 Queen St E, Sault Ste. Marie, ON , Canada, P6A 2E5, lisa.venier@canada.ca

INTRODUCTION

Forest birds are effective indicators of forest ecological integrity and sustainable forest management. They are highly diverse and have the capacity to capture ecosystem processes, their occurrence can be measured effectively with standardized methods, their identification is relatively easy for skilled observers, and the rich information on their life history makes the interpretation of patterns more robust (Venier and Pearce 2004). Songbirds are most often sampled using point-count surveys (Rosenstock et al. 2002), where an observer records all birds heard or seen at a station for a specified period of time (Ralph et al. 1995). Automated recording units (ARUs) have been suggested, assessed, and implemented as a means of conducting surveys of birds and amphibians (Acevedo and Villanueva-Rivera 2006, Swiston and Mennill 2009, Goyette et al. 2011, Venier et al. 2012, Holmes et al. 2014, Sidie-Slettedahl et al. 2015, Leach et al. 2016). Comparisons of recordings with field observations have varied in their approaches and have reached different conclusions. For instance, Hutto and Stutzman (2009) concluded that automated recording units do not provide a cost-effective alternative to field point counts because they fail to observe a large proportion of the detections recorded by human observers in the field and are more expensive and time consuming to use. In contrast, several other studies argued that the value of automated recording units depends on their implementation and study objectives and that they can be both cost effective and effective at observing birds because of their ability to collect more data than field observers (Haselmayer and Quinn 2000, Hobson et al. 2002, Acevedo and Villanueva-Rivera 2006, Celis-Murillo et al. 2009, Venier et al. 2012).

ARUs offer many advantages. For instance, these devices can be left in the field unattended for long periods of time, they accumulate data that can be assessed repetitively or by multiple experts if necessary, and they do not require sending trained observers in the field during the bird breeding season (Hobson et al. 2002, Rempel et al. 2005). Recordings can be interpreted by one or a few experienced observers during the off-season reducing observer effect (Venier et al. 2012). In addition to increasing the sampling effort at each point, automated recording units can be deployed at any time of day, regardless of weather, and record simultaneously at a suite of sites. The deployment and retrieval of recorders can occur outside of the breeding season, making field work more flexible (Venier et al. 2012). Automatic recording units can generate massive amounts of recordings that would not be possible to obtain from point count surveys conducted by observers. This can improve potential detectability at a site, as well as improve estimates of detectability that can reduce bias in occupancy estimates (MacKenzie and Royle 2005, Bailey et al. 2007), but also incurs additional postprocessing costs associated with interpreting additional or longer recordings.

An alternative to having trained observers interpret recordings is to scan the recorded data using automated recognition software. Automated recognition of bird songs is currently a very active area of research (Kogan and Margoliash 1998, Briggs et al. 2012, Potamitis et al. 2014, de Oliveira et al. 2015, Katz et al. 2016). This research has been translated into practical, easy to use, song recognition tools such as Song Scope (Wildlife Acoustics Inc., Concord, MA, USA; I. Agranat, 2009, *unpublished manuscript*, https://wildlifeacoustics.com/images/documentation/Automatically-Identifying-Animal-Species-from-their-Vocalizations.pdf), Raven (Cornell Laboratory of Ornithology; Charif et al. 2006), or monitoR (Katz et al. 2016). In this study, we employed Song Scope software that features an algorithm based on Hidden Markov Models (HMM) devised to evaluate spectral and temporal features of individual syllables as well as how syllables are structured to form complex songs (Kogan and Margoliash 1998, Somervuo et al. 2006). Users can develop and validate automated recognizers for species of interest and apply them to generate lists of "suspected" positive identifications that can be subjected to postprocessing review. The review process does not require the same level of expertise as a complete audio interpretation of a recording of multiple species. Potamitis et al. (2014) found that an automatic species recognition approach could reduce the search time for an observer by up to 98%, but that the current recognition algorithms still produce many false positives and false negatives requiring postprocessing of the data (Wimmer et al. 2013).

Song Scope recognizers have been developed and tested by Wildlife Acoustics in an unpublished study where 37% of the target vocalizations were detected on new test data, with at least one vocalization detected on 74% of all target recordings with a false positive rate of 0.4% (I. Agranat, 2009, *unpublished manuscript*). Additional recognizers developed through Song Scope have been used in several published studies to identify bird species in recordings (Holmes et al. 2014, Zwart et al. 2014), and have been found to be particularly useful for detecting rare species from many hours of recordings (Holmes et al. 2014, Zwart et al. 2014). However, there is currently insufficient published data to fully evaluate the utility of the Song Scope recognizers.

In this paper, we compare the detection probability for a suite of 10 selected forest songbirds typically occurring in boreal forest habitats, between manually interpreted recordings and those processed using automated recognizers built using Song Scope software (Wildlife Acoustics, Maynard, MA). We hypothesized that (1) trained observers listening to 10-min recordings would have a greater ability to detect bird species than automated recognizers applied to the same 10-min recordings, but that (2) using multiple 10-minute recordings, i.e., using a week-long sample of recorded data, would increase the autorecognizer detection probability for a given species to a level equal or greater than that achieved by experts listening to 10-min recordings. In addition, we estimated the cumulative detection probability to provide recommendations for when song recognizers could have greater utility relative to manual interpretation.

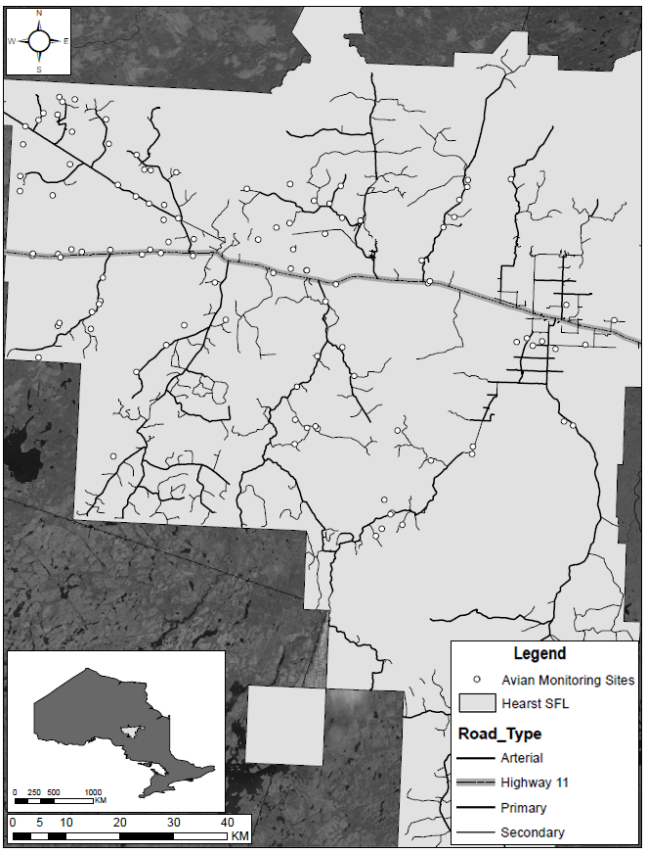
We used single-season site occupancy models to estimate the influence of the method (manual versus song recognizer) and site characteristics on the detection probability of each species for each period at each station (MacKenzie et al. 2002, 2006, Furnas and Callas 2015). We included a small suite of site characteristics in the model to account for potential heterogeneity in occupancy to meet the assumption that heterogeneity in occupancy is appropriately modeled with covariates. We chose variables that have been found to predict bird habitat use in boreal forests (Venier et al. 2005, 2007).

METHODS

Study area

The study was conducted within the Hearst Sustainable Forest Licence (Hearst SFL), which is an area of the boreal forest in northern Ontario of approximately 1.2 M ha size, centered at 49° 36'N and 83°39'W (Fig. 1). Our sampling focused on the north and central portions of the forest, which are dominated by flat clay and silty clay soils and predominantly lowland black spruce (*Picea mariana* Mill. B.S.P.) stands (Fig. 1). Black spruce is also the dominant tree species across the entire Hearst SFL, occupying approximately 67% of the overall forest by area. Such sites are typically characterized by moderately deep (20–40 cm) to deep (> 40 cm) organic soils over clays, with relatively poor drainage and low productivity. Nearly pure stands of black spruce compose approximately 34% of the forest land base or as mixtures with other conifers including larch (*Larix laricina* (Du Roi) K. Koch) and cedar (*Thuja occidentalis* L.), and less frequently with intolerant hardwood species including trembling aspen (*Populus tremuloides* Michx.) and balsam poplar (*Populus balsamifera* L.).

Fig. 1. Location of sampling sites where automated recording units were deployed to detect forest bird species in 2013 and 2014 at 108 stations in Hearst Forest Sustainable Forest License, Ontario, Canada.



We randomly selected 109 spruce-dominated lowland stands. These stands spanned across gradients of age and vertical structural composition (Fig. 1) based on the following criteria: (a) ≥= 50% black spruce in the overstory, (b) ≥ 10 ha in area, and (c) within 2 km of a secondary or tertiary forest road where access to the stand was not blocked by a feature such as a river or large swamp. Potential sampling sites were identified based on spatial analysis of available forest inventory data, road and water layers provided by Hearst Forest Management Inc. on an ArcGIS platform (ArcMap 9.3, ESRI 2008). Some potential sites were rejected or moved as plots were being established because of unforeseen issues with stand access including roads being flooded by beaver activity or culverts being washed out during spring runoff events. Overall, we sampled 74 stands in 2013 and 35 different stands in 2014. A single bird sampling station was positioned in each stand at least 150 m from the edge of the stand with a minimum of 350 m between stations in any one year. Using georeferenced forest inventory data, we measured the proportional cover of black spruce within a 100 m radius circle around each sampling point. We used airborne Light Detection and Ranging (LiDAR) data with an average sampling density of 1.1 returns/m² that were acquired between 4 July and 4 September 2007 (Table 1) to capture metrics of vegetation structure (Table 2; see Pitt et al. 2014 for LiDAR sampling details). Airborne LiDAR data across the entire study area were subdivided into a 400-m² (20 m × 20 m) grid. Height distribution statistics for the point-clouds in each grid cell were then calculated using all returns, without any height threshold to filter point data. We identified five vegetation structure metrics to include in our models as shown in Table 2.

Table 1. Light Detection and Ranging (LiDAR) acquisition specifications for the Hearst Forest.

Parameter	Description
Sensor	Leica ALS50
Platform	Cessna 310
Pulse Rate	119,000 Hz
Scan Rate	32 Hz
Field of view	30°
Flying height	2400 m
Track spacing	1000 m
Overlap	20%
Vertical accuracy	< 30 cm
Return density	1.1/m²

Table 2. Light Detection and Ranging (LiDAR) metrics used to characterize forest vertical structure within sample plots (100 m radius circles) in each of 109 lowland black spruce stands.

Variable	Description
P90	Ninth decile (M) of vegetation returns-index of stand age
P90_SD	Index of canopy height heterogeneity
Veg.less4m	The number of vegetative returns in the < 4 m level
Veg.great4m	The number of vegetative returns in the > 4m level
Veg.less2m	The number of vegetative returns in the < 2m level

Collecting bird recording data

Automated recording units (SM2; Wildlife Acoustics, Inc., Concord, MA) were placed at each sampling station. Recording

units were set to sample at 24000Hz in stereo using the wav format. The gain was set to factory defaults. In 2013, recorders were programmed to record for 10 minutes during five periods each day: half an hour before sunrise, at sunrise, as well as 30 min, 1.5 h, and 3.5 h following sunrise. The recorders collected 50 minutes of data per day for 29 or 30 days from 5 June to 3 or 4 July 2013 (Table 3). This sums to approximately 24 h of data per site or 1896 h of recorded data across all 74 sites sampled in 2013. In 2014, a total of 90 min were captured daily at each site with the same schedule as 2013 plus additional 10-min recordings: 1 h, 2 h, 2.5 h, and 3 h following sunrise. Recordings in 2014 started on 11 June and continued to 13 or 14 July for a total of approximately 47 hours per site and 1645 hours of recorded data across all 35 sites sampled in that year (Table 3). We chose 10-minute recording times as recommended by Howe et al. (1997). The timing of the recordings was chosen based on our knowledge of singing frequency and in an attempt to capture the span of time when the vast majority of singing takes place during the day. Recordings were made in the breeding season for forest passerines in our study area. We waited until June to sample to reduce the influence of nonterritorial birds on the observations. We conducted comparisons between manual and semiautomated detection for 10 bird species representing a range of common and rarer species typical of boreal forest bird communities.

Table 3. Sampling periods considered for the analysis of recording data collected from automated recording units deployed at 109 stations in Hearst Forest Sustainable Forest Licence, Ontario, Canada.

Data set	Year	Visit 1	Visit 2	Visit 3	Visit 4
Single 10-minute recordings	2013	5 June	14 June	24 June	2 July
	2014	12 June	18 June	25 June	2 July
Recordings pooled across week	2013	5-11 June	12-18 June	19-25 June	26 June-2 July
	2014	12-18 June	19-25 June	26 June-3 July	4-14 July

Building song recognizer

Each species recognizer was built separately using a suite of high quality training recordings without background noise such as rain, cars, or other birds. Song recognizers were parameterized in a process guided by the Song Scope documentation (Wildlife Acoustics Inc.), but also included a substantial amount of trial and error using both cross-training statistics and results from scanning of test data. Our protocol for building recognizers, the parameter settings for each recognizer, the cross-validation statistics for final models, the minimum score and quality settings for running recognizers and the source of cross training data are presented in Appendix 1. Song Scope Recognition (SSR) files were created and used within the Wildlife Acoustics Song Scope Software to recognize individual species songs (Appendix 2). See Appendix 1 for a description of the development and use of SSR files, the source data for training recognizers, the parameter settings for our song recognizers, and performance estimates..

Comparing manual vs song recognizer on single 10 minute recordings

We chose a single 10-minute recording taken half an hour after sunrise for each of four dates within the breeding season of each bird species. These four recordings at each station were processed using the manual and song recognizer approaches (Table 3). Thus, four recordings for each of 109 stations (436 recordings) were interpreted manually and scanned using a song recognizer. Dates were chosen to be evenly spaced throughout the breeding season. Recordings that contained excessive noise from wind or rain that obscured the audio signal were substituted by those from the next or previous day.

Manual interpretation of the recordings consisted of viewing and listening to all 10 minutes of each recording using spectrogram software (Song Scope, Wildlife Acoustics, Inc., Concord, MA), and noting each unique species heard or seen on the spectrogram. Only the first observation within the 10-minute recording of each species was noted. This process produced a single datum (detection or nondetection) for each of four separate days for each station and each species. Recordings were interpreted by one of two technicians very experienced in conducting auditory bird surveys. Both technicians have more than five years of experience conducting bird surveys and interpreting recordings in a professional capacity. They both found that using the spectrograms in conjunction with listening improved their ability to detect species.

The song recognizer processing of the recordings consisted of scanning each of the same four recordings per station used on the manual detection (four recordings on four separate days), with a song-recognizer built by one of the authors (AR) for each of the 10 species (Table 4). Song recognizer scanning of all 436 recordings (109 stations x 4 recordings = 4360 minutes) was conducted overnight in a batch scan for a single species at a time. This scanning process creates a results file of potential positives for the target species that includes the identity of the source data, i.e., which recording the hit came from, and links it to the location of the hit. This location on the recording must then be validated by a technician by listening to the audio signal and visually examining the spectrogram signal. This is a relatively simple recognition task that does not require expert level song recognition and takes only a few seconds per hit. The hits are examined sequentially and as soon as a true positive is validated then the bird is noted as present and the postprocessing is stopped.

In this study, we were interested in the ability of the recognizer to identify the presence of the target bird on each 10-min recording. Each recognizer was built separately using a suite of high quality recordings of the species of interest and specifications for the recognizer are adjusted to optimize the cross-validation statistics (see Appendix 1 for more details). The song recognizer processing yielded a single datum consisting of a detection or nondetection for each of the 10 bird species, for each of four recordings at each sampling station.

Song recognizer performance with multiple recordings pooled across week

One of the biggest advantages of song recognition software is the low cost incurred with scanning additional recordings. To quantify the gain in information from using multiple recordings, we used our recognizers to scan the complete recordings from the breeding season at each station in each year. This included five 10-min recordings each day in 2013 and nine 10-min recordings each day in 2014. We divided the recordings into four (week long)

Table 4. Species list with number of detections in recordings out of 436 and in stations out of 109 for each of (1) manual detection in single 10-minute recordings, (2) song recognition in single 10-minute recordings, and (3) song recognition in each pooled week of 10-minute recordings. Values in parentheses represent the number of occupied stations based on four recordings.

Species Code	Species Common Name	Species Latin Name	Number of recordings (stations)[†] with manual detections: based on single 10 min recording	Number of recordings (stations) with recognizer detections: based on single 10 min recording	Number of weeks (stations) with recognizer detections: based on 35 to 63 ten minute recordings
BBWA	Bay-breasted Warbler	*Setophaga castanea*	22 (14)	18 (11)	84 (36)
BRCR	Brown Creeper	*Certhia americana*	19 (15)	6 (6)	70 (39)
GCKI	Golden-crowned Kinglet	*Regulus satrapa*	161 (68)	84 (49)	262 (78)
HETH	Hermit Thrush	*Catharus guttatus*	230 (95)	86 (55)	302 (96)
LISP	Lincoln's Sparrow	*Melospiza lincolnii*	42 (17)	33 (14)	48 (16)
MOWA	Mourning Warbler	*Geothlypis philadelphia*	34 (17)	18 (9)	18 (7)
NOWA	Northern Waterthrush	*Parkesia noveboracensis*	32 (16)	13 (7)	33 (15)
RCKI	Ruby-crowned Kinglet	*Regulus calendula*	244 (95)	119 (68)	296 (93)
SWTH	Swainson's Thrush	*Catharus ustulatus*	164 (79)	33 (28)	187 (74)
YBFL	Yellow-bellied Flycatcher	*Empidonax flaviventris*	186 (81)	95 (52)	237 (85)

[†]Four recordings per station.

time windows matching the four dates that were used for the individual scans of 10-min recordings in the previous section (Table 3). These week-long sampling periods included either 35 or 63 10-minute recordings depending on year. For each time window and station, the multiple recordings were scanned and then postprocessed to validate "hits" or positives as described in the previous section. This yielded the detection data (a single presence or absence) for each of four time windows, 109 stations, and 10 species. We conducted a graphical comparison of these detection probabilities to the manual detection probabilities of 10-minute recordings.

Statistical analyses

Data sets

We created two data sets for analysis. The first consisted in combining the detection data from the manual and song recognizer processing of 10-min recordings into detection histories for each station. Thus, each station had two observations (one for manual, one for song recognizer) for each visit (recording). The second data set consisted solely of the multiple recordings pooled across each of four weeks. In both cases, the data are in a format amenable to site occupancy analysis.

Comparison on individual 10-minute recordings

We used single-species, single-season site occupancy models (MacKenzie et al. 2002, 2006) to estimate the influence of the method and five site characteristics on the detection probability and occupancy of each species for each period at each station. This model type uses detection data to estimate parameters on the probabilities of detection and occupancy. The main assumptions of the model include the following: (1) the occupancy state remains static between the first and last visit (no extinctions or colonizations during the study), (2) the heterogeneity in detection probability and occupancy is appropriately modeled with covariates, (3) detections at a given station and visit are independent, and (4) there are no false positives (species misidentification). We are confident that model assumptions were met because we collected the data during the breeding season of the species, we used covariates to model heterogeneity, and an experienced technician postprocessed the detection data to remove false-positives. Each of the 10 species was analyzed separately.

We used a model selection and multimodel inference framework to assess the influence of different covariates on detection probability as well as occupancy of a given species (Burnham and Anderson 2002, Mazerolle 2006). We included the effect of site characteristics on occupancy to account for potential heterogeneity in occupancy to meet model assumptions, although we do not examine the occupancy results in this paper. We conducted two series of analyses. The first involved a comparison of the detections from the single 10-minute recordings (four 10-min recordings on each of four separate days for each station) using the manual and semiautomated approaches. To do so, we tested eight hypotheses on the detection probability involving the method, the visit, the vegetation structure, and the stand age (Table 5). We tested three hypotheses on the occupancy probability involving the effects of the proportion of black spruce cover, the vegetation structure, as well as the stand age and canopy height heterogeneity (Table 6). We built models for each scenario on detection probability and occupancy, yielding a total of 24 candidate models. We computed model-averaged predictions based on the observed explanatory variables.

Cumulative detection probabilities from individual 10 minute recording data

We calculated cumulative detection probability estimates for each species. The first four estimates (1–4 visits) were computed using our data and models. The later estimates (> four visits) were computed using average detection probabilities from the first four visits; we computed the cumulative detection probability of detecting a given species for different scenarios of the number of visits using the equation, $1 - (1 - p)^t$, where p is the probability of detecting the species conditional on its presence at a station during a single visit, and t is the number of visits.

Table 5. Biological hypotheses tested on detection probability (p) and occupancy (ψ) of forest birds from the detection data obtained from four 10-min recordings in 2013 and 2014 at 109 stations in Hearst Forest Sustainable Forest License, Ontario, Canada.

Parameter Model structure	Biological hypothesis
Detection probability	
p(.)	constant detection probability
p(Method)	detection probability is greater with the manual method than the semiautomated method
p(Year + Visit)	detection probability varies with year and visit
p(Year + Visit + Method)	detection probability varies with year, visit, and method
p(Method + Veg.less2m)	detection probability varies with method and decreases with increasing vegetation structure below 2 m (additive effects)
p(Method + Veg.less2m + Method: Veg.less2m)	detection probability of a method depends on vegetation structure below 2 m (interactive effects of method and vegetation structure below 2 m)
p(Method + Stand.age)	detection probability varies with method and stand age (additive effects)
p(Method + Stand.age + Method: Stand.age)	detection probability of a given method depends on stand age (interactive effects of stand age and method)
Occupancy	
ψ(Spruce.cover)	occupancy probability varies with spruce cover in the stand
ψ(Veg.less4m + Veg.great4m)	occupancy probability varies with vegetation structure below 4 m and above 4 m
ψ(Stand.age + Canopy.diversity)	occupancy probability varies with stand age and canopy diversity

Table 6. Biological hypotheses tested on detection probability (p) and occupancy (ψ) of forest birds from the detection data obtained from recordings spanning each of the four weeks in 2013 and 2014 at 108 stations in Hearst Forest Sustainable Forest Licence, Ontario, Canada.

Parameter Model structure	Biological hypothesis
Detection probability	
p(.)	constant detection probability
p(Year + Visit)	detection probability varies with year and visit
p(Veg.less2m)	detection probability decreases with increasing vegetation structure below 2 m
p(Stand.age)	detection probability varies with stand age
Occupancy	
ψ(Spruce.cover)	occupancy probability varies with spruce cover in the stand
ψ(Veg.less4m + Veg.great4m)	occupancy probability varies with vegetation structure below 4 m and above 4 m
ψ(Stand.age + Canopy.diversity)	occupancy probability varies with stand age and canopy diversity

Song recognizer performance with multiple recordings pooled across week

The second analysis focused exclusively on the semiautomated approach and involved the detections obtained from the data spanning each week (based on 35 to 63, 10-min recordings for each of four time periods for each station). Here, we tested four hypotheses on detection probability involving vegetation structure and stand age, as well as three hypotheses on the probability of occupancy identical to the ones for the first analysis (Table 6). We did not consider the "method" variable in this second exercise because only data from the song recognizers were used for this part. We considered a total of 12 candidate models in the second analysis.

Modeling protocols for both data sets

We standardized all numerical variables before entering them in models. We checked the correlations between variables and never included variables with ($|r| > 0.7$) in the same model. Models were fit with maximum likelihood estimation in the unmarked package for R 3.3.0 (Fiske and Chandler 2011, R Core Team 2016). Model selection and multimodel inference based on the second-order Akaike information criterion (AIC_c) was implemented with the AICcmodavg package (Mazerolle 2016). The ratio of observations to the number of estimated parameters was < 40 therefore we used the small sample AIC (aka AICc; Burnham and Anderson 2002).

We checked the fit of the top-ranking model for each species with the MacKenzie and Bailey (2004) goodness of fit test with 10,000 iterations. When there was evidence for overdispersion (i.e., c-hat > 1), we used the $QAIC_c$ for our inferences and adjusted the standard errors by multiplying with the square-root of the overdispersion estimate.

RESULTS

Comparison on individual 10-minute recordings

Based on the 10-min recordings with manual detection, the 10 species that we evaluated were detected at between 13% (14/109) and 87% (95/109) of the stations (Table 4). Detection patterns agreed between 67 and 99% of the 436 survey periods, between the manual method on 10-min recordings and of the song recognizer on the 10-min recordings, depending on the species (Table 7). The manual detection method often detected the species when the recognizer (on the 10-minute recording) did not, but the converse, i.e., the recognizer detecting the species when the manual method did not, was not common (Table 7). Overall agreement between methods was strongly influenced by instances of no-detection using either method. This represents an important indicator of the postprocessing effectiveness, however, because false positives are relatively common using recognizers and postprocessing appears to be effective in distinguishing between false and true positives. Without consideration of the 0-0 case, detection patterns agreed between 19 and 82% of the time between the manual method on 10-min recordings and of the song recognizer on the 10-min recordings (Table 7).

Table 7. Summary of detection (1) and nondetection (0) patterns for two comparisons. The first compares manually interpreted 10-minute recordings against 10-minute recordings processed by a recognizer. The second comparison contrasts manually interpreted 10-minute recordings against week-long compiled recordings (between 350 and 630 10-minute recordings). In all three methods (10-minute manual, 10-minute recognizer, and week-long recognizer) we have 436 observations of presence or absence. (0-0 = true negatives for both methods, 0-1 = false negative with manual approach and true positive for recognizer, 1-0 = true positive for manual approach and false negative for recognizer, 1-1 = true positive for both methods). As a measure of agreement between methods, we calculated accuracy and sensitivity. Accuracy was computed as the number of true positives and true negatives (0-0 or 1-1) divided by the total number of observations (109 sites x 4 visits = 436), whereas sensitivity was estimated as the number of true positives (1-1) divided by the number of true positives (1-1) and false negatives (0-1). See Table 4 for species codes.

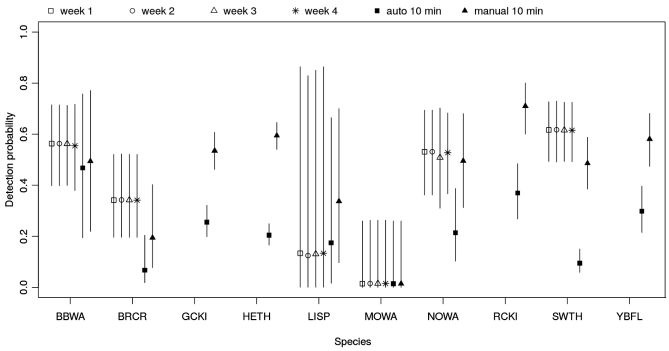
Species	Manual 10 min vs Recognizer 10 min						Manual 10 min vs Recognizer (35-63 10-min recordings)					
	0-0	0-1	1-0	1-1	Agreement (accuracy)	Agreement (sensitivity)	0-0	0-1	1-0	1-1	Agreement (accuracy)	Agreement (sensitivity)
BBWA	414	0	4	18	0.99	0.82	349	65	3	19	0.84	0.22
BRCR	416	1	14	5	0.97	0.25	357	60	9	10	0.84	0.13
GCKI	274	1	78	83	0.82	0.51	160	115	14	147	0.70	0.53
HETH	206	0	144	86	0.67	0.37	91	115	43	187	0.64	0.54
LISP	393	1	10	32	0.97	0.74	379	15	9	33	0.94	0.58
MOWA	402	0	16	18	0.96	0.53	397	5	21	13	0.94	0.38
NOWA	405	0	18	13	0.96	0.42	390	15	13	18	0.94	0.39
RCKI	188	4	129	115	0.69	0.46	103	89	37	207	0.71	0.62
SWTH	271	1	132	32	0.69	0.19	190	82	59	105	0.68	0.43
YBFL	248	2	93	93	0.78	0.49	158	92	41	145	0.69	0.52

The analysis (single season site occupancy models) of the single 10-min recordings, (one recording collected on each of four separate visits) suggested adequate model fit for all 10 species with little to no overdispersion (Appendix 3). For every species except BBWA (Bay-breasted Warbler, *Setophaga castanea*), the detection probability component of the top-ranked model included the effect of method (manual vs recognizer), either alone or with the additive effects of stand age, year, visit, or vegetation structure < 4 m (Appendix 3). For all species except BBWA and MOWA (Mourning Warbler, *Geothlypis philadelphia*), the manual detection method had a greater detection probability than the semiautomated approach alone (comparing solid shapes within species in Fig. 2, Table 8). Detection probability also varied with stand age and visit for certain species (Fig. 3, Table 8).

Cumulative detection probabilities from individual 10-minute recording data

Based on the cumulative detection probability of 10-min recordings, as an example, for Bay-breasted warbler, as few as five 10-min recordings processed with the song recognizer yielded a 90% chance of detecting the species at least once (Fig. 4). In contrast, semiautomated processing of at least 26 10-minute recordings of MOWA are required to yield comparable cumulative detection probabilities to manually processing four 10-minute recordings. At least five species were predicted to require six or more visits to reach an 80% detection probability with song recognizers. With the manual approach using 10-min recordings, most species are predicted to reach a 90% chance of being detected after only five visits (based on the equation 1-(1-p)t to predict cumulative detection probabilities) and all but three species is predicted to need no more than three visits to reach an 80% detection probability.

Fig. 2. Detection probability of forest birds based on single 10-min recordings and multiple recordings pooled within each week collected with automated recording units in 2013 and 2014 at 109 stations in Hearst Forest Sustainable Forest License, Ontario, Canada. Solid symbols indicate detection probability based on a single 10-minute recording (triangles represent manual detection and squares represent automated song recognition). Open symbols indicate estimated detection probability for multiple recordings pooled into four separate weeks. These symbols represent detection probability based on 35 to 63 10-minute recordings). Error bars represent 95% confidence intervals. See Table 4 for species codes.



Song recognizer performance with multiple recordings pooled across week

Based on the pooled weekly data species were detected at between 6% (7/109) and 88% (96/109) of the stations (Table 4). Detection patterns agreed between 64 and 94% of the time (Table 7). The

Table 8. Summary of detection (1) and nondetection (0) patterns by species comparing manual 10-minute approach (a single recoding) to each of recognizer 10-minute approach (a single recording) and recognizer week-long approach (between 35 and 63 10-minute recordings). (0-0 = not detected with either method, 0-1 = not detected with manual approach but detected with recognizer, 1-0 = detected with manual approach but not detected with recognizer, 1-1 = detected with both methods). See Table 4 for species codes.

Species	Manual 10 min vs Recognizer 10 min				Manual 10 min vs Recognizer (35-63 10-min recordings)			
	0-0	0-1	1-0	1-1	0-0	0-1	1-0	1-1
BBWA	414	0	4	18	349	65	3	19
BRCR	416	1	14	5	357	60	9	10
GCKI	274	1	78	83	160	115	14	147
HETH	206	0	144	86	91	115	43	187
LISP	393	1	10	32	379	15	9	33
MOWA	402	0	16	18	397	5	21	13
NOWA	405	0	18	13	390	15	13	18
RCKI	188	4	129	115	103	89	37	207
SWTH	271	1	132	32	190	82	59	105
YBFL	248	2	93	93	158	92	41	145

Fig. 3. Effect of stand characteristics on detection probability of forest birds based on 10-min recordings collected with automated recording units in 2013 and 2014 at 109 stations in Hearst Forest Sustainable Forest License, Ontario, Canada. See Table 4 for species codes.

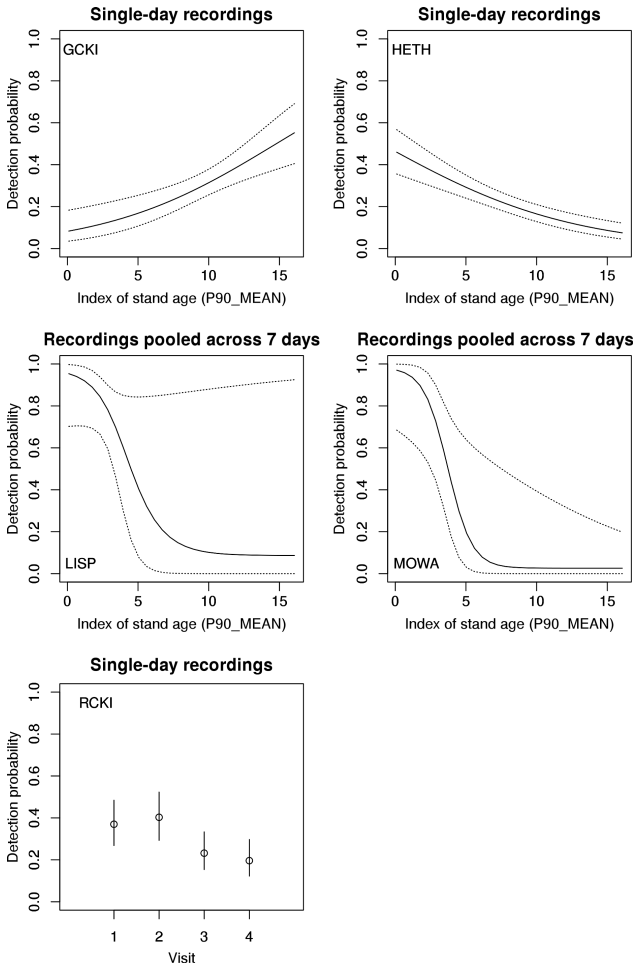
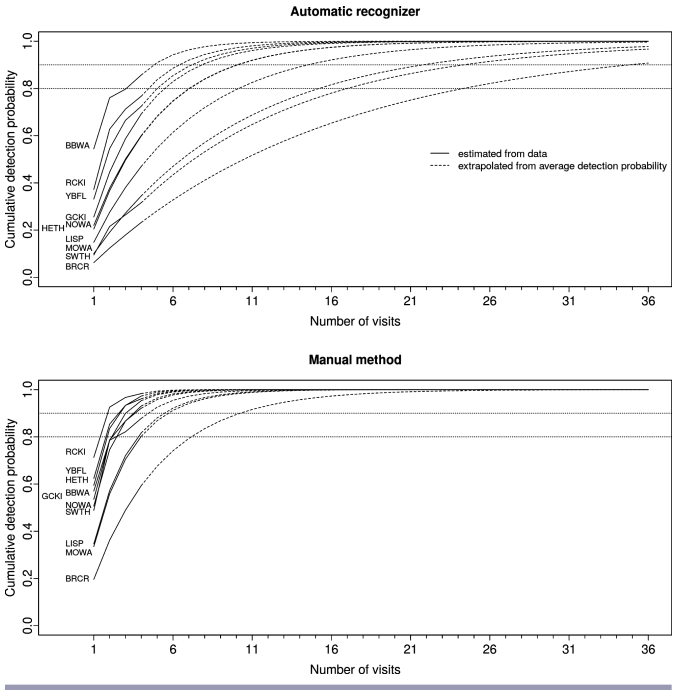


Fig. 4. Cumulative probability of detecting the species at least once with a given number of visits computed from model-averaged estimates of detection probability for the automatic song recognizer and manual methods. The first four cumulative detection estimates were computed using the estimates for each visit based on the data and models (solid line). The later visits (dotted line) were computed using the equation 1-(1-p)^t, where p was the average detection probability derived from for the first four visits. Although we processed 350 to 630 minutes of recordings using the automated recognizers, we could not use this data to build cumulative detection past 4 visits because the recorded sample was processed as a single long recording rather than individual 10-minute recordings. Each species name appears next to its curve. See Table 4 for species codes.



recognizer detection on pooled weekly data often detected the species when the manual interpretation on the 10-minute recording did not indicate true positives for the recognizer approach and false negatives for the manual approach, but the converse was much less frequent (Table 7). In a classic confusion matrix, one observation is considered truth and the other is observed. In our case, we assume that if the species is observed by either method then it is actually present. Either method can miss species so we do not assume that one is more correct than the other. Thus a 0-1 case is a false negative for the first approach and a true positive for the second approach. A 1-0 case is a true positive for the first approach and a false negative for the second approach. Agreement between methods in this comparison (manual on 10-minute recording vs recognizer on week-long recording) was generally lower, mostly because of improved detection rates of the song recognizer method when more recorded data are used.

For the recording data pooled for the week exclusively based on the song recognizer technique, we found substantial lack of fit of the models for GCKI (Golden-crowned Kinglet, *Regulus satrapa,* $X^2 = 60.533$, $P < 0.0001$), HETH (Hermit Thrush, *Catharus guttatus*, $X^2 = 54.526$, $P < 0.0001$), RCKI (Ruby-crowned Kinglet, *Regulus calendula*, $X^2 = 119.13$, $P = 0.0001$), and YBFL, Yellow-bellied Flycatcher, *Empidonax flaviventris*, $X^2 = 82.67$, $P < 0.0001$). Thus, we excluded these species from subsequent analysis. Among the six species remaining for analysis, there was considerable model selection uncertainty regarding the variables included on detection probability and occupancy of forest birds, revealing weak or no effects (Appendix 4). We found evidence of a negative effect of stand age on detection probability of MOWA (model-averaged shrinkage estimate: -4.24, 95% CI: -7.24, -1.23). We found a similar but weaker effect of stand age on LISP (Lincoln's Sparrow, *Melospiza lincolnii*, model-averaged shrinkage estimate: -2.79, 95% CI: -5.92, 0.33). Based on visual inspection of Figure 2, the detection probability of most species was slightly higher with data pooled within each week, than that using the manual approach based on the single 10-minute recording.

DISCUSSION

Our first hypothesis was that the manual approach has higher detection probabilities than the song recognizer approach for the same 10-minute recordings. The results provide no evidence to reject the hypothesis. The manual detection approach had a greater detection probability than the song recognizer for six out of 10 species investigated. These results indicate that for an equivalent sampling time, i.e., a 10-minute recording), song recognizers are inferior to an experienced human interpreter. This result is not surprising because there has been some skepticism about the ability of these algorithms to consistently recognize songs of individual species as effectively as a human observer (Swiston and Mennill 2009, Goyette et al. 2011; *personal observation*).

Creating a generalized song recognition algorithm for real-world field conditions is a complex task. Our recordings included singing and calling from multiple individuals and species, with overlapping songs, and background noise including other nonavian species, wind, rain, and traffic. Thus, the signal-to-noise ratio in the data in some cases could be weak. To deal with these issues, the Song Scope algorithms preprocess the recorded data to reduce the effects of noise (I. Agranat, 2009, *unpublished manuscript*). Also, songs from individuals within species can be quite variable. As a result, the algorithm must be flexible enough to recognize individuals that did not form part of the training set. We have found that the trade-off with flexibility is the generation of false positives, whereby individuals from other species are misclassified as the target species. Our song recognition approach accepts high levels of false positives in the interest of generating fewer false negatives. Our postprocessing protocol effectively removes the false positives with limited cost in time and effort. Because of the requirement of post-processing, we term this approach semiautomated.

Our results highlight that the effectiveness, i.e., detection probability, of the song recognizers varies among different species. Song recognizer detectability for the 10-minute recordings ranged from less than 0.1 for BRCR (Brown Creeper, *Certhia americana*) to as high as 0.5 for BBWA. Narrow-band whistled vocalizations, lacking any distinctive spectral features, are expected to be difficult to identify, whereas broadband vocalizations with complex spectral properties should be easier. Longer vocalizations are also expected to be more easily recognized because of their greater information content. There is, however, also a lot of variability in the manual detection probabilities, where detectability ranges from 0.2 to 0.7. Investigators should estimate detection probability explicitly whenever detection probability of any method is lower than 1 (Mazerolle et al. 2007, Williams et al. 2002). This is a strong argument for the use of sampling protocols that can estimate detectability (repeated measures) and yield useful estimates of resource use, occupancy, or species richness, such as protocols using automated recording units.

The comparison of 10-min recordings quantifies the gap in detectability between the two methods. Manual detection is much better than song recognition when based on a 10-minute recording. However, this is not the most relevant comparison because although 10-minute counts are regularly used to assess bird presence and abundance (Ralph et al. 1995, Howe et al. 1997, Venier and Pearce 2005, 2007), recording and automated recognition protocols would be unlikely to use a single 10-minute recording to assess the presence of a species at a site (e.g., Holmes et al. 2014). One of the most significant advantages of the recording and automated song recognition approach is the ability to collect and process large amounts of recorded data at a site with very little additional cost compared to collecting and processing a single 10-minute sample. Although computer processing time is directly related to the amount of recorded data being used, the semiautomated song recognizer approach greatly reduces the time required for a technician to review the recordings for suspected detections and remove false positives. A better comparison would have been to interpret all of the recorded data and compare that to the song recognizer interpretations from the same data although we did not have the resources to make that comparison. As an alternative, we developed our second hypothesis that compared single 10-minute recordings using manual interpretation with multiple 10-minute recordings using song recognition software.

Our second hypothesis was that increasing the length of the recording (in this case to 350 to 630 minutes over seven days)

increases the detection probability of the species by the song recognition software to equal or surpass that from the 10-min recordings processed by trained listeners. The data for four species yielded models that lacked fit and were removed from analysis. Of the remaining six species, three had higher detection probability using recognizers on multiple recordings from entire weeks (BRCR, NOWA [Northern Waterthrush, *Parkesia noveboracensis*], SWTH [Swainson's Thrush, *Catharus ustulatus*]) than the manual detection on single 10-min recordings. In contrast, LISP, BBWA, and MOWA detection probability did not improve using additional recording time. For some species, pooling recordings within each week greatly improved the detection probability of the song recognizers to similar or greater levels as the manual method from a single 10-minute recording.

We found that using recognizers on more recorded data provides much higher detection probabilities than using recognizers on a single 10-minute recording and often produced similar or higher detection probabilities than the manual technique. Other studies have found that increasing survey time, which is strength of (automated) recorded data, can produce more detections than using data acquired in the field by technicians (Venier et al. 2012, Klingbeil and Willig 2015). There are many advantages to using a recording protocol rather than field observations, especially the ability to collect much more auditory data (Venier et al. 2012), but it provides less information on abundance. Using automated recognizers provides only presence/absence data, while manual detection may allow for some minimal abundance data to be collected. Once a decision has been made to use recorded data, however, the use of automated processing over manual detection allows the processing of orders of magnitude more recorded data in the same time frame.

Four species were excluded from the analysis of the week-long recording data. Although being detected at a large number of sites (72%–88%), these species yielded overdispersed detection data, as c-hat estimates were much larger than 4. The lack of fit was mainly due to an underrepresentation of certain detection histories relative to those expected under the model, namely those with three detections across four visits (0111, 1011, 1101, 1110). This lack of fit could stem from heterogeneity in detection probabilities or occupancy not explained by the variables at hand. As with any statistical technique, this also highlights the importance of assessing model fit for each species before making inferences instead of applying the technique to every species without discernment.

Comparisons of manual detection vs automated recognition approaches are relatively rare (Swiston and Mennill 2009, Goyette et al. 2011, Towsey et al. 2012, Stowell and Plumbley 2014). When comparisons are made based on a paired approach where the same recording is analyzed by the different methods, the manual interpretation is consistently better and usually used as the benchmark against which automated identification is measured (Goyette et al. 2011, Towsey et al. 2012; our data). At an even finer scale, comparisons are sometimes made at the level of individual songs (Goyette et al. 2011). This may be important if one is interested in measuring something like song rate or other song metrics, i.e., where correctly identifying individual songs is the goal (Swiston and Mennill 2009). At this point in the technology development, however, it is not possible to consistently recognize all songs of a target species. In a study of nocturnal tropical birds, Goyette et al. (2011) found sensitivity (the proportion of known calls of target species identified by recognizers) ranging from 0.17 to 0.79 and a positive predictive value (the proportion of detected sounds that corresponded to target species) ranging from 0.39 to 0.60.

Song recognizers can perform effectively when the objective is to identify the presence of a species at a site based on a minimum amount of recorded data. Based on our results, we can conclude that 10 minutes is not enough but 350–630 minutes is probably more than enough. The cost of using a song recognizer on multiple recordings is minimally more than for a single recording. Thus, the reduced detection probability of the recognizers compared to the manual method can be offset by the capacity of the recognizers to deal with more data relatively quickly. In addition, where the goal is to identify rare species, it will be advantageous to be able to sample much longer than what can be reasonably accomplished with manual detection (Swiston and Mennill 2009, Holmes et al. 2014). The lower detection probability for the automatic recognizer can be mostly offset by sampling fewer than 30 recordings (300 minutes) for even the poorest recognizers that we examined.

We expect that the ecology of species will influence their detectability. We observed large differences in detectability by species ranging from less than 10% to 70% for a single 10-minute count. There are a number of ecological factors that are expected to influence detectability including differences in habitat preferences, habitat use, abundance, song behavior, and song phenology (McShea and Rappole 1997, Alldredge et al. 2007, Royle and Nichols 2003). The cumulative detection probability curves suggest that differences in detection probability are greater for the automated recognizer method than the manual detection method, but that increasing the amount of recorded data can greatly reduce differences between methods. But our primary interest in estimating detection probability here is to compare methods with the assumption that increasing probability of detection increases the quality of our data.

Based on our experience, the time required for the technician to postprocess (check for false positives and confirm the true positives) for 1400–2520 (35–63 recordings x 10 minutes x 4 visits) minutes of recordings per station per species ranges from around 1 minute/site when detection probability for the species is high to 20 minutes when the species is not detected at the site. Rare or absent species take longer to process because the technician is required to review all of the false positives in the results files. In contrast, common species appear early in the results file and the postprocessing can stop after confirmation of the technician of the first detection. The average human processing time is 5.7 minutes/species/station for the 1400–2520 minutes of recordings. Manual interpretation of 40 minutes of recordings to assess the entire community takes about 1 hour per station. The time to assess 10 species semiautomatically is approximately equivalent to the time taken for the manual approach to assess the entire community.

The time needed to process recorded data with recognizers is largely affected by the number of species considered. If the goal is to monitor all species in the community, then recognizers will probably not be an efficient choice because many forest bird

communities have more than 40 species. On the other end of the spectrum, conservation work focused on a single species will be quicker to implement with recognizers. (Swiston and Mennill 2009, Holmes et al. 2014, Zwart et al. 2014). Rare or individual species detection is clearly a very effective use of recognizers because it allows the sampling of much more recorded time that increases the probability of detection necessary for uncommon species. In between these two approaches is the potential use of recognizers to monitor a suite of indicator species that may provide enough community information necessary to assess sustainability or ecological integrity. Some studies have suggested an indicator framework to select such a suite of species that could act as a bioassay to evaluate the sustainability of forest management (Venier and Pearce 2004, Rempel 2007). For this purpose, Rempel (2007) proposed a suite of 13 songbirds capturing a broad range of habitat conditions to represent the boreal songbird community in Ontario. For this type of application, song recognizers may be a more efficient approach for acquiring data. Another relevant factor is the lack of expert interpreters to process the data manually. Some organizations including Canadian national parks have a requirement to monitor ecological integrity, but do not necessarily have qualified staff to complete interpretations of recordings. Recognizers could provide an alternative that requires much less expertise and the additional benefits of an approach to scan for species at risk, and archived acoustic data.

CONCLUSIONS

Based on our results, we suggest that recording and automated recognition methods can be useful tools for generating occupancy and detectability information for forest bird communities. Using song recognizers on recorded data from a sufficient number of surveys, we can achieve detection probabilities similar to listening to recordings manually, i.e., manual detection. In addition, song recognizers offer several advantages over manual detection including the ability to process large amounts of recorded data relatively quickly, less stringent requirements for technical expertise to identify songs, and potentially a reduction in variability among observers. However, one significant pitfall is the additional time required to process multiple species. We suggest that the use of song recognizers is sensible when individual or small suites of species are being considered, when larger quantities of recorded data are feasible to collect, and particularly when bird identification expertise is limited. In contrast, where knowledge of full community diversity is a requirement, we suggest that manual interpretation by an expert remains the most efficient and accurate method of assessing avian bioacoustics data acquired by automated recording devices. Next steps should include the manual interpretation of more individual recording to develop an understanding how standard error of occupancy estimates are related to increasing sampling times.

Alternative algorithms have been developed and tested for the automated recognition of bird songs from continuous recordings (Kogan and Margoliash 1998, Acevedo et al. 2009, Katz et al. 2016) and it is likely that there will be continued progress on the development of new and improved algorithms over time, including multispecies approaches (Briggs et al. 2012). However, results from our study suggest that current algorithms are effective for many existing applications when sufficient recording data is assessed, and can, in broad context, enhance our collective ability to achieve both conservation and sustainable forest management goals when used appropriately.

Responses to this article can be read online at:
http://www.ace-eco.org/issues/responses.php/1029

Acknowledgments:

For computations, we used the Colosse supercomputer at Université Laval, managed by Calcul Québec and Compute Canada. Thanks to Andrea Drosdowska, Kevin Good, Kevin Barber, and Derek Chartrand for field support. Tom Swystun and Kerrie Wainio-Keizer provided logistic support on data management and audio processing. Song recognizers were developed using recordings from Borror Laboratory of Bioacoustics (https://blb.osu.edu/), and the Macaulay Library from the Cornell Lab of Ornithology (http://macaulaylibrary.org/).

LITERATURE CITED

Acevedo, M. A., C. J. Corrada-Bravo, H. Corrada-Bravo, L. J. Villanueva-Rivera, and T. M. Aide. 2009. Automated classification of bird and amphibian calls using machine learning: a comparison of methods. *Ecological Informatics* 4:206-214. http://dx.doi.org/10.1016/j.ecoinf.2009.06.005

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. http://dx.doi.org/10.2193/0091-7648(2006)34[211:UADRSA]2.0.CO;2

Alldredge, M. W., K. H. Pollock, T. R. Simons, and S. A. Shriner. 2007. Multiple-species analysis of point count data: a more parsimonious modelling framework. *Journal of Applied Ecology* 44(2):281-290. https://doi.org/10.1111/j.1365-2664.2006.01271.x

Bailey, L. L., J. E. Hines, J. D. Nichols, and D. I. MacKenzie. 2007. Sampling design trade-offs in occupancy studies with imperfect detection: examples and software. *Ecological Applications* 17:281-290. http://dx.doi.org/10.1890/1051-0761(2007)017[0281:SDTIOS]2.0.CO;2

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. http://dx.doi.org/10.1121/1.4707424

Burnham, K. P., and D. R. Anderson. 2002. *Model selection and multimodel inference: a practical information-theoretic approach.* Second edition. Springer-Verlag, New York, New York, USA. http://dx.doi.org/10.1007/b97636

Celis-Murillo, A., J. L. Deppe, and M. F. Allen. 2009. Using soundscape recordings to estimate bird species abundance, richness, and composition. *Journal of Field Ornithology* 80:64-78. http://dx.doi.org/10.1111/j.1557-9263.2009.00206.x

Charif, R. A., D. W. Ponirakis, and T. P. Krein. 2006. *Raven Lite 1.0 user's guide.* Cornell Laboratory of Ornithology, Ithaca, New York, USA.

de Oliveira, A. G., T. M. Ventura, T. D. Ganchev, J. M. de Figueiredo, O. Jahn, M. I. Marques, and K.-L. Schuchmann. 2015. Bird acoustic activity detection based on morphological filtering of the spectrogram. *Applied Acoustics* 98:34-42. http://dx.doi.org/10.1016/j.apacoust.2015.04.014

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. http://dx.doi.org/10.18637/jss.v043.i10

Furnas, B. J. and R. L. Callas. 2015. Using automated recorders and occupancy models to monitor common forest birds across a large geographic region. *Journal of Wildlife Mangement* 79:325-337. http://dx.doi.org/10.1002/jwmg.821

Goyette, J. L., R. W. Howe, A. T. Wolf, and W. D. Robinson. 2011. Detecting tropical nocturnal birds using automated audio recordings. *Journal of Field Ornithology* 82:279-287. http://dx.doi.org/10.1111/j.1557-9263.2011.00331.x

Haselmayer, J., and J. S. Quinn. 2000. A comparison of point counts and sound recording as bird survey methods in Amazonian southeast Peru. *Condor* 102:887-893. http://dx.doi.org/10.1650/0010-5422(2000)102[0887:ACOPCA]2.0.CO;2

Hobson, K. A., R. S. Rempel, H. Greenwood, B. Turnbull, and S. L. Van Wilgenburg. 2002. Acoustic surveys of birds using electronic recordings: new potential from an omnidirectional microphone system. *Wildlife Society Bulletin* 30:709-720.

Holmes, S. B., K. A. McIlwrick, and L. A. Venier. 2014. Using automated sound recording and analysis to detect rare bird species in southwestern Ontario woodlands. *Wildlife Society Bulletin* 38:591-598. http://dx.doi.org/10.1002/wsb.421

Howe, R. W., G. J. Niemi, S. J. Lewis, and D. A. Welsh. 1997. A standard method for monitoring songbird populations in the Great Lakes region. *Passenger Pigeon* 59:183-192.

Hutto, R. L. and R. J. Stutzman. 2009. Human versus autonomous recording units: a comparison of point-count results. *Journal of Field Ornithology* 80:387-398. http://dx.doi.org/10.1111/j.1557-9263.2009.00245.x

Katz, J., S. D. Hafner, and T. Donovan. 2016. Tools for automated acoustic monitoring within the R package monitoR. *Bioacoustics* 25:197-210. http://dx.doi.org/10.1080/09524622.2016.1138415

Klingbeil, B. T., and M. R. Willig. 2015. Bird biodiversity assessments in temperate forest: the value of point count versus acoustic monitoring protocols. *PeerJ* 3:e973. http://dx.doi.org/10.7717/peerj.973

Kogan, J. A., and D. Margoliash. 1998. Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden Markov models: a comparative study. *Journal of the Acoustical Society of America* 103:2185-2196. http://dx.doi.org/10.1121/1.421364

Leach, E. C., C. J. Burwell, L. A. Ashton, D. N. Jones, and R. L. Kitching 2016. Comparison of point counts and automated acoustic monitoring: detecting birds in a rainforest biodiversity survey. *Emu* 116:305-309. http://dx.doi.org/10.1071/MU15097

MacKenzie, D. I., and L. L. Bailey. 2004. Assessing the fit of site-occupancy models. *Journal of Agricultural, Biological, and Environmental Statistics* 9:300-318. http://dx.doi.org/10.1198/108571104x3361

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. http://dx.doi.org/10.1890/0012-9658(2002)083[2248:ESORWD]2.0.CO;2

MacKenzie, D. I., J. D. Nichols, J. A. Royle, K. H. Pollock, L. L. Bailey, and J. E. Hines. 2006. *Occupancy estimation and modeling: inferring patterns and dynamics of species occurrence.* Academic Press, New York, New York, USA.

MacKenzie, D. I., and J. A. Royle. 2005. Designing occupancy studies: general advice and allocating survey effort. *Journal of Applied Ecology* 42:1105-1114. http://dx.doi.org/10.1111/j.1365-2664.2005.01098.x

Mazerolle, M. J. 2006. Improving data analysis in herpetology: using Akaike's Information Criterion (AIC) to assess the strength of biological hypotheses. *Amphibia-Reptilia* 27:169-180. http://dx.doi.org/10.1163/156853806777239922

Mazerolle, M. J. 2016. *AICcmodavg: model selection and multimodel inference based on (Q)AIC(c). R package version 2.0-4.* R Foundation for Statistical Computing, Vienna, Austria. [online] URL: http://CRAN.R-project.org/package=AICcmodavg

Mazerolle, M. J., L. L. Bailey, W. L. Kendall, J. A. Royle, S. J. Converse, and J. D. Nichols. 2007. Making great leaps forward: accounting for detectability in herpetological field studies. *Journal of Herpetology* 41:672-689. http://dx.doi.org/10.1670/07-061.1

McShea, W. J., and J. H. Rappole. 1997. Herbivores and the ecology of forest understory birds. Pages 298-309 *in* W. J. McShea, H. B. Underwood, and J. H. Rappole, editors. *The science of overabundance: deer ecology and population management.* Smithsonian Institution Press, Washington, D.C., USA.

Pitt, D. G., M. Woods, and M. Penner. 2014. A comparison of point clouds derived from stereo imagery and airborne laser scanning for the area-based estimation of forest inventory attributes in boreal Ontario. *Canadian Journal of Remote Sensing* 40:214-232. http://dx.doi.org/10.1080/07038992.2014.958420

Potamitis, I., S. Ntalampiras, O. Jahn, and K. Riede. 2014. Automatic bird sound detection in long real-field recordings: applications and tools. *Applied Acoustics* 80:1-9. http://dx.doi.org/10.1016/j.apacoust.2014.01.001

R Core Team. 2016. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. [online] URL: https://www.R-project.org/

Ralph, C. J., J. R. Sauer, and S. Droege. 1995. *Monitoring bird populations by point counts.* U.S. Forest Service General Technical Report PSW-GTR-149, Pacific Southwest Research Station, Albany, California, USA. http://dx.doi.org/10.2737/psw-gtr-149

Rempel, R. S. 2007. Selecting focal songbird species for biodiversity conservation assessment: response to forest cover

amount and configuration. *Avian Conservation and Ecology* 2 (1):6. http://dx.doi.org/10.5751/ace-00140-020106

Rempel, R. S., K. A. Hobson, G. Holborn, S. L. Van Wilgenburg, and J. Elliott. 2005. Bioacoustic monitoring of forest songbirds: interpreter variability and effects of configuration and digital processing methods in the laboratory. *Journal of Field Ornithology* 76:1-11. http://dx.doi.org/10.1648/0273-8570-76.1.1

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. http://dx.doi.org/10.1642/0004-8038(2002)119[0046:LCTCPA]2.0.CO;2

Royle, J. A., and J. D. Nichols. 2003. Estimating abundance from repeated presence-absence data or point counts. *Ecology* 84 (3):777-790. http://dx.doi.org/10.1890/0012-9658(2003)084[0777:eafrpa]2.0.co;2

Sidie-Slettedahl, A. M., K. C. Jensen, R. R. Johnson, T. W. Arnold, J. E. Austin, and J. D. Stafford. 2015. Evaluation of autonomous recording units for detecting 3 species of secretive marsh birds. *Wildlife Society Bulletin* 39:626-634. http://dx.doi.org/10.1002/wsb.569

Somervuo, P., A. Harma, and S. Fagerlund. 2006. Parametric representations of bird sounds for automatic species recognition. *IEEE Transactions on Audio, Speech and Language Processing* 14:2252-2263. http://dx.doi.org/10.1109/tasl.2006.872624

Stowell, D., and M. D. Plumbley. 2014. Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning. *PeerJ* 2:e488. https://doi.org/10.7717/peerj.488 http://dx.doi.org/10.7717/peerj.488

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. http://dx.doi.org/10.1111/j.1557-9263.2009.00204.x

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. http://dx.doi.org/10.1080/09524622.2011.648753

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

Venier, L. A. and J. L. Pearce. 2004. Birds as indicators of sustainable forest management. *Forestry Chronicle* 80:61-66. http://dx.doi.org/10.5558/tfc80061-1

Venier, L. A. and J. L. Pearce. 2005. Boreal bird community response to jack pine forest succession. *Forest Ecology and Management* 217:19-36. http://dx.doi.org/10.1016/j.foreco.2005.05.058

Venier, L. A., and J. L. Pearce. 2007. Boreal forest landbirds in relation to forest composition, structure, and landscape: implications for forest management. *Canadian Journal of Forest Research* 37:1214-1226 http://dx.doi.org/10.1139/x07-025

Williams, B. K., J. D. Nichols, and M. J. Conroy. 2002. *Analysis and management of animal populations.* Academic Press, New York, New York, USA.

Wimmer, J., M. Towsey, P. Roe, and I. Williamson. 2013. Sampling environmental acoustic recordings to determine bird species richness. *Ecological Applications* 23:1419-1428. http://dx.doi.org/10.1890/12-2088.1

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

Editor-in-Chief: Keith A.Hobson
Subject Editor: Steven L.Van Wilgenburg

Sponsored by the Society of Canadian Ornithologists and Bird Studies Canada
Parrainée par la Société des ornithologistes du Canada et Études d'oiseaux Canada

Appendix 1 Protocols, parameters and cross validation for song recognizers.

Each species recognizer was built separately using a suite of high quality training recordings (clean recordings with little background noise including rain, cars, other birds etc.). Song recognizers were parameterized in a process guided by the Song Scope documentation (Wildlife Acoustics Inc.), but also included a substantial amount of trial and error using both cross-training statistics and results from scanning of test data. Training recordings were downloaded from online digital libraries of audio recordings; Borror Library of Bioacoustics http://blb.osu.edu/database) as the primary source, and xeno-canto (http://www.xeno-canto.org/) as a secondary source) in MP3 format (see Table A1 on source data below). MP3s were converted into .wav files using free online software (Freemake Audio Converter; freemake.com). Target species songs were annotated in the digital file and uploaded into a new recognizer in Song Scope. Songs were inspected in the spectrogram window using log frequency scale with normalized power levels. The resolution of the time scale was altered to optimize the view of individual syllables (uninterrupted segment of song) in the song. Uncharacteristic songs were excluded from the training data.

Parameter values for each song recognizer model are found in Table A2; Parameter Values below. The maximum complexity was set to the default setting of 32 because we found that the results were not sensitive to this parameter. Maximum resolution should also reflect complexity of the song such that higher resolution is needed to model more complex song. Overfitting with a resolution that is too high will result in highly specific recognizers that generate higher levels of false negatives, whereas a resolution that is too low will result in higher levels of false positives. Sampling rate is set to reflect the frequency of the song of the target species. The default is 16000 Hz which is sufficient for most species but too low for the BBWA and GCKI that sing at 8,000 to 10,000 Hz. The sampling rate is split between 2 channels (stereo) so a bird that sings at 10,000Hz requires a sampling rate of at least 20,000 Hz. Larger FFT (fast Fourier transform) sizes will show more frequency resolution at the expense of detail on the time axis. For example, songs with rapid variation in frequency require smaller FFT values to optimize temporal resolution. In practice, we used trial and error to set FFT values to optimize recognizers. We set minimum frequency and frequency range to limit the window of frequencies to those specific to individual species. We set the background filter to 1 second as recommended by the software documentation to reduce background noise. Maximum syllable size, maximum inter syllable gaps and max song length (in milliseconds), were initially estimated from inspection of the spectrogram of the training data and adjusted to optimize cross-training and test scan results. Dynamic range adjusts the sensitivity of the recognizer to the signal to noise ratio. A larger dynamic range value is used when signal to noise ratio was low.

When running test data, each ‘hit’ is assigned a score and quality value which act as sensitivity filters. Score and quality values were examined from the test data results. Normally the lowest score and quality values are associated with false positives so a minimum score and minimum quality criterion is set to filter the results. This has the effect of filtering our many false positives. The minimum score and quality settings used for each recognizer run are found in Table A3:

Score and Quality table below. During the model development phase, the training data are used to generate cross-validation statistics (Table A3). These values are then used to assess the model fit while the parameters are adjusted. When the technician feels that the cross validation cannot be improved, the recognizer is tested against a test recording that has been manually interpreted and that includes multiple species and background noise. A second round of parameter adjustments are then made to maximize the agreement between the recognizer results and the know bird songs.

Table A1.1: Source of training data for song recognizers.

Species	Recordings from Borror	Recordings from xeno-canto	# of individual songs
BBWA	3402, 4641, 6364, 17527	XC133258, XC137479	28
BRCR	14773,84786,100884, 119458, 133327		16
GCKI	17541	XC144683, XC161131, XC168196,XC189416	12
HETH	3671, 3673, 3675, 3691		12
LISP	17551, 18807, 29218, 29378		26
MOWA	13957, 13961, 22525		17
NOWA	10554, 10569, 11189, 12597	XC189085, XC192527, XC195608	14
RCKI	4729, 5937, 10457, 11451, 29059		10
SWTH	10503, 10601, 11458, 14412		9
YBFL		XC110097, XC110098, XC187564, XC189407	33

Table A1.2: Parameter values for each of 10 song recognizers. See text above for descriptions of the parameters.

Species	Max Complex	Max Res	Sample Rate	FFT Size	Min Freq	Freq Range	Filter	Max Syllable (ms)	Max Syllable Gap (ms)	Max Song Length	Dynamic Range
BBWA	32	4	20000	512	152	104	1s	410	154	1306	12
BRCR	32	9	16000	256	53	68	1s	328	216	1720	30
GCKI	32	8	20000	512	164	85	1s	282	333	3866	20
HETH	32	11	16000	256	13	88	1s	400	152	1624	24
NOWA	32	10	16000	512	64	144	1s	144	112	1936	20
LISP	32	7	16000	256	21	128	1s	216	136	1944	16
MOWA	32	6	16000	256	29	70	1s	192	128	1296	30
RCKI	32	4	16000	512	72	70	1s	96	64	1392	16
SWTH	32	11	16000	128	10	40	1s	288	176	1176	25
YBFL	32	6	16000	256	32	59	1s	96	8	200	22
Default	32	6	16000	256	0	128	0	500	500	3000	20

All other settings were left at default

Table A1.3: Performance of song recognizers. Cross training indicates the average and standard deviation of the fit of excluded annotation identifications. Total training indicates the average and standard deviation of the fit of all the training data in the final model which includes the training data.

Recognizer	Minimum Quality	Minimum Score	Cross Training with standard deviation
BBWA	50	70	78.85 +/- 5.77
BRCR	40	60	73.62 +/- 3.05
GCKI	40	60	71.27 +/- 5.48
HETH	50	50	74.42 +/- 6.79
LISP	50	70	77.49 +/- 5.16
MOWA	60	60	75.17 +/- 1.87
NOWA	50	65	72.24 +/- 2.78
RCKI	35	65	83.98 +/- 6.06
SWTH	40	60	72.98 +/- 2.00
YBFL	35	60	87.16 +/- 2.27

Appendix 2. zipped folder with SRC files

Please click here to download file 'Spec_SSR.zip'.

Appendix 3. Model selection results (delta $QAIC_c < 4$) for the bird detection data obtained from four 10-min recordings collected by automated recording units in 2013 and 2014 at 109 stations in Hearst Forest Sustainable Forest Licence, Ontario, Canada.

Species	Model	K	$QAIC_c$	Δ $QAIC_c$	Akaike weight
BBWA (c-hat = 1.61)	ψ(Spruce.cover) p(Year + Visit)	8	145.78	0	0.30
	ψ(Spruce.cover) p(.)	4	147.37	1.59	0.14
	ψ(Spruce.cover) p(Year + Visit + Method)	9	147.73	1.95	0.11
	ψ(Veg.less4m + Veg.great4m) p(Year + Visit)	9	147.83	2.05	0.11
	ψ(Spruce.cover) p(Method)	5	149.19	3.41	0.06
	ψ(Stand.age + Canopy.diversity) p(Year + Visit)	9	149.45	3.67	0.05
BRCR (c-hat = 1.53)	ψ(Stand.age + Canopy.diversity) p(Method)	6	140.90	0	0.17
	ψ(Spruce.cover) p(Method)	5	141.51	0.62	0.13
	ψ(Spruce.cover) p(Method + Veg.less2m)	6	141.92	1.02	0.10
	ψ(Stand.age + Canopy.diversity) p(Method + Veg.less2m)	7	142.04	1.15	0.10
	ψ(Spruce.cover) p(Method + Stand.age)	6	142.43	1.54	0.08
	ψ(Stand.age + Canopy.diversity) p(Method + Stand.age)	7	142.78	1.89	0.07
	ψ(Veg.less4m + Veg.great4m) p(Method)	6	143.57	2.67	0.05
	ψ(Veg.less4m + Veg.great4m) p(Method + Veg.less2m)	7	143.87	2.97	0.04
	ψ(Stand.age + Canopy.diversity) p(.)	5	143.98	3.09	0.04

	ψ(Spruce.cover) p(Method + Veg.less2m + Method:Veg.less2m)	7	144.01	3.11	0.04
	ψ(Stand.age + Canopy.diversity) p(Method + Veg.less2m + Method:Veg.less2m)	8	144.16	3.27	0.03
	ψ(Spruce.cover) p(Method + Stand.age + Method:Stand.age)	7	144.36	3.47	0.03
	ψ(Spruce.cover) p(.)	4	144.65	3.75	0.03
	ψ(Stand.age + Canopy.diversity) p(Method + Stand.age + Method:Stand.age)	8	144.78	3.89	0.02
GCKI (c-hat = 1.17)	ψ(Veg.less4m + Veg.great4m) p(Method + Stand.age)	7	685.70	0	0.76
	ψ(Veg.less4m + Veg.great4m) p(Method + Stand.age + Method:Stand.age)	8	687.97	2.27	0.24
HETH*	ψ(Veg.less4m + Veg.great4m) p(Method + Stand.age)	6	957.65	0	0.55
	ψ(Stand.age + Canopy.diversity) p(Method + Stand.age)	6	959.89	2.25	0.18
	ψ(Veg.less4m + Veg.great4m) p(Method + Stand.age + Method:Stand.age)	7	959.92	2.28	0.18
LISP (c-hat = 1.50)	ψ(Veg.less4m + Veg.great4m) p(Method + Stand.age)	7	161.48	0	0.53
	ψ(Veg.less4m + Veg.great4m) p(Method + Stand.age + Method:Stand.age)	8	162.44	0.96	0.33

MOWA*	ψ(Stand.age + Canopy.diversity) p(Method + Stand.age + Method:Stand.age)	7	237.08	0	0.56
	ψ(Stand.age + Canopy.diversity) p(Method + Stand.age)	6	237.60	0.52	0.43
NOWA (c-hat = 1.34)	ψ(Stand.age + Canopy.diversity) p(Method + Stand.age)	7	189.93	0	0.15
	ψ(Veg.less4m + Veg.great4m) p(Method + Stand.age)	7	189.97	0.05	0.15
	ψ(Spruce.cover) p(Method + Stand.age)	6	190.52	0.60	0.11
	ψ(Stand.age + Canopy.diversity) p(Method + Stand.age + Method:Stand.age)	8	190.96	1.04	0.09
	ψ(Veg.less4m + Veg.great4m) p(Method + Stand.age + Method:Stand.age)	8	191.03	1.10	0.09
	ψ(Spruce.cover) p(Method + Stand.age + Method:Stand.age)	7	191.44	1.51	0.07
	ψ(Stand.age + Canopy.diversity) p(Method)	6	191.61	1.69	0.07
	ψ(Veg.less4m + Veg.great4m) p(Method)	6	191.72	1.80	0.06
	ψ(Spruce.cover) p(Method)	5	191.81	1.88	0.06
	ψ(Veg.less4m + Veg.great4m) p(Method + Veg.less2m)	7	193.65	3.72	0.02
	ψ(Stand.age + Canopy.diversity) p(Method + Veg.less2m)	7	193.89	3.97	0.02
RCKI (c-hat = 1.72)	ψ(Veg.less4m + Veg.great4m) p(Year + Visit + Method)	10	606.53	0	0.63
	ψ(Spruce.cover) p(Year + Visit + Method)	9	608.27	1.73	0.26

SWTH (c-hat = 1.35)	ψ(Veg.less4m + Veg.great4m) p(Year + Visit + Method)	10	575.84	0	0.50
	ψ(Veg.less4m + Veg.great4m) p(Method + Stand.age)	7	578.21	2.38	0.15
	ψ(Veg.less4m + Veg.great4m) p(Method)	6	578.67	2.83	0.12
	ψ(Veg.less4m + Veg.great4m) p(Method + Stand.age + Method:Stand.age)	5	579.70	3.86	0.07
YBFL (c-hat = 2.20)	ψ(Spruce.cover) p(Method + Veg.less2m)	6	444.07	0	0.21
	ψ(Spruce.cover) p(Year + Visit + Method)	9	445.31	1.24	0.12
	ψ(Spruce.cover) p(Method)	5	445.51	1.43	0.10
	ψ(Veg.less4m + Veg.great4m) p(Method + Veg.less2m)	7	446.06	1.98	0.08
	ψ(Spruce.cover) p(Method + Veg.less2m + Method:Veg.less2m)	7	446.27	2.19	0.07
	ψ(Stand.age + Canopy.diversity) p(Method + Veg.less2m)	7	446.31	2.24	0.07
	ψ(Spruce.cover) p(Method + Stand.age)	6	446.53	2.46	0.06
	ψ(Veg.less4m + Veg.great4m) p(Year + Visit + Method)	10	447.33	3.25	0.04
	ψ(Veg.less4m + Veg.great4m) p(Method)	6	447.34	3.26	0.04
	ψ(Stand.age + Canopy.diversity) p(Year + Visit + Method)	10	447.71	3.63	0.03
	ψ(Stand.age + Canopy.diversity) p(Method)	6	447.71	3.64	0.03

*There was no evidence for overdispersion for HETH and MOWA, thus AIC_c was used for model selection of these species.

Appendix 4. Model selection results (delta $QAIC_c < 4$) for the bird detection data obtained from recordings spanning each of the four weeks in 2013 and 2014 collected by automated recording units in 2013 and 2014 at 109 stations in Hearst Forest Sustainable Forest License, Ontario, Canada. Note that MOWA occurred too infrequently for analysis and that models for GCKI, HETH, RCKI, YBFL lacked fit and were excluded from analysis.

Species	Model	K	$QAIC_c$	Δ $QAIC_c$	Akaike weight
BBWA (c-hat = 2.81)	ψ(Veg.less4m + Veg.great4m) p(.)	5	123.43	0	0.30
	ψ(Spruce.cover) p(.)	4	124.59	1.16	0.17
	ψ(Veg.less4m + Veg.great4m) p(Veg.less2m)	6	124.73	1.31	0.16
	ψ(Veg.less4m + Veg.great4m) p(Stand.age)	6	125.67	2.24	0.10
	ψ(Spruce.cover) p(Veg.less2m)	5	126.13	2.71	0.08
	ψ(Stand.age + Canopy.diversity) p(.)	5	126.48	3.05	0.07
	ψ(Spruce.cover) p(Stand.age)	5	126.74	3.32	0.06
BRCR (c-hat = 1.61)	ψ(Veg.less4m + Veg.great4m) p(Stand.age)	6	212.17	0	0.30
	ψ(Veg.less4m + Veg.great4m) p(.)	5	213.24	1.06	0.18
	ψ(Stand.age + Canopy.diversity) p(.)	5	213.49	1.32	0.16
	ψ(Stand.age + Canopy.diversity) p(Stand.age)	6	213.52	1.35	0.15
	ψ(Veg.less4m + Veg.great4m) p(Veg.less2m)	6	214.22	2.05	0.11
	ψ(Stand.age + Canopy.diversity) p(Veg.less2m)	6	215.53	3.36	0.06

LISP (c-hat = 1.31)	ψ(Stand.age + Canopy.diversity) p(Stand.age)	6	94.62	0	0.35
	ψ(Spruce.cover) p(Stand.age)	5	94.64	0.02	0.34
	ψ(Veg.less4m + Veg.great4m) p(Stand.age)	6	95.64	1.02	0.21
MOWA*	ψ(Stand.age + Canopy.diversity) p(Stand.age)	5	69.78	0	0.50
	ψ(Spruce.cover) p(Stand.age)	4	70.52	0.74	0.34
	ψ(Veg.less4m + Veg.great4m) p(Stand.age)	5	72.20	2.43	0.15
NOWA*	ψ(Veg.less4m + Veg.great4m) p(Stand.age)	5	170.08	0	0.31
	ψ(Stand.age + Canopy.diversity) p(Stand.age)	5	170.50	0.41	0.25
	ψ(Spruce.cover) p(Stand.age)	4	171.93	1.85	0.12
	ψ(Stand.age + Canopy.diversity) p(.)	4	173.01	2.93	0.07
SWTH (c-hat = 3.74)	ψ(Veg.less4m + Veg.great4m) p(.)	5	144.97	0	0.31
	ψ(Veg.less4m + Veg.great4m) p(Stand.age)	6	146.49	1.52	0.15
	ψ(Spruce.cover) p(.)	4	146.54	1.57	0.14
	ψ(Veg.less4m + Veg.great4m) p(Veg.less2m)	6	147.11	2.14	0.11
	ψ(Stand.age + Canopy.diversit) p(.)	5	147.74	2.78	0.08
	ψ(Spruce.cover) p(Stand.age)	5	147.84	2.87	0.07
	ψ(Spruce.cover) p(Veg.less2m)	5	148.48	3.51	0.05

*There was no evidence for overdispersion for MOWA and NOWA, thus AIC_c was used for model selection of this species.