

Using paired acoustic sampling to enhance population monitoring of New Zealand's forest birds

Sara P. Bombaci*  and Liba Pejchar 

Colorado State University, 1474 Campus Delivery, Fort Collins, CO, 80523, USA

*Author for correspondence (Email: sara.bombaci@colostate.edu)

Published online: 23 November 2018

Abstract: Large-scale bird monitoring can provide valuable insights about drivers of population change across different spatial and temporal scales. Yet, challenging terrain and survey costs hinder the collection of data needed to estimate absolute abundance or population densities for New Zealand's forest birds. Acoustic sampling is being used more frequently to increase efficiency in avian monitoring and paired sampling facilitates robust density estimation from acoustic data. In paired sampling, point counts are conducted simultaneously by human observers and autonomous recording units (ARUs) to allow estimation of statistical offsets that correct biases in ARU data relative to human observers. These offsets can then be used to calibrate count data collected only by ARUs in a larger sampling scheme. However, the effectiveness of paired sampling has not yet been evaluated in New Zealand. We assessed bias in bird counts from ARUs relative to traditional point counts and evaluated whether paired sampling reduced ARU bias, when present, at 280 count stations in six indigenous forest patches on the North Island from January to April 2017. For 13 forest bird species, we estimated δ , a statistical offset that represents the ratio of the effective detection radius (EDR) of the ARU data to human count data and compared bias in density estimates from ARUs relative to human observers between models with and without δ offsets. We found that δ estimates of EDR ratios were near 1.0 and 95% confidence intervals around δ overlapped 1.0 for nine of 13 species. Furthermore, densities produced by ARU counts were unbiased relative to human point counts for nine of 13 species. When models included δ offsets, ARU density estimate bias was removed for all species. Thus, paired acoustic sampling offers a promising strategy for increasing the efficiency, and spatial and temporal coverage of bird population monitoring across New Zealand.

Keywords: 5-minute bird counts, abundance estimation, acoustic survey bias, ARU, audio monitoring, autonomous recording unit, avian monitoring, bioacoustics, point counts, population density estimation

Introduction

Sustained monitoring is critical for evaluating how animal populations respond to land use change and conservation interventions, especially when monitoring is designed to address specific *a priori* hypotheses (Bart 2005; Nichols & Williams 2006; MacLeod et al. 2012). In New Zealand, management actions frequently focus on reducing the abundance of invasive mammals (e.g. ship rats *Rattus rattus*, Norway rats *R. norvegicus*, mustelids *Mustela* spp., possums *Trichosurus vulpecula*; Parkes & Murphy 2003; Towns et al. 2013), to protect native bird populations, e.g. the Department of Conservation 'Battle for our Birds' campaign (Elliott & Kemp 2016). These actions are meant, in part, to reduce mammal predation on native forest birds and browsing-related impacts to their habitats (Towns et al. 2013).

Robust monitoring is necessary to evaluate whether these and other management actions are meeting conservation objectives (MacLeod et al. 2012). Although a multi-species national monitoring program is underway (see www.doc.govt.nz/our-work/monitoring-and-reporting-system), such

programs require intensive data collection that could benefit from strategies that increase the efficiency or cost-effectiveness of population monitoring. Citizen science programs (e.g. New Zealand eBird, iNaturalist NZ) provide low-cost monitoring data useful for addressing questions about species distributions and occurrence patterns (Scofield et al. 2012), and robust methods have been developed that allow researchers to model changes in species abundance over time from citizen science list data (Szabo et al. 2010). However, these methods do not provide measures of absolute abundance and are considered a complementary method to more intensive monitoring programs (Szabo et al. 2010). Also, these methods do not explicitly account for imperfect detection, and in some cases, indices that do not account for detection bias can generate spurious conclusions that may lead to incorrect management actions being taken (Greene & Pryde 2012; Iknayan et al. 2014). Further, citizen science monitoring programmes often over-sample accessible areas near urban centres and under-sample remote areas with challenging terrain, which limits the value of these data for assessing nation-wide population trends (MacLeod et al. 2012).

Distance sampling (Buckland et al. 2001) and time removal sampling (Farnsworth et al. 2002; Solyomos et al. 2013) methods are frequently used to estimate bird abundance/population density and monitor changes in bird populations, but these methods require observers to collect additional information during bird counts (i.e. distance between the observer and the detected bird or time-of-detection) that require extra training for observers to meet analytical assumptions. Distance sampling analysis also demands numerous observations per species to allow proper estimation of the detection function (Buckland et al. 2001). The need for reliable distance and/or time-of-detection measurements and abundant data can limit the usefulness of these techniques for monitoring bird populations in New Zealand, where the terrain is often steep, thickly vegetated, and difficult to traverse (MacLeod et al. 2012; Allen et al. 2013). Thus, sending trained observers to multiple points distributed across the landscape, or to the same points over repeated visits, to acquire sufficient survey data for distance or removal sampling may be costly or impractical. Furthermore, some bird species that are present, but difficult to detect using standard point counts, e.g. nocturnal residents, human-sensitive species and migrants, may not be recorded by transient observers (Steer 2010; Van Wilgenburg et al. 2017). Alternative population monitoring solutions that address these challenges are needed to facilitate a robust population monitoring program in New Zealand.

A potential tool for increasing the efficiency, and spatial and temporal coverage of bird population monitoring efforts across New Zealand is the use of autonomous recording units, or ARUs. ARUs have been used in New Zealand and elsewhere to supplement data collected by human observers with acoustic data (Haselmayer & Quinn 2000; Hobson et al. 2002; Francis et al. 2009; Steer 2010; Klingbeil & Willig 2015; Van Wilgenburg et al. 2017). ARUs may provide a cost-effective population monitoring solution because they only require humans to visit a location once to deploy the ARU and once to retrieve it (Yip et al. 2017). However, transcribing the audio recordings is time consuming, which may reduce the cost-effectiveness of using ARUs, and these extra costs should be considered. The ARU can be programmed to collect repeated surveys following almost any protocol, and thus can increase the quantity of data collected in an area substantially with relatively little effort (Steer 2010; Van Wilgenburg et al. 2017). Acoustic recording files also provide a permanent record that can be verified by multiple experts, facilitating accurate and reproducible research (Steer 2010). Finally, software is available to aid in bird call identification (e.g. Raven Pro, Bioacoustics Research Program 2014) and advances in machine learning will soon enable accurate automated identification of bird calls from audio recordings (Brandes 2008; Acevedo et al. 2009; Digby et al. 2013; Stowell & Plumbeley 2014).

Most detections of forest birds in point count surveys are auditory (Haselmayer & Quinn 2000; Hutto & Stutzman 2009), so it is reasonable to assume that the data collected using ARU- and human-based surveys would be comparable in these habitats. In fact, point count data collected by ARUs and human observers has frequently produced comparable abundance and diversity estimates (Hobson et al. 2002; Blumstein et al. 2011; Venier et al. 2012; Klingbeil & Willig 2015). However, these two approaches do not always produce similar outcomes (Hutto & Stutzman 2009; Venier et al. 2012), and the efficacy of ARUs has rarely been evaluated for New Zealand birds (MacLeod et al. 2012; but see Steer 2010; Digby et al. 2013). A potential shortcoming of using ARUs is that biases in detection of song

cues between human observers and ARU-based point counts may occur due to differences in the detection radius for each method, which could lead to biased abundance estimates if not corrected for (Van Wilgenburg et al. 2017).

A novel study design and analytical approach was recently developed to allow ARU data to be calibrated with human point count data to estimate bird densities/abundance from both types of surveys, while accounting for imperfect detection and species availability (Van Wilgenburg et al. 2017). By conducting synchronous point count and ARU surveys, researchers can estimate statistical offsets that account for the differences in detection radius between human observers and ARUs when estimating population densities (Solyomos et al. 2013; Van Wilgenburg et al. 2017). These offsets can then be used to correct biases in count data collected by ARUs only, which would allow acoustic recorders to be deployed over large areas or long time periods and calibrated by observers visiting sampling points over a small subset of the survey period to collect paired data. There is also potential for integration of paired ARU sampling with established citizen-science monitoring programs (e.g. eBird <https://ebird.org/newzealand/home> and The Cacophony Project <https://cacophony.org.nz>). Scientists or trained citizen scientists could conduct paired ARU and human observer sampling across different regions that can be used to correct acoustic data from citizen science repositories.

To evaluate the potential for ARUs to be used to improve bird population monitoring in New Zealand, we conducted a field test with the following objectives: (1) to assess whether human point count surveys and ARU-based surveys produced similar density estimates for multiple New Zealand forest bird species, and (2) to test whether the sampling framework of Van Wilgenburg et al. (2017) removes bias in estimated densities between the two approaches, when it occurs. We hypothesised that ARU-based surveys would underestimate forest bird densities relative to human point counts because the detection radius for ARUs is smaller than for human point counts (Van Wilgenburg et al. 2017; Yip et al. 2017). We further hypothesised that the paired acoustic sampling approach (Van Wilgenburg et al. 2017) would remove these biases, when present.

Materials and methods

Study area

Our study was conducted at six sites in the North Island, including two sites in each of Auckland, Waikato and Taranaki Regions (Fig. 1), which ranged in size from approximately 100 ha to 3300 ha, and included three fenced mammal free sanctuaries (Tawharanui Regional Park 36°22'18" S, 174°50'33" E, Maungatautari Ecological Reserve 38°02'58" S, 175°33'36" E, and Rotokare Scenic Reserve 39°27'14" S, 174°24'35" E), and three forest patches with minimal predator control (McElroy Scenic Reserve 36°27'32" S, 174°41'32" E, Te Tapui Scenic Reserve 37°48'38" S, 175°37'23" E, and Tarata Conservation Area 39°10'05" S, 174°21'24" E). We included mammal-free sanctuary sites in our study design to obtain data for rare species, i.e. North Island robin *Petroica longipes*, North Island saddleback *Philesturnus rufusate*, whitehead *Mohoua albigilla*.

Forest cover was dominated by mānuka (*Leptospermum scoparium*) mixed with other indigenous trees in Tawharanui Regional Park and McElroy Scenic Reserve, and tawa



Figure 1. The six study areas (black triangles) where we conducted bird point counts to assess the effectiveness of ARUs relative to human observers for estimating bird densities in New Zealand. From north to south: Tawharanui Regional Park, McElroy Scenic Reserve, Te Tapui Scenic Reserve, Maungatautari Ecological Reserve, Tarata Conservation Area, and Rotokare Scenic Reserve.

(*Beilschmiedia tawa*)–rimu (*Dacrydium cupressinum*) mixed indigenous forest in Maungatautari Ecological Reserve, Te Tapui Scenic Reserve, Tarata Conservation Area, and Rotokare Scenic Reserve. Median annual total rainfall ranged from 1200 mm to 1800 mm, median annual mean temperature ranged from 13°C to 16°C (rainfall and temperature from National Institute of Water and Atmospheric Research 1981–2010 data), and mean site elevation ranged from 44 m to 442 m (see Appendix S1 in Supplementary Material).

Bird surveys

Surveys were conducted from January to April 2017 at 280 point count stations. Points were established at random locations >200 m apart (MacLeod et al. 2012; Allen et al. 2013) along randomly-selected pest monitoring lines (fenced sanctuary sites) or along randomly-placed transects (non-sanctuary sites). We revisited each point 2–11 times over the 4-month

sampling period for a total of 589 surveys. Surveys were conducted between 15 minutes and 5 hours after sunrise by one trained observer (SB). The observer collected distance data following point transect distance sampling protocol (Buckland et al. 2001) and time-of-detection data following time removal sampling protocol (Farnsworth et al. 2002; Sólymos et al. 2013). Specifically, the observer recorded exact distances to detected birds using a rangefinder and recorded the minute interval of initial detection for each individual bird observation over a 5-min observation period.

Before starting each point count, the observer placed an Olympus (Olympus Corporation, Center Valley, PA) DM-620 digital voice recorder with a built-in three microphone system in a quiet location approximately 5 m away from the point location to avoid impacts to recording quality from noises made by observer movement. We chose to use a basic voice recorder to test the application of the paired sampling method under a

cost-effective sampling regime. ARU settings were: MP3 file format at 320 kbps, all three microphones at high sensitivity, low cut filter off, and all other recording options set to defaults. We ran a pilot study in early 2016 to determine optimal ARU settings necessary to retain audio quality while minimising memory requirements. We found that we were able to detect the same individual birds in audio from an MP3 file format at 320 kbps as we could using a higher-quality uncompressed PCM file format (48 kHz, 16 bit), so we chose to use the lower quality setting to reduce file size. The recorder was placed in a small plastic container to protect it from rainfall and moisture. The observer then started the audio recording and walked to the established point to immediately begin the point count survey. After the 5-min bird survey, the observer stopped the audio recordings. We did not conduct surveys during precipitation above a light drizzle or when winds exceeded 20 kph.

Audio transcription

ARU recordings were transcribed after the field season without access to field data by the same observer (SB), as recommended by Van Wilgenburg et al. (2017). We did not transcribe audio files with high sound interference – i.e. when bird calls could not be heard well over background wind or other noise. Excluding noisy audio files reduced our dataset from 589 to 352 paired surveys. We chose to exclude noisy audio files from our analysis because they provided less data per effort, but they can be included if the transcriber incorporates a covariate to account for variation that arises from background noise, e.g. a categorical classification or direct measure of background noise (Van Wilgenburg et al. 2017). Human counts also only occurred during periods of low background noise, e.g. low wind or rain. For each audio detection of individuals, the transcriber would record the species and exact time of detection. The transcriber could replay audio segments to confirm identifications. Times for both human and audio count data were later converted to three time intervals (0–2 m, 2–4 m, and 4–5 m) for the time removal analysis (Farnsworth et al. 2002). We used Raven Software (Raven Lite 2.0.0, Bioacoustics Research Program 2014) to help distinguish cues made by different individuals of the same species. Birds of the same species singing at the same kilohertz (kHz) range and intensity were considered the same individual and thus ‘removed’ from being further detected.

Analysis summary

To evaluate the efficacy of using ARUs for monitoring birds in New Zealand, we followed the framework of Van Wilgenburg et al. (2017) to generate maximum likelihood δ offsets that represent the relationship between ARU and human observer bird counts. We used δ estimates to assess bias in bird counts and density estimates derived from ARU data relative to human point count data, and we validated our findings using repeated random subsampling (Van Wilgenburg et al. 2017). In the section below, we provide a brief overview of the theoretical background of the paired sampling method. In the following section, we describe how we applied the method to assess ARU to human observer bias in New Zealand while noting any modifications we made.

Theoretical background

The paired acoustic sampling approach (Van Wilgenburg et al. 2017) requires observers to conduct simultaneous audio and human point counts and collect time-of-detection and distance data during the human point counts. Thus, paired sampling

is similar to the double observer method in that it utilises a ‘secondary observer’, i.e. an ARU, to estimate differences in detection among observers (Nichols et al. 2000). This method assumes that the population present in a surveyed area for a given species is equal for both the human observer and the audio recorder and that both are exposed to the same acoustic signals. Thus, differences between human and ARU counts should arise chiefly from differences in the area sampled by the two methods (Van Wilgenburg et al. 2017).

Paired acoustic sampling allows biologists to produce population density or abundance estimates that incorporate two components of the detection process: (1) the probability that a bird in the survey area gives a visual or audio cue that is available to be detected (probability of availability, p); and (2) the probability that a bird was detected, given it was available for detection (probability of perceptibility, q) (Allredge et al. 2007; Nichols et al. 2009; Van Wilgenburg et al. 2017). Availability (p) is estimated for both ARUs and the human point count data using removal or time-of-detection methods (Farnsworth et al. 2002; Allredge et al. 2007; Sólymos et al. 2013). Perceptibility (q) is estimated from the human point count data using distance sampling methods (Buckland et al. 2001) and modelled with a conditional maximum likelihood approach that accounts for differences in the area sampled, or effective detection radius (EDR), between human point counts and ARUs (Van Wilgenburg et al. 2017).

See Van Wilgenburg et al. (2017) for the complete theory but briefly, the expected value of a count from human observer data is represented as:

$$E[Y_H] = Npq \quad (1)$$

where Y_H is the count; N is the species’ abundance; p is the probability of species’ availability, given presence, during the cumulative duration of the count; and q is the probability a bird is detected, given availability. This can be rewritten when replacing perceptibility (q) with the area sampled (A_H) to:

$$E[Y_H] = Npq = DA_H p_H \quad (2)$$

where D is the species’ density at the point, and A_H is the area sampled in the human point counts, which can be estimated using distance sampling methods as $\hat{A}_H = \pi \hat{\tau}_H^2$, where τ = effective detection radius. Assuming equal density ($D_H = D_A = D$) and equal availability ($p_H = p_A = p$) among survey methods, and replacing perceptibility (q) with the area sampled (A_H), we can take the mean of the expected ARU to human observer counts to represent the relationship between the areas sampled between the two methods:

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

If we set $\delta = \frac{\hat{\tau}_A}{\hat{\tau}_H}$ then $\hat{\tau}_A = \delta \hat{\tau}_H$ and $\hat{\tau}_A^2 = \delta^2 \hat{\tau}_H^2$, so we could rewrite equation 3 as:

$$\frac{E[Y_A]}{E[Y_H]} = \frac{\delta^2 \hat{\tau}_H^2}{\hat{\tau}_H^2} = \delta^2 \quad (4)$$

which is the squared scaling constant that relates the effective detection radius of the human count data to the unknown effective detection radius of the ARU. The δ^2 estimate can be derived as above, or as a maximum likelihood estimate that accounts for differences in sampling strategies, which is calculated by back-transforming a ‘survey type’ coefficient

($\delta^2 = \exp(\beta)$) in a Poisson or negative binomial regression model (Van Wilgenburg et al. 2017). Here ‘survey type’ is a fixed effect for ARU or human surveys, with human as the reference category, and $\delta^2 = \exp(\beta)$ represents the ratio of bird counts for a species between ARUs and human observers and thus is a measure of bias in τ between the two methods, derived from the generalised linear mixed effects model (GLMM) maximum likelihood coefficients.

The δ^2 estimate is then included as a statistical offset in Poisson or negative binomial GLMMs to model count data from both the human point counts and audio counts, where δ^2 accounts for the differences in the effective detection radius between both survey methods (Van Wilgenburg et al. 2017). The GLMM models also incorporate a random ‘sampling point’ effect to account for multiple visits to sampling points and a statistical offset that accounts for imperfect detection across the different survey types through a correction factor ($C = \pi \hat{\tau}_H^2 \hat{p}_H$), where τ and p are estimated from the human observer data using distance sampling and time removal sampling, respectively (Sólymos et al. 2013; Van Wilgenburg et al. 2017). The δ offset is fit with an indicator function (I_A) in the Poisson or negative binomial GLMM models that takes the value of zero for human observer data and a value of one for ARU data. Thus, the mean count for a point location i with the correction factor C and squared scaling constant δ is written as:

$$\log(\lambda_i) = \log(D_i) + \log(C_i) + I_A \log(\delta^2) \quad (5)$$

Data analysis

We followed the approach described above and in Van Wilgenburg et al. (2017) to conduct our analyses for multiple forest bird species (see Appendix S2 for a list of species and scientific names), excluding those with less than 20 detections and those commonly detected as flyovers (e.g. New Zealand falcon *Falco novaeseelandiae*). We fit count removal models using the ‘detect’ package (Sólymos et al. 2016) in R version 3.4.2 (R Development Core Team 2008) with a ‘survey type’ factor to estimate availability (p) and assess differences in p between ARU and human point count data. We excluded species that did not meet the $p_H = p_A$ assumption (i.e. if the 95% confidence intervals between p_H and p_A did not overlap) from further analyses. We used p estimates from the remaining species’ models to calculate the correction factor, C .

We used program Distance version 7.1 (Thomas et al. 2010) to estimate τ . Unlike Van Wilgenburg et al. (2017), we did not bin exact distance measurements into distance intervals. We fit models with the following key detection functions and series expansions: a half normal function with a hermite expansion, a hazard rate function with a simple polynomial expansion, and a uniform function with a cosine expansion. Distance sampling models also included a survey effort correction to account for repeated visits to plots, where the number of sampling points, K , in the density equation was multiplied by the number of visits to each point (See Section 1.6 in Buckland et al. 2015). Prior to analysis, we viewed detection function histograms and truncated data to visually identified distances when truncation improved model fit (i.e. increased the p -value in Kolmogorov–Smirnov goodness-of-fit tests) over untruncated models. We used Akaike’s Information Criterion with a small sample size correction (AICc) to identify the most parsimonious model (Burnham & Anderson 2002) and used the τ estimate from this model in the correction factor (C) calculation.

Two species (North Island robin and New Zealand fantail *Rhipidura fuliginosa*) frequently approached observers, which

violated the distance sampling assumption of no movement in response to observers. Since all observers were trained to note the initial location of any animals that moved in response to the observer, this issue should have been minimised by our field methods. Yet, because some individuals may have been missed, we also used a grouping analysis method outlined in Buckland et al. (2015) to address this potential issue, where the width of the first distance interval was chosen to encompass the distance over which animals will respond to observers. From field trials, we identified these distances to be 12 m for the New Zealand fantail and 20 m for the North Island robin and we grouped all detections between 0–12 m and 0–20 m for these species, respectively, during the distance sampling model fitting process.

We used repeated random subsampling to validate the models and evaluate bias in density estimates, as in Van Wilgenburg et al. (2017). We randomly selected 70% of the point count locations over 50 iterations and used these subsamples to estimate δ^2 and 95% confidence intervals across the replicates using the GLMM modelling approach described above, where $\delta^2 = \exp(\beta)$. These δ^2 estimates were used in the same iteration with the remaining 30% data subset to fit models for the human observer and ARU data that included a statistical offset with both C and δ^2 incorporated. We also used the 30% validation data subsampled over 50 replicates to calculate δ^2 as an empirical ratio of mean bird count totals from ARU surveys to mean count totals from human surveys. We compared δ estimates and 95% confidence intervals for δ produced from maximum likelihood estimates using the full dataset to those produced with the 70% calibration data and to the empirical estimates produced with the 30% validation data subsample. We also used the 30% subset validation data to evaluate bias in density estimates derived from ARUs relative to human point counts. Bias was calculated as the difference in mean densities predicted by ARU models vs. human observer models ($\hat{D}_A - \hat{D}_H$). We estimated D_H by fitting GLMM models to the human count data for each 30% validation subsample. The model included a random intercept for repeated visits to sampling points and a statistical offset that adjusts for biases in availability and perceptibility, i.e. $\log(\pi \hat{\tau}_H^2 \hat{p})$. We used the same GLMM modelling procedure to estimate D_A , except we fit two models to the ARU data, a model with the same statistical offset used for the human observer data and a model that incorporated δ^2 (estimated from the calibration data) in the statistical offset, i.e. $\log(\delta^2 \pi \hat{\tau}_H^2 \hat{p})$. We calculated bias in ARU density estimates for both offset methods. We also evaluated whether models with a ‘survey type’ effect were supported over null models for all species using AIC model selection to select the most parsimonious models ($\Delta AIC < 2.0$; Burnham & Anderson 2002). We calculated variance explained by the null and ‘survey type’ models using Nakagawa and Schielzeth’s (2013) conditional R² (R²_{GLMM(c)}).

Results

We detected 29 species across all sites (Appendix S2), of which 14 species met our selection criteria including the Australasian magpie *Gymnorhina tibicen*, bellbird *Anthornis melanura*, chaffinch *Fringilla coelebs*, common myna *Acridotheres tristis*, Eurasian blackbird *Turdus merula*, grey warbler *Gerygone igata*, kererū *Hemiphaga novaeseelandiae*, New Zealand fantail, North Island robin, North Island saddleback, silvereye *Zosterops lateralis*, tomtit *Petroica macrocephala*, tūī

Prothemadera novaeseelandiae, and whitehead. All species except the Eurasian blackbird met the assumption of equal availability, (95% confidence intervals overlapped for p_H and p_A estimates), so all but blackbirds were considered in further analyses. However, the tomtit and tūi had confidence intervals that only slightly overlapped for ARU and human observer estimates (Table 1). Removal model estimates of availability between human and ARU count data were highly correlated (Pearson's $r = 0.84$, $p = 0.0003$) across all species.

The effective detection radius, τ , from the human observer data ranged from 16 m to 85 m (Appendix S3), and the scaling constant, δ , ranged from 0.814 to 1.247 (median = 0.954) (Fig. 2, Appendix S4). Calibration δ estimates from the GLMMs fit to 70% of the data over 50 iterations were similar to the maximum likelihood estimates (range = 0.793 to 1.239; median

= 0.943) (Fig. 2, Appendix S4). A δ value of 1.0 indicates that ARU count data and human count data are the same, and values <1.0 indicate that fewer birds were detected during ARUs than during human surveys. Most species' maximum likelihood δ estimates were just below 1.0, indicating that they were detected slightly less during ARU surveys than human point count surveys (Fig. 2, Appendix S4). However, the δ estimates for the bellbird, kererū, and silveryeye ranged from 0.8 to 0.9 (Fig. 2, Appendix S4), which indicates that they were detected 10–20% less during ARU surveys. Tomtit and chaffinch δ estimates were 1.12 and 1.25, respectively (Fig. 2, Appendix S4), which indicates that they were detected 12–25% more during ARU surveys. The confidence intervals from the maximum likelihood δ estimates overlapped 1.0 for

Table 1. Probability of availability estimates from simultaneous point counts conducted by ARUs (p_A) and human observers (p_H) for 13 bird species detected across six native forest sites in 2017 on the North Island of New Zealand. Availability (p) was estimated using time removal methods.

LCL	p_H	UCL	LCL	p_A	UCL
Bellbird	0.861	0.923	0.964	0.873	0.934
Chaffinch	0.422	0.771	0.981	0.667	0.852
Fantail	0.835	0.915	0.966	0.909	0.954
Grey warbler	0.779	0.868	0.933	0.737	0.837
Kererū	0.001	0.147	1.000	0.165	0.506
Magpie	0.819	0.922	0.978	0.747	0.879
Myna	0.607	0.868	0.987	0.868	0.968
North Island robin	0.536	0.788	0.956	0.715	0.887
Saddleback	0.527	0.766	0.940	0.717	0.854
Silveryeye	0.320	0.509	0.732	0.477	0.638
Tomtit	0.663	0.820	0.933	0.903	0.948
Tūi	0.730	0.817	0.891	0.890	0.929
Whitehead	0.610	0.839	0.971	0.360	0.672

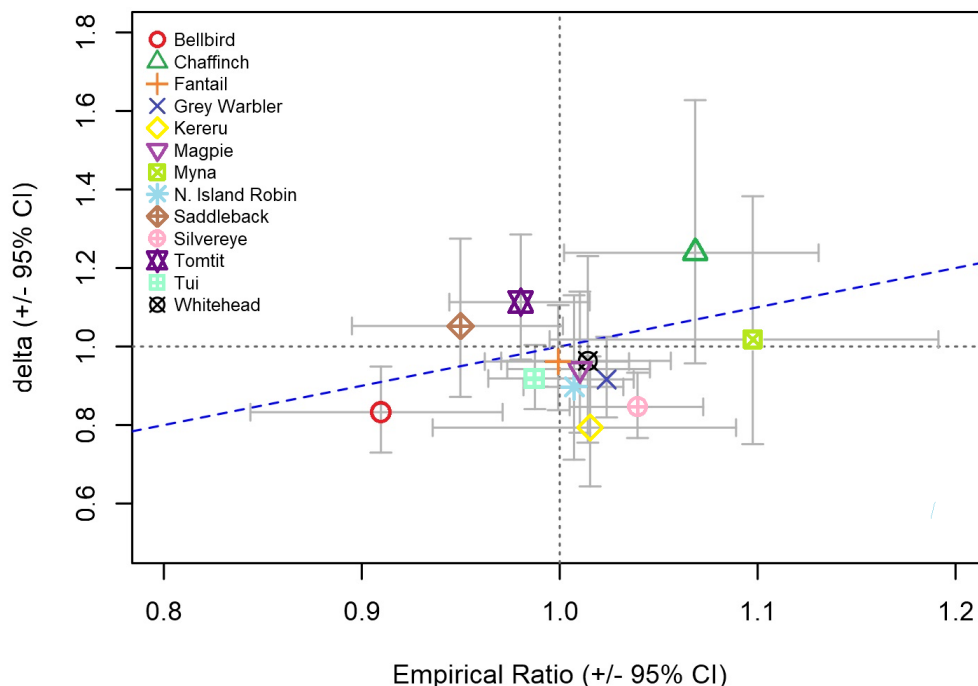


Figure 2. Comparison of maximum likelihood δ estimates (a measure of bias between ARU and human survey bird counts) produced using GLMMs (y-axis) and empirically-estimated δ values (x-axis). GLMM estimates are derived from Poisson model regression coefficients as $\delta = \sqrt{\exp(\beta)}$ from models fit to 70% of the data. Empirical estimates were produced using the withheld 30% validation data by taking the square root of the ratio of mean bird count totals from ARU surveys to mean count totals from human surveys. For both GLMM and empirical approaches, values <1.0 indicate that fewer birds were detected using ARUs than during human surveys.

nine of 13 species (excluding the bellbird, kererū, silvereye, or tūī) (Fig. 2, Appendix S4). The square root of the empirical ratio ranged from 0.910 to 1.10 (median = 1.01; Fig. 2). The confidence intervals around the maximum likelihood δ estimates overlapped with the confidence intervals for the empirical ratio δ estimates (Eqn. 4) for all species except the grey warbler and the silvereye (Appendix S4).

Models that included δ^2 statistical offsets reduced bias in density estimates for 10 of 13 species compared to models that included the offset for availability and perceptibility ($\log(\pi\hat{\tau}_{HP}^2)$) only (Fig. 3). Three species, the kererū, silvereye, and tūī, had negatively biased density estimates (95% confidence intervals around bias estimates did not overlap zero) when δ offsets were not incorporated and one species, the chaffinch, had a positively biased density estimate without δ offsets. Density estimates for all species were unbiased (95% confidence intervals overlapped zero) when δ offsets were incorporated (Fig. 3).

Models with the ‘survey type’ effect were supported over null models for seven of 13 species, but both null models and models with the ‘survey type’ effect explained similar variance (R^2_{GLMM}) and had $\Delta AICc < 2.0$ for all but the bellbird, kererū, silvereye, and tūī (Table 2). The ‘survey type’ model was the only model with $\Delta AICc < 2.0$ for them, suggesting that ARU data may produce biased estimates relative to human observer data for these species.

Discussion

Our findings suggest that ARUs offer a promising tool for increasing the efficiency of forest bird population monitoring in New Zealand. We found that densities produced by ARU counts and human observer counts were generally equivalent for most species, and when present, ARU bias can be corrected for using the paired acoustic sampling method (Van Wilgenburg et al. 2017). We found partial support for our hypothesis that

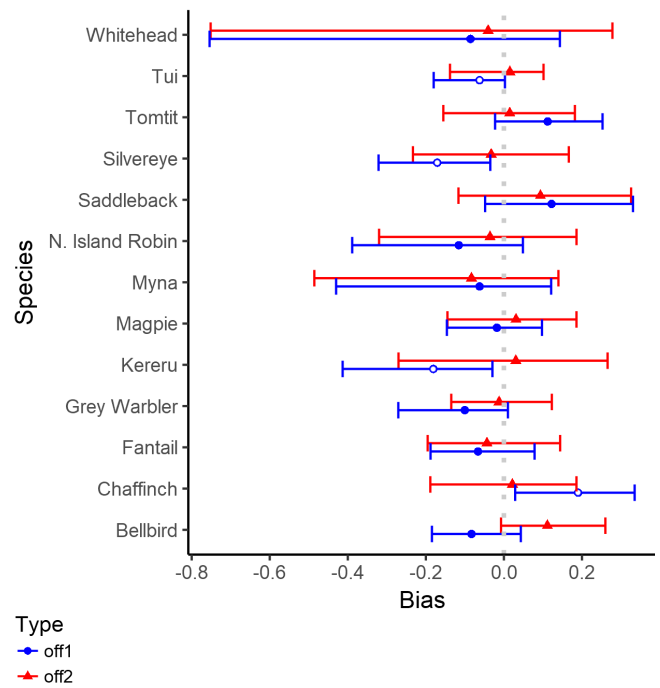


Figure 3. Bias in density estimates (birds ha⁻¹) and 95% confidence intervals from ARUs relative to estimates from human point counts ($\hat{D}_A - \hat{D}_H$) for ARU models with δ (red triangles) and without δ (blue circles) incorporated into statistical offsets that adjust for biases in availability and perceptibility, i.e.

$$\log(\delta\pi\hat{\tau}_{HP}^2) \text{ vs. } \log(\pi\hat{\tau}_{HP}^2)$$

The δ constant corrects for differences in the area sampled between ARUs and human observers and was calculated as $\delta = \sqrt{\exp(\beta)}$ from GLMMs fit to 70% of the data over 50 iterations. Densities were estimated using the withheld 30% validation data. Estimates with 95% confidence intervals that overlapped zero are drawn with closed circles and estimates with confidence intervals that did not overlap zero are drawn with open circles.

Table 2. Model selection results ($\Delta AICc$ values of model rank relative to the model with the lowest AICc value) for Poisson GLMM comparisons between null models (random effect for point transect only) and models with both a point transect random effect and a ‘survey type’ fixed effect to account for differences between bird surveys conducted by ARUs and human observers. Also shown are conditional R^2_{GLMM} values for GLMMs.

Species	Competing models			
	Null	R^2_{GLMM}	Survey type	R^2_{GLMM}
Bellbird	7.575	0.615	0.000	0.618
Chaffinch	1.808	0.470	0.000	0.512
Fantail	0.000	0.104	1.701	0.107
Grey warbler	1.000	0.087	0.000	0.091
Kererū	3.253	0.712	0.000	0.716
Magpie	0.000	0.236	1.656	0.236
Myna	0.000	0.839	1.975	0.842
North Island robin	0.000	0.589	0.532	0.587
Saddleback	0.000	0.429	1.713	0.426
Silvereye	11.579	0.827	0.000	0.850
Tomtit	1.411	0.629	0.000	0.631
Tūī	2.599	0.297	0.000	0.305
Whitehead	0.000	0.832	1.902	0.832

ARU bird counts would be low relative to human observer counts, as the δ estimate (representing the ratio of ARU counts to human observer counts) from both maximum likelihood and empirical approaches was slightly below 1.0 (Fig. 2, Appendix S4), and the bias in density estimates between ARU and human observer counts was negative for most species (Fig. 3). However, 95% confidence intervals around δ overlapped 1.0 for nine of 13 species (Fig. 2, Appendix S4), 95% confidence intervals around density bias estimates overlapped zero for nine of 13 species (Fig. 3), null models were supported by the data ($\Delta\text{AICc} < 2.0$) for nine of 13 species (Table 2), and models with 'survey type' fixed effects explained little extra variation in the data beyond null models for most species (Table 2), suggesting that the bias in ARU bird count data is negligible for most species. This finding that ARUs and human counts produce similar results is consistent with several other studies that found comparable bird abundance, occupancy, or community composition between ARU and human observer counts (Hobson et al. 2002; Celis-Murillo et al. 2009; Blumstein et al. 2011; Digby et al. 2013; Van Wilgenburg et al. 2017).

Although ARU and human observer data were comparable for most species, we found that ARUs may underrepresent densities for the bellbird, kererū, silvereeye, and tūī. Upper confidence limits for δ were < 1.0 and δ values ranged from 0.80 to 0.91 for these species (Fig. 2, Appendix S4). Density estimates were negatively biased for all but the bellbird (Fig. 3), and null models were not supported in AIC model selection comparisons for all four species (Table 2). These lines of evidence suggest that ARU data may underestimate population densities of these species by approximately 10–20% if the bias is not accounted for. Yet, the upper confidence limit for the tūī δ estimate was just below 1.0 (UCL = 0.99), so bias in ARU counts for the tūī may be minimal.

Based on our knowledge of bird behaviour and our experience conducting field surveys, we can speculate why some species were detected less in ARU recordings than in human point counts. First, the bellbird and tūī often produced short single-note calls from distances over 70 m away in our study. These distant calls were easily detected by the observer but were likely missed by the audio recorder, a problem also encountered in automated birdsong detection (Priyadarshani et al. 2018). Second, the kererū was generally detected by the observer upon hearing its strong wingbeats while flying, or by its low frequency (2–3 kHz) coos, but these sounds were easily drowned out by even low background noise during point counts. The observer was able to see birds when flying, which seemed to compensate for missed sound cues. Without the added visual cues, sound cues from kererū may be missed in acoustic recordings in all but those environments with the least ambient noise, resulting in under-estimated densities, if uncorrected. Similarly, Steer (2010) had fewer detections of bird calls < 5 kHz. The silvereeye also produces relatively low-frequency calls and songs (< 10 kHz), but they call or sing with regularity (Steer 2010), which generally made them easy for the observer to detect in our study and in the audio recordings. However, they also frequently occurred in large flocks that were easy to count by visual observation but would be impossible to accurately count in an audio recording. Thus, we likely undercounted silvereeyes during the audio transcription process. Distance sampling can be adapted to accommodate species that frequently occur in clusters by including an estimate of cluster size (Buckland et al. 2001); however, we did not collect data to estimate cluster size and could not apply this technique. We recommend that observers

collect data to estimate cluster size when silvereeye populations are being monitored with ARUs. Based on our findings and field observations, we speculate that ARUs may underestimate bird counts in the following conditions: (1) when species are frequently detected from far distances by single-note calls; (2) when species only produce low-frequency sounds; or (3) when species occur in clusters if information on cluster size is not included during distance sampling analysis.

Two species, the tomtit and chaffinch, had δ estimates above 1.0, suggesting that they were slightly more detectable in ARU recordings than in counts by human observers, but confidence intervals around δ slightly overlapped 1.0. We speculate that these species were more detectable in ARU recordings because transcribers can pause and rewind recordings, which may facilitate detection of species that are easily missed, e.g. those with short duration calls. Although the tomtit and chaffinch both have long duration songs, their calls are relatively short, and some calls may have been missed during human observer counts that were detected in audio recordings. Celis-Murillo et al. (2009) also found higher detection probability for some birds when comparing ARU data to field survey data, which suggests that paired acoustic sampling may improve detection in some cases. Another explanation for the positive bias in ARU recordings may relate to the consistency of bird calls – Steer (2010) indicated that species that call regularly could be overrepresented in audio recordings. The chaffinch and tomtit sometimes call repeatedly, but other species that call regularly (e.g. silvereeyes) were not overrepresented in the audio recordings during our study, so calling regularity may not explain the positive bias for these species. Furthermore, we carefully viewed spectrograms when listening to audio recordings to reduce the chance of double-counting an individual that calls regularly during the recording.

We found that the inclusion of δ statistical offsets reduced bias in density point estimates from ARU recordings relative to human point counts for most species, including the kererū, silvereeye, and tūī, which had δ estimates below 1.0 and negatively biased density estimates when models did not include δ . Importantly, confidence intervals around the density bias estimate overlapped zero when the δ offset was incorporated into GLMMs, which suggests that the paired acoustic sampling approach (Van Wilgenburg et al. 2017) can be used to produce unbiased density estimates relative to human point counts for many New Zealand forest birds. This strategy uses common Poisson or negative binomial GLM or GLMM models with 'survey type' effects to generate offsets that correct for potential biases from ARU-derived data. The ability to correct ARU bias allows researchers to deploy ARUs over a larger sampling scheme while only visiting a subset of sampling points or sampling occasions to collect paired data. Given the relative ease of adding an audio recorder to traditional point counts, and the familiarity of the GLMM modelling procedure, this approach provides a promising opportunity for researchers and scientists to advance population monitoring nation-wide.

We emphasize that ARUs should be considered a supplemental monitoring strategy to field-based data collection. If reliable population monitoring estimates are to be produced, researchers will need to regularly conduct paired counts to test the assumption of equal availability and validate the performance of these models over time and in different contexts. In particular, since we tested this approach in indigenous forest habitat, caution should be used when extrapolating our findings to different habitat types, as the acoustic environment may

change in other habitats and impact the relationship between ARU counts and human observer counts for some species (Van Wilgenburg et al. 2017; Yip et al. 2017). However, we suggest that this method warrants exploration and field testing in other habitat types and encourage inclusion of ‘habitat’ fixed effects into the GLMMs to account for habitat variation if multiple habitats are surveyed for a species. Furthermore, sampling dates and times must be similar between audio recordings and human point counts, i.e. dawn recordings cannot be used to calibrate evening recordings, and recordings made during the breeding season cannot be used to calibrate those made at other times of the year. Since our study was conducted from mid to late summer, our findings may reflect post-breeding season differences between ARUs and human observers and could differ from those produced using data from breeding seasons. However, Van Wilgenburg et al. (2017) found that paired sampling corrected ARU bias in bird surveys conducted during the breeding season in North America, thus it is reasonable to expect that this method could also be useful during breeding season surveys in New Zealand. We encourage the adoption of this method more widely, which would allow generalisable delta offsets to be produced, and conclusions to be made about the suitability of paired sampling across a range of habitats, times, and seasons. Until these generalities can be verified, we discourage researchers from deploying acoustic recorders alone, without also conducting paired human point counts at a subset of points that can be used to estimate ARU bias in their study. Nevertheless, we believe that the ability to collect more data for a smaller field effort justifies further exploration of the usefulness of this method.

It is also important to note that although the data collection component of this method requires minimal investment relative to regular point counts, transcription of audio recordings post-data collection requires an added time investment by skilled observers (Priyadarshani et al. 2018). However, in our study, the cost of labour for transcription was much lower than the cost of sending field observers to collect the same amount of data because the considerable time spent navigating to sampling points was reduced. Furthermore, advancements in machine learning technology may soon make automated detection of bird calls from acoustic datasets a cost-effective and reliable solution (Brandes 2008; Acevedo et al. 2009; Digby et al. 2013; Stowell & Plumbly 2014). However, automated detection does have limitations. These include increased false-positive detections, misclassification issues, difficulties in dealing with noisy recordings or calls from distant birds, and the need for researchers to have knowledge of the complex modelling techniques used in these methods (Priyadarshani et al. 2018). Thus, manual audio transcription may continue to be used by most researchers for some time, and the costs of transcription should be considered.

We tested the paired acoustic sampling method with a relatively low-cost acoustic recording setup. We chose to use a basic voice recorder (Olympus DM-620) without an added microphone as opposed to a professional bioacoustic monitoring setup (e.g. Song Meter SM4, Wildlife Acoustics, Inc., Maynard, MA, USA) because we wanted to test the application of the paired sampling method under a high cost efficiency scenario. The recorders we chose cost approximate \$250 NZD and allow for automated recording at pre-set time intervals. Comparatively, a professional recorder, e.g. SongMeter SM4 (Wildlife Acoustics 2018), costs over \$1000 NZD, and microphone, power, and storage accessories may add several hundred dollars in additional costs. However,

there is a compromise in using our low-cost approach in terms of battery life and storage capacity. Our recorders can store 36GB of data, whereas the SongMeter SM4 can store up to a terabyte of data (Wildlife Acoustics 2018). Our recorders ran for approximately 25 hours of recording time at our settings, but at higher quality settings (i.e. 44.1 kHz 16-bit wav files), the recording time is reduced to 6 hours. The SongMeter SM4 can record up to 400 hours of high-quality stereo recordings before internal batteries need to be replaced (Wildlife Acoustics 2018). Thus, our low-cost setup will require more frequent visitation by field workers to download data and change batteries than with professional setups, and these added costs need to be considered when selecting an appropriate acoustic monitoring scheme. Although biases may exist for different acoustic recording setups (Yip et al. 2017), paired acoustic sampling can remove these biases because it corrects for differences in detection probabilities between a given ARU and human observers (Van Wilgenburg et al. 2017).

Other methods are available to estimate bird densities from acoustic data, e.g. acoustic arrays (Dawson & Efford 2009; Mennill et al. 2012) and acoustic localisation (Collier et al. 2010), or to estimate the bias in density between ARU and human observers with experimental playbacks (Yip et al. 2017). Yet, these methods are logistically more challenging, costlier, and in the case of the playbacks, require extra assumptions to be made about a species’ singing amplitude (Van Wilgenburg et al. 2017). Thus, the paired acoustic sampling approach can provide a cost-effective, relatively easy to implement alternative to other sampling strategies that correct biases in ARU data when estimating bird abundance or population densities. Our study and others have shown that ARUs can produce unbiased abundance and presence-absence data for many bird species in New Zealand and globally (Hobson et al. 2002; Celis-Murillo et al. 2009; Blumstein et al. 2011; Digby et al. 2013), and our study provides additional support that the paired sampling approach can be used to correct for ARU bias when present (Van Wilgenburg et al. 2017). Thus, we recommend increased use of ARUs for bird population monitoring programs in New Zealand and additional testing to refine our estimates and assess our findings in a variety of habitats and contexts.

ARUs offer promising new opportunities to expand research on bird conservation and ecology in New Zealand. Increased efficiency may allow bird surveys to be conducted at greater spatial and temporal scales, which could lead to advances in our understanding of forest bird population responses to invasive mammal eradications, habitat fragmentation or loss, or changes in metapopulation dynamics. Paired acoustic sampling can also be used in a meta-analysis of historic point count data (Van Wilgenburg et al. 2017), which further highlights the potential for this method to be used to expand research opportunities and generate new insights for bird conservation and ecology in New Zealand.

Acknowledgements

This work was supported by funding from a National Science Foundation Graduate Research Fellowship, a Ford Foundation Dissertation Fellowship, a National Geographic Research and Exploration Grant, an Explorer’s Club Exploration Fund Grant, and a Conservation Support Fund Grant from the Riverbanks Zoo. Field research protocol was reviewed and approved by the Colorado State University Institutional Animal

Care and Use Committee (Protocol ID 14-5445A), and site access was granted by the Auckland Council (Application CS66), and the Department of Conservation Research and Collection Authorisation 41860-FLO. We also thank the conservation managers and staff at Tawharanui Regional Park, Maungatautari Ecological Reserve, and Rotokare Scenic Reserve for approving permits to access sanctuary sites and providing valuable information to guide this study. We thank J. Innes and D. Kelly for guidance during project development. We thank the following field assistants, volunteers, and lab assistants for additional support: Breanna Dodge, Dan Hawkins, Tori Flaherty, Brendan Bombaci, Logan Bashford, Brianne Lauro, Cassandra Brown, and Anthony Erwin. Comments from two anonymous reviewers improved an earlier draft of this manuscript.

References

- Acevedo MA, Corrada-Bravo CJ, Corrada-Bravo H, Villanueva-Rivera LJ, Aide TM 2009. Automated classification of bird and amphibian calls using machine learning: a comparison of methods. *Ecological Informatics* 4: 206–214.
- Allredge MW, Pollock KH, Simons TR, Collazo JA, Shriner SA 2007. Time-of-detection method for estimating abundance from point-count surveys. *The Auk* 124: 653.
- Allen RB, Wright EF, MacLeod CJ, Bellingham PJ, Forsyth DM, Mason NWH, Gormley AM, Marburg AE, MacKenzie DI, McKay M 2013. Designing an inventory and monitoring programme for the Department of Conservation's Natural Heritage Management System. Landcare Research Contract Report: LC1730. 250 p.
- Bart J 2005. Monitoring the abundance of bird populations. *The Auk* 122: 15–25.
- Bioacoustics Research Program 2014. Raven Pro: interactive sound analysis software (Version 1.5) [Computer software]. Ithaca, NY: The Cornell Lab of Ornithology. Available from <http://www.birds.cornell.edu/raven>.
- Blumstein DT, Mennill DJ, Clemins P, Girod L, Yao K, Patricelli G, Deppe JL, Krakauer AH, Clark C, Cortopassi KA, Hanser SF, McCowan B, Ali AM, Kirschel ANG 2011. Acoustic monitoring in terrestrial environments using microphone arrays: applications, technological considerations and prospectus. *Journal of Applied Ecology* 48: 758–767.
- Brandes TS 2008. Automated sound recording and analysis techniques for bird surveys and conservation. *Bird Conservation International* 18: S163–S173.
- Buckland ST, Anderson DR, Burnham KP, Laake JL, Borchers DL, Thomas L (Eds.) 2001. Introduction to distance sampling: estimating abundance of biological populations. Oxford, Oxford University Press. 432 p.
- Buckland ST, Rexstad EA, Marques TA, Oedekoven CS 2015. Distance sampling: methods and applications. Springer International Publishing. 277 p.
- Burnham KP, Anderson DR 2002. Model selection and multimodel inference: a practical information-theoretic approach. New York, Springer. 488 p.
- Celis-Murillo A, Deppe JL, Allen MF 2009. Using soundscape recordings to estimate bird species abundance, richness, and composition. *Journal of Field Ornithology* 80: 64–78.
- Collier TC, Kirschel ANG, Taylor CE 2010. Acoustic localization of antbirds in a Mexican rainforest using a wireless sensor network. *The Journal of the Acoustical Society of America* 128: 182–189.
- Dawson DK, Efford MG 2009. Bird population density estimated from acoustic signals: population density from acoustic signals. *Journal of Applied Ecology* 46: 1201–1209.
- Digby A, Towsey M, Bell BD, Teal PD 2013. A practical comparison of manual and autonomous methods for acoustic monitoring. *Methods in Ecology and Evolution* 4: 675–683.
- Elliott G, Kemp J 2016. Large-scale pest control in New Zealand beech forests. *Ecological Management & Restoration* 17: 200–209.
- Farnsworth GL, Pollock KH, Nichols JD, Simons TR, Hines JE, Sauer JR, Brawn J 2002. A removal model for estimating detection probabilities from point-count surveys. *The Auk* 119: 414–425.
- Francis CM, Blancher PJ, Phoenix RD 2009. Bird monitoring programs in Ontario: what have we got and what do we need? *The Forestry Chronicle* 85: 202–217.
- Greene TC, Pryde M 2012. Three population estimation methods compared for a known South Island robin population in Fiordland, New Zealand. *New Zealand Journal of Ecology* 36: 340–352.
- Haselmayer J, Quinn JS 2000. A comparison of point counts and sound recording as bird survey methods in Amazonian southeast Peru. *The Condor* 102: 887–893.
- Hobson KA, Rempel RS, Greenwood H, Turnbull B, Van Wilgenburg SL 2002. Acoustic surveys of birds using electronic recordings: new potential from an omnidirectional microphone system. *Wildlife Society Bulletin* (1973–2006) 30: 709–720.
- Hutto RL, Stutzman RJ 2009. Humans versus autonomous recording units: a comparison of point-count results. *Journal of Field Ornithology* 80: 387–398.
- Iknayan KJ, Tingley MW, Furnas BJ, Beissinger SR 2014. Detecting diversity: emerging methods to estimate species diversity. *Trends in Ecology & Evolution* 29: 97–106.
- Klingbeil BT, Willig MR 2015. Bird biodiversity assessments in temperate forest: the value of point count versus acoustic monitoring protocols. *PeerJ* 3: e973.
- MacLeod CJ, Greene T, MacKenzie D, Allen RB 2012. Monitoring widespread and common bird species on New Zealand's conservation lands: a pilot study. *New Zealand Journal of Ecology* 36: 300–311.
- Mennill DJ, Battiston M, Wilson DR, Foote JR, Doucet SM 2012. Field test of an affordable, portable, wireless microphone array for spatial monitoring of animal ecology and behaviour. *Methods in Ecology and Evolution* 3: 704–712.
- Nakagawa S, Schielzeth H 2013. A general and simple method for obtaining R² from generalized linear mixed-effects models. *Methods in Ecology and Evolution* 4: 133–142.
- Nichols JD, Williams B 2006. Monitoring for conservation. *Trends in Ecology & Evolution* 21: 668–673.
- Nichols JD, Hines JE, Sauer JR, Fallon FW, Fallon JE, Heglund PJ 2000. A double-observer approach for estimating detection probability and abundance from point counts. *The Auk* 117: 393–408.
- Nichols JD, Thomas L, Conn PB 2009. Inferences about landbird abundance from count data: recent advances and future directions. In: Thomson DL, Cooch EG, Conroy MJ eds. *Modeling demographic processes in marked populations*. Environmental and Ecological Statistics, vol 3.

- Boston, MA, Springer US. Pp. 201–235.
- Parkes J, Murphy E 2003. Management of introduced mammals in New Zealand. *New Zealand Journal of Zoology* 30: 335–359.
- Priyadarshani N, Marsland S, Castro I 2018. Automated birdsong recognition in complex acoustic environments: a review. *Journal of Avian Biology* 49: jav-01447.
- R Development Core Team 2008. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. www.R-project.org.
- Scofield RP, Christie D, Sagar PM, Sullivan BL 2012. eBird and avifaunal monitoring by the Ornithological Society of New Zealand. *New Zealand Journal of Ecology* 36: 1–8.
- Sólymos P, Matsuoka SM, Bayne EM, Lele SR, Fontaine P, Cumming SG, Stralberg D, Schmiegelow FKA, Song SJ 2013. Calibrating indices of avian density from non-standardized survey data: making the most of a messy situation. *Methods in Ecology and Evolution* 4: 1047–1058.
- Sólymos P, Moreno M, Lele SR 2016. detect: Analyzing wildlife data with detection error. R package version 0.4-0. The R Project for Statistical Computing, Vienna, Austria. <https://CRAN.R-project.org/package=detect> (accessed 06 January 2018).
- Steer J 2010. Bioacoustic monitoring of New Zealand birds. *Notornis* 57: 75–80.
- Stowell D, Plumbley MD 2014. Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning. *PeerJ* 2: e488.
- Szabo JK, Vesk PA, Baxter PWJ, Possingham HP 2010. Regional avian species declines estimated from volunteer-collected long-term data using list length analysis. *Ecological Applications* 20: 2157–2169.
- Thomas L, Buckland ST, Rexstad EA, Laake JL, Strindberg S, Hedley SL, Bishop JRB, Marques TA, Burnham KP 2010. Distance software: design and analysis of distance sampling surveys for estimating population size. *Journal of Applied Ecology* 47: 5–14.
- Towns DR, West CJ, Broome KG 2013. Purposes, outcomes and challenges of eradicating invasive mammals from New Zealand islands: an historical perspective. *Wildlife Research* 40: 94.
- Van Wilgenburg SL, Sólymos P, Kardynal KJ, Frey MD 2017. Paired sampling standardizes point count data from humans and acoustic recorders. *Avian Conservation and Ecology* 12: 69–83.
- Venier LA, Holmes SB, Holborn GW, McIlwrick KA, Brown G 2012. Evaluation of an automated recording device for monitoring forest birds. *Wildlife Society Bulletin* 36: 30–39.
- Wildlife Acoustics 2018. Song Meter SM4 User Guide. Wildlife Acoustics Inc, Concord, Massachusetts.
- Yip DA, Leston L, Bayne EM, Sólymos P, Grover A 2017. Experimentally derived detection distances from audio recordings and human observers enable integrated analysis of point count data. *Avian Conservation and Ecology* 12: 11.

Editorial board member: Dean Anderson

Received 31 May 2018; accepted 8 October 2018

Supplementary material

Additional supporting information may be found in the supplementary material file for this article:

Appendix S1. Characteristics of the six study areas in which we conducted bird point counts using both human observers and ARUs.

Appendix S2. Species detected across all study sites in 2017, indigenous status, and whether density was estimated or not for the species and the reason.

Appendix S3. Estimates of effective detection radius, τ (Buckland et al. 2001), and 95% confidence intervals for 13 bird species detected across six indigenous forest sites in 2017 on the North Island of New Zealand.

Appendix S4. Estimates of the scaling constant δ (a measure of the ratio of the effective detection radius between bird count data derived from ARU and human survey methods) and 95% confidence intervals produced using (1) the maximum likelihood approach (MLE δ) of Van Wilgenburg et al. (2017), (2) using the MLE approach over 50 repeated subsamples of 70% of the data (calibration δ), and (3) by estimating empirical ratios of mean bird count totals from ARU surveys to mean count totals from human surveys over 50 repeated subsamples of 30% of the data.

The New Zealand Journal of Ecology provides supporting information supplied by the authors where this may assist readers. Such materials are peer-reviewed and copy-edited but any issues relating to this information (other than missing files) should be addressed to the authors.