

# A novel field evaluation of the effectiveness of distance and independent observer sampling to estimate aural avian detection probabilities

Mathew W. Alldredge<sup>1</sup>, Krishna Pacifici<sup>1</sup>, Theodore R. Simons<sup>1</sup> and Kenneth H. Pollock<sup>2\*</sup>

<sup>1</sup>US Geological Survey, NC Cooperative Fish and Wildlife Research Unit, Department of Zoology, and

<sup>2</sup>Zoology, Biomathematics and Statistics, North Carolina State University, Campus Box 7617, Raleigh, NC 27695, USA

## Summary

1. The validation of field sampling techniques is a concern for applied ecologists due to the strong model assumptions implicit in all methods. Computer simulations make replication easy, but they do not give insights into how much bias occurs in real populations. Testing sampling methods on populations of known size can establish directly how well estimators perform, but such populations are very hard to find, and replicate, and they may have unusual attributes.

2. We present a field validation of distance and double-observer methods of estimating detection probabilities on aural avian point counts. Our research is relevant to conservation agencies worldwide who design thousands of avian monitoring programmes based primarily on auditory point counts. The programmes are a critical component in the management of many avian species.

3. Our validation used a simulation system which mimics birds calling in a field environment. The system allowed us to vary singing rate, species, distance, the complexity of points, and other factors.

4. Distance methods performed poorly, primarily due to large localization errors, and estimates did not improve for simplified points.

5. For the double-observer method, two pairs of observers tended to underestimate true population size, while the third pair tended to double-count birds which overestimated the population. Detection probabilities were always higher and population estimates lower when observers subjectively matched birds compared to an objective rule and showed a slight negative bias and good precision. A simplified 45-degree matching rule did not improve the performance of double-observer estimates which had a slight positive bias and much lower precision. Double-observer estimates did improve on the simplified points.

6. *Synthesis and applications.* We encourage ecologists working with sampling methods to develop similar methods of working with simulated populations through use of technology. Our simulated field evaluation has demonstrated the difficulty of accurately estimating population size when limited to aural detections. Problems are related to limitations in the ability of observers to localize sound, estimate distance, and accurately identify birds during a count. Other sources of error identified are the effects of observers, singing rate, singing orientation and background noise.

**Key-words:** avian point counts, aural detections, detection probability, distance sampling, field tests, multiple observers

## Introduction

Point counts are used extensively to monitor spatial and temporal patterns of bird abundance, to assess species–habitat relationships, to evaluate the response of populations to environmental change or management, and to estimate species diversity. The data are easy to collect at large spatial

scales, compared to mark and recapture methods that are frequently too costly. Surveys of breeding songbirds rely heavily on auditory detections, which can make up 70–90% of observations in most environments (Simons *et al.* 2007). Hundreds of thousands of point counts are conducted annually around the world across a spectrum of scales, from short-term site-specific studies to long-term continental scale surveys such as the Breeding Bird Surveys in North America and Great Britain (Ralph, Droege & Sauer 1995; Sauer, Hines

\*Correspondence author. E-mail: pollock@unity.ncsu.edu

& Fallon 2005; BTO 2006). Although these large-scale surveys are best-known, many additional point-count surveys are conducted each year by local, state, federal, and private land-management agencies. Bart (2005) estimated that between 1000 and 2000 independent programmes currently gather long-term data on bird abundance in North America alone. Bird abundance estimates are also a common feature of applied ecological research. A recent survey of 355 research papers in the *Journal of Wildlife Management*, *Ecological Applications*, and *Conservation Biology* revealed that 23% reported on birds, and of those, 36% reported abundance estimates based on count data (Simons *et al.* 2007). However, despite the substantial effort and money expended counting birds for research and monitoring, there is still considerable disagreement over the validity of various survey methods. Most of the disparity concerns the importance of estimating detection probabilities associated with individual counts (Thompson 2002; Rosenstock *et al.* 2002; Williams, Nichols & Conroy 2002; Bart *et al.* 2004).

Detection probability consists of two components (Marsh & Sinclair 1989; Pollock *et al.* 2004). First, a bird has a probability of being available for detection and then, conditional on availability, a probability of detection. For example, in aural surveys a bird has to sing to be available. Further, the bird's conditional probability of detection may decline with distance from the observer. There are a variety of methods that are available to estimate detection probability in avian point counts, with the most important including distance sampling (Buckland *et al.* 2001), use of multiple observers (Nichols *et al.* 2000; Alldredge 2004; Alldredge, Pollock & Simons 2006), time of detection methods (Farnsworth *et al.* 2002; Alldredge 2004; Alldredge *et al.* 2007a), and repeated count methods (Royle & Nichols 2003; Kery, Royle & Schmid 2005). The first two methods assume all animals are available during the count interval, whereas the latter two methods estimate the overall detection probability. For aural point counts, the focus of this study, distance and multiple-observer methods require that individual birds in the area sing at some time during the count.

Validation of field sampling techniques to estimate important population parameters such as those described above is a crucial research concern for many applied ecologists. Possible serious biases in the estimators may result because of failure of model assumptions. Statisticians often resort to computer simulations to study the bias and precision properties of the population parameter estimators under various kinds of model failure (Otis *et al.* 1978; Buckland *et al.* 2001). The simulation approach has the advantage that Monte Carlo replicates can be run. Nevertheless, simulations do not address how much bias occurs in real populations.

One alternative is to sample real populations of birds when population size has been estimated through intensive surveys of essentially closed populations. For example, Fancy (1997) validated point-count-based population size estimates for two intensively studied species of birds in Hawaii. Another is to use real populations of known size and establish directly how well the estimators perform. Edwards & Eberhardt (1967)

reported on estimates of closed population capture–recapture models on a penned population of 135 cottontail rabbits *Sylvilagus floridanus*. Mares, Streilein & Willig (1981) looked at similar estimates for a known population of 82 eastern chipmunks *Tamias striatus*. Novel artificial populations have been sometimes also been used. For example, Carrothers (1973) used a population of 420 taxis in Edinburgh, Scotland, to study the performance of closed population capture–recapture models. Various authors have used artificial populations to study line transect sampling (artificial fish, Witzig 1988; stakes, Laake 1978; bricks, Bergstedt & Anderson 1990).

Many of the studies that used known populations only looked at one population (no replications). Here we have used technology to realistically simulate singing bird populations and communities under field conditions (Simons *et al.* 2007). Our first field experiments using this system focused on factors influencing detection probability (Alldredge, Simons & Pollock 2007b) and measurement error in detection distance (Alldredge, Simons & Pollock 2007c). Alldredge *et al.* (2007d) evaluated the time of detection method, while here we present a field evaluation of the distance sampling and independent double-observer methods for aural point counts.

## Methods

We used a bird song simulation system (Simons *et al.* 2007) to simulate aural point counts in a field setting. This system uses a laptop computer to control remote MP3 players and amplified speakers distributed around a point. The system can simulate conditions on actual point counts while controlling variables of interest.

Thirty-five MP3 players were uniformly distributed with respect to radial area surrounding a single point in a mixed pine–hardwood forest at Howell Woods Environmental Science Center in the Piedmont Region of North Carolina. The forest has a dense understorey that limits visibility to 30 m or less in most directions. All players were set 1 m above ground at radial distances between 5 m and 120 m. Previous experiments at this site demonstrated little effect of speaker height (Alldredge *et al.* 2007c); therefore, we chose to eliminate this variable from our experiments. Speaker height could be important under other forest conditions. Songs for all species were played at a volume of approximately 90 dB at a distance of 1 m. The maximum distance of a speaker from the point was 120 m due to practical constraints and because preliminary studies showed that at least for some species detections were possible at this distance. Each speaker was checked before each experiment to ensure that it was working properly.

A total of sixty 3-min point counts were simulated on 2 days for both a simple and complex experiment. The complex experiment was conducted first and each count had exactly 12 unique birds of up to eight species. There were six species of primary interest (Acadian flycatcher *Empidonax vireescens*; black and white warbler *Mniotilta varia*; black-throated blue warbler *Dendroica caerulescens*; black-throated green warbler *Dendroica virens*; hooded warbler *Wilsonia citrine*; and scarlet tanager *Piranga olivacea*), which were simulated with a population size of 100 each. This meant that for each count, there were either one or two individuals of a target species present. Individuals of an additional two species were added to increase the complexity of the point to the desired 12 birds per point. The 100 birds for each species of interest were played across the

range of distances to approximate a population uniformly distributed with respect to area. The simple experiment was similar to the complex experiment except that there were fewer simulated birds at each point. Simple experiments consisted of exactly eight birds per point and three species of interest (Acadian flycatcher, black-throated blue warbler, and black-throated green warbler). The complex points were similar to what would be expected on real point counts in this area while the simple points were lower in terms of the number of species and total individual birds present.

Previous work (Allredge *et al.* 2007c) demonstrated the importance of singing rate and singing orientation (singing towards or away from the observer) on detection probability; therefore, we incorporated these sources of variability for some of the species of interest. A high singing rate (18 songs per 3-min point count) and a low singing rate (two songs per 3-min point count with a 10-s inter-song interval) were used to simulate singing rate variability for Acadian flycatcher and scarlet tanager. Half of the simulated population was placed in each of the singing rate classes for each species. We simulated variability in singing orientation for black-throated blue warbler and hooded warbler by playing songs out of speakers oriented towards or away from observers. Half of the birds simulated for both of these species were oriented toward observers, while the other half were oriented away from observers. Singing rate (18 songs per 3-min point-count) and orientation (towards the observer) were constant for black and white warbler and black-throated green warbler.

All observers had extensive experience, and most had participated in previous experiments. As the observers had to estimate distance and direction to each bird of each species detected, they participated in 1 day of distance estimation training for the species of interest at the same site used in these experiments. All but one observer had also participated in a previous distance estimation experiment (Allredge *et al.* 2007c) and, therefore, had considerable distance estimation training.

Following the experiments, birds detected during the experiment were scored against the actual bird song played during the experiment; therefore, both estimated distance and true distance were known for each observation. This scoring provided measurements of error rates resulting from misidentification, and double-counting.

A total of 720 individual birds were played on the complex experiment, and 480 individual birds were played on the simple experiment. Thirty points were simulated each day for a total of 60 points in each experiment. Six observers (three pairs) participated on both days of the complex experiment. Four observers (two pairs) participated on both days of the simple experiment.

Distance sampling methodology (Buckland *et al.* 2001) assumes that detection is certain at the point; detection is a decreasing function of distance; there is no measurement error in distances; and there is no movement of birds before detection. We used the program DISTANCE (version 5, Beta 5, <http://www.ruwpa.st-and.ac.uk/distance/>) to analyse our distance data. We analysed the complex and simple experiment separately to estimate density with detection a function of distance from the observer. Each analysis was run with three candidate models for the detection function; half-normal, uniform, or hazard-rate, with a maximum of one adjustment term; simple polynomial or cosine polynomial. A chi-square goodness-of-fit (GOF) test was used to assess model fit, and Akaike's Information Criterion (AIC; Burnham & Anderson 2002) was used to select the most parsimonious model. Data sets were right-truncated by 10% of the maximum distance recorded. A second analysis was run with much greater truncation in an attempt to eliminate problems associated with severe heaping of distance estimates between 60 m and 80 m (S. Buckland and L. Thomas, personal communication). By examining the probability density function, we were able to identify a truncation

distance which excluded distances where heaping was identified. This truncation distance varied but was always between 50 m and 70 m.

While true distance is unknown in actual field studies, we ran additional analyses using the true distance instead of observer-estimated distances. This allowed us to assess the effects of other sources of variability on detection probability when there was no error in distance estimation. These sources of variability included both controlled (orientation of player or singing rate) and uncontrolled sources, such as micro-habitat differences associated with the location of each player.

Independent double-observer approaches require that two observers detect birds simultaneously on the same sample area. Estimation is based on a modified Lincoln–Petersen capture–recapture model based on the birds detected by both observers, detected by observer 1 only, or detected by observer 2 only (Allredge 2004; Allredge *et al.* 2006). This approach assumes that the population is closed and there is no undetected movement out of the area; the counts of the two observers are independent; observers match their detections accurately; and there are equal detection probabilities of all individual birds of each species for each observer. The last assumption can be relaxed if covariates such as distance are used to model variation in the detection probability.

All observers were familiar with the multiple-observer point count method, and understood the importance of accurately matching birds. Observers were paired, and allowed to develop their own subjective matching rules. These rules often applied information about the timing and frequency of songs as well as the perceived location of the song. Observers understood the problems with signal bounce and directionality, and thus, on occasion, observers matched birds that were mapped in very different locations. A second set of detection histories were also developed by using a more conservative 45-degree matching rule based on mapped locations on the data sheets. Under this rule, if both observers mapped two individuals of the same species within 45 degrees of each other, the birds were considered a match.

The independent-observer data were analysed in the program MARK (White & Burnham 1999), using the Huggins Closed Captures model (Huggins 1989, 1991), which allows the incorporation of distance as a covariate. Parameterized models included;  $M_{00}$ , equal detection probability among all observations and observers, and  $M_{obs}$ , equal detection probability among all observations for a given observer, but different detection probabilities among observers. In cases where heterogeneity was incorporated through song orientation or singing rate, the effect was treated as a group effect, similar to sex or age. This was possible because observations were matched to known conditions, which allowed us to assign a group (high or low singing rate, orientation toward or away from observers) to each observation. It is not possible to fit unobservable heterogeneity models (Allredge *et al.* 2006) with only two observers. Observable heterogeneity of detection probabilities due to distance can be modelled. We used estimated distances as model covariates treated as both an additive and multiplicative effect. Akaike's Information Criterion corrected for small sample size ( $AIC_c$ , Burnham & Anderson 2002) was used for model selection. Models with a  $\Delta AIC_c$  (difference in  $AIC_c$  between a model and the most parsimonious model) of  $< 4$  were considered as having some support for a particular data set. We also repeated our analyses using true distance as a model covariate.

## Results

### COUNTING ERRORS

Between 533 and 682 of 720 total birds were counted by individual observers during the complex experiment (Table 1). These

	Complex	Simple
No. of birds simulated	720	480
% birds mapped within true quadrant	75.0 (66.7–83.3)	82.8 (75.1–88.9)
% birds double-counted	7.7 (2.1–11.9)	2.2 (1.2–3.5)
% imagined birds	0.3 (0–1.0)	0.4 (0–0.7)
% birds off 180 degrees	1 (0.5–2.1)	0.2 (0–0.5)
% observers match	70.9 (66.7–83.3)	86.3 (86.2–86.4)
% observers match in same quadrant	65.2 (63.5–68.4)	81.4 (78.0–84.9)

**Table 1.** Summary of information on birds detected plus various identification and matching errors for double-observer and distance sampling point counts under the complex and under the simple experiment. Where appropriate, the mean is given first and then the range is given in parentheses

include double-counts of individual birds and ‘imagined’ birds that did not actually occur on the count. Observers’ ability to accurately map locations of birds was poor. On average, only 75% (range 66.7–83.3%) of birds were mapped within 45 degrees of their true location. Double-counting was substantial (8%, range 2.1–11.9%). Imagined birds (0.3%) and birds that mapped 180 degrees from their true location (1.0%) were less common errors.

Between 419 and 434 of 480 total birds were counted by individual observers during the simple experiment (Table 1). Observers accurately mapped 82.8% (range 75.1–88.9%) of birds to within 45 degrees of their true location. Of the remaining observations 14.4% (range 8.6–20.5%) were not recorded within 45 degrees of their true location, 2.2% (range 1.2–3.5%) of observations were double-counts, 0.4% (range 0.0–0.7%) were imagined birds, and 0.2% (range 0.0–0.5%) were birds mapped 180 degrees from their true location.

#### DISTANCE ANALYSES

Distance analyses were run for six observers on each of the six primary species for the complex experiment, and for four observers and each of the three primary species for the simple experiment, for a total of 36 and 12 data sets, respectively. There were a minimum of 70 detections in each data set. The hazard rate model was selected based on  $AIC_c$  model selection for all data sets of observed distances. The uniform

key function was selected as a competitive model for 15 of the 36 data sets in the complex experiment and five of the 12 data sets of observed distances in the simple experiment. Similarly, the half-normal key function was selected as a competitive model for 12 of the observed distance data sets in the complex experiment and five in the simple experiment. All three key functions resulted in highly competitive models ( $\Delta AIC_c < 2$ ) for all but three of the data sets in the complex experiment and in all but one of the data sets in the simple experiment, when true distances were used. Differences in model selection between observed and true distances were probably due to heaping in observed distances. Observers tended to heap observations at distances between 60 and 80 m. Extreme truncation to distances around 50 to 60 m did not substantially improve density estimates from observed distances. Models based on truncated data tended to favour uniform detection functions, which overestimated the true abundance because the uniform model does not account for any decline in detection probability as distance increases.

The true density of birds was 0.37 birds  $ha^{-1}$  for all primary species in both the simple and complex experiments. Although some individual estimates underestimated true density, estimated densities averaged across species and observers showed a strong positive bias, with the estimated densities ranging from 0.51 (SE = 0.033) birds  $ha^{-1}$  for the simple experiment to 0.85 (SE = 0.054) birds  $ha^{-1}$  for the complex experiment (Table 2). An exception was the hooded warbler in the complex experiment

**Table 2.** Density estimates for the simple and complex experiments based on distance sampling analyses. Averages and ranges for the complex experiment were based on six observers, and those for the simple experiment were based on four observers. True density for all species was 0.37. Estimates were averaged based on the selected model for an individual observer. Estimated distances represent the estimate made by the observer during the point count, and true distances were determined by the actual locations of the players. The species were Acadian flycatcher (ACFL), black and white warbler (BAWW), black-throated blue warbler (BTBW), black-throated green warbler (BTNW), hooded warbler (HOWA), and scarlet tanager (SCTA)

Species	Observed distance				True distance			
	Average	SE	Min	Max	Average	SE	Min	Max
Complex experiment								
ACFL	0.51	0.033	0.23	0.71	0.38	0.023	0.33	0.43
BAWW	0.76	0.060	0.39	1.30	0.39	0.021	0.28	0.46
BTBW	0.62	0.040	0.33	1.11	0.48	0.026	0.35	0.56
BTNW	0.76	0.040	0.36	1.18	0.46	0.024	0.32	0.58
HOWA	0.38	0.020	0.33	0.43	0.37	0.021	0.28	0.41
SCTA	0.52	0.020	0.44	0.62	0.34	0.014	0.30	0.37
Simple experiment								
ACFL	0.63	0.044	0.33	0.88	0.32	0.025	0.23	0.38
BTBW	0.62	0.039	0.48	0.94	0.40	0.027	0.33	0.42
BTNW	0.85	0.054	0.45	1.49	0.35	0.009	0.34	0.37

[average density 0.38 (SE = 0.02) birds ha<sup>-1</sup>]. This large positive bias was caused by substantial heaping of distance estimates between 60 m and 80 m. For true distances between 60 m and 80 m the observed distances were biased high while for true distances beyond 80 m the observed distances were biased low. Observers had a very difficult time estimating distance when the true distance was beyond 60 m even in the simple experiment (see also Alldredge *et al.* 2007c).

Using true distances markedly improved the density estimates for both the simple and complex experiments, with estimates averaged across observers ranging from 0.32 (SE = 0.025) birds ha<sup>-1</sup> to 0.48 (SE = 0.026) birds ha<sup>-1</sup> (Table 2). The overestimates in this ideal case were probably caused by some double-counting and imagined birds, while underestimates could have resulted from some undercounting of birds at intermediate distances because the birds were missed due to masking of songs by other bird songs.

#### TWO INDEPENDENT OBSERVER ANALYSES

The mean number of observations for each observer group in the complex experiment was 696 birds (range 623–786), with one group recording more birds than were actually simulated (720). Observers matched 70.9% (range from 66.7–73.0%) of their observations using their own matching rules and 65.2% (range from 63.5–68.4%) of observations using the 45-degree rule (Table 1). Matching rates were much higher in the simple experiments, averaging 86.3% when observers used their own matching rules and 81.4% using the 45-degree rule.

Model selection using the AIC criteria is summarized in Table 3. The black and white warbler and black-throated green warbler had constant singing rates and all songs were orientated towards observers. Models for these two species incorporated observer, distance, or both observer and distance effects in both the complex and simple experiments. Most data sets for Acadian flycatcher and scarlet tanager, which included singing rate as a source of variability, supported a group effect (*g* represents singing rate) in selected models for both the complex and simple experiments. Observer differences for these two species were minimal in the complex experiment, suggesting that these louder more distinct songs were either easily heard or completely missed due to masking by other songs. The effect of song orientation was less clear for the two species (black-throated blue warbler and hooded warbler) where this factor varied. Singing orientation (the group effect) was important for all hooded warbler data sets, but it was important for only one observer group in the black-throated blue warbler experiment.

Estimated population sizes ranged from 67 birds (SE = 1.66) for the black and white warbler with group A to 143 (SE = 3.85) birds for the black-throated green warbler with group B based on observer matches, compared to the true population size of 100 birds. Observer groups A and C generally underestimated the true population size in the complex experiment, while observer group B, which had more double-counting errors, generally overestimated population size. However, all three observer groups overestimated

**Table 3.** (a) Multiple-observer results from the complex experiment where there were 12 birds calling per point with 60 points. There were three pairs or groups of observers studying six species of interest. The species were Acadian flycatcher (ACFL), black and white warbler (BAWW), black-throated blue warbler (BTBW), black-throated green warbler (BTNW), hooded warbler (HOWA), and scarlet tanager (SCTA). Observer matching rules were either subjective or based on a 45-degree match. Models were chosen based on the AIC<sub>c</sub> procedure and population size estimates with their standard errors are presented. Observed distance is a covariate (*d*), observers are denoted as *obs*, and *g* represents a singing rate group effect, while + indicates that the effects were additive and \* indicates that the effects were interacting

Species	Group	Observer match			45-degree match		
		Model	$\hat{N}$	SE ( $\hat{N}$ )	Model	$\hat{N}$	SE ( $\hat{N}$ )
ACFL	A	M <sub>obs,g,+d</sub>	87	1.79	M <sub>obs,g,+d</sub>	118	14.02
	B	M <sub>d</sub>	102	3.56	M <sub>g,+d</sub>	111	5.88
	C	M <sub>g,+d</sub>	81	8.29	M <sub>g,+d</sub>	93	11.33
BAWW	A	M <sub>obs</sub>	67	1.66	M <sub>obs</sub>	77	0.27
	B	M <sub>obs,d</sub>	90	4.29	M <sub>obs,d</sub>	116	20.45
	C	M <sub>obs,d</sub>	74	4.12	M <sub>obs,d</sub>	92	6.92
BTBW	A	M <sub>obs</sub>	97	2.07	M <sub>obs</sub>	97	2.57
	B	M <sub>obs,g,+d</sub>	142	4.18	M <sub>obs,g,+d</sub>	143	4.54
	C	M <sub>obs,d</sub>	86	1.12	M <sub>obs,d</sub>	86	1.15
BTNW	A	M <sub>d</sub>	122	2.89	M <sub>o</sub>	139	4.74
	B	M <sub>obs,d</sub>	143	3.85	M <sub>obs,d</sub>	157	5.64
	C	M <sub>d</sub>	112	3.2	M <sub>obs,d</sub>	115	3.83
HOWA	A	M <sub>g,+d</sub>	97	2.73	M <sub>g</sub>	110	4.74
	B	M <sub>g,+d</sub>	104	1.65	M <sub>g,+d</sub>	110	2.96
	C	M <sub>g,+d</sub>	88	1.86	M <sub>g,+d</sub>	99	4.11
SCTA	A	M <sub>o</sub>	96	0.50	M <sub>g</sub>	101	1.33
	B	M <sub>g,+d</sub>	104	1.04	M <sub>g,+d</sub>	111	1.87
	C	M <sub>g,+d</sub>	99	5.49	M <sub>g,+d</sub>	111	10.00

**Table 3.** (b) Multiple-observer results from the simple experiment where there were eight birds calling per point with 60 points. There were two pairs or groups of observers studying three species of interest. The species were Acadian flycatcher (ACFL), black-throated blue warbler (BTBW) and black-throated green warbler (BTNW). Observer matching rules were either subjective or based on a 45-degree match. Models were chosen based on the AIC<sub>c</sub> procedure, and population size estimates with their standard errors are presented. Observed distance is a covariate (*d*), observers are denoted as *obs*, and *g* represents a singing rate group effect, while + indicates that the effects were additive and \* indicates that the effects were interacting

Species	Group	Observer match			45-degree match		
		Model	$\hat{N}$	SE ( $\hat{N}$ )	Model	$\hat{N}$	SE ( $\hat{N}$ )
ACFL	AB	M <sub>obs,g,+d</sub>	90	1.42	M <sub>obs,g,+d</sub>	102	4.85
	C	M <sub>obs,g,+d</sub>	81	3.98	M <sub>g,+d</sub>	91	10.23
BTBW	AB	M <sub>obs,g,+d</sub>	89	0.99	M <sub>obs,+d</sub>	97	2.45
	C	M <sub>d</sub>	99	1.18	M <sub>d</sub>	99	1.18
BTNW	AB	M <sub>d</sub>	102	1.34	M <sub>d</sub>	122	4.55
	C	M <sub>obs,d</sub>	103	0.99	M <sub>o</sub>	104	1.08

population size for black-throated green warbler in the complex experiment, due to the higher double-counting errors and lower matching probabilities associated with this species. Population estimates based on observer matches were always

lower than estimates based on the 45-degree matching rule. Observers successfully matched more birds using their own criteria than they did using a rigid 45-degree rule. Higher matching success resulted in higher detection probabilities for observer matched data sets. Overall, 95% CIs included the true population size in only seven of the 24 data sets based on the observer matches. All estimates showed a slight negative bias and good precision. Ninety-five per cent confidence intervals for 13 of the data sets based on the 45-degree matching rule included the true population size, but these estimates had a slight positive bias and much larger standard errors. Unlike the distance estimates, double-observer estimates for the simplified points were far less variable than the complex points and much closer to the true population size of 100.

We repeated our analyses of the multiple observer data using true distance as a covariate. No reduction in bias was achieved so we do not report the results in detail here. As discussed above, the main cause of bias with the multiple observer method appears to be matching errors.

## Discussion

A crucial step in estimation is the evaluation of model assumptions and estimation bias. Often this is not possible because true populations are not known and we are limited to comparing repeated estimates, comparing different methods, or making descriptive assessments of model assumptions. We were able to simulate a field situation for aural point count surveys and make a realistic evaluation of both distance- and multiple-observer point count methods. An advantage is that it is possible to evaluate several variables of interest while controlling other factors that obscure patterns in the data. The most important limitation is the lack of spatial and temporal replication. Although we would expect our results to vary in different habitats, seasons, and environmental conditions, we do not believe these factors would change our fundamental conclusions. These findings are relevant to conservation practitioners involved in literally thousands of point-count-based avian monitoring programmes around the world. Proper interpretation and application of the data collected through these programmes depends on a thorough understanding of the sources of bias and measurement in current sampling methods.

Aural detections often comprise 95% of all bird detections in forested habitats (Simons *et al.* 2007). Our findings indicate that common assumptions about the accuracy of distance estimation, and the ability of observers to accurately map and match observations, are often not met on auditory point counts. Although we think that our results are realistic, we would expect bias and measurement errors on actual point counts to be higher because we simplified a number of factors that influence detection in these experiments. One approach to reducing counting errors might involve reducing the number of species of interest, thus reducing the amount of data that an observer must collect. Increased observer performance on our simplified counts suggests this might occur. However, in field situations, reducing the number of birds singing at a

point is not possible. Instead, observers would have to filter out birds of interest from the remaining 'noise' on a count and this might cause some extra 'imagined' birds and species misidentifications.

The distance method performed poorly in these experiments, producing large positive biases in estimated density. This bias was due in part to double-counting and imagined birds; however, even in cases where these errors did not occur, estimates still showed a large positive bias. Alldredge *et al.* (2007c) demonstrated large errors in distance estimation for aural detections of birds. They demonstrated that observers tended to heap most of their distance estimates beyond 60 m at the same distance. True distances from about 60 m to 80 m were overestimated while distances beyond 80 m were underestimated. This was due to the observer's inability to localize or pinpoint the bird's location from a call cue alone. The large positive bias in density estimates found in these experiments is probably due to similar heaping errors. To examine the importance of accurate distance measurements, we re-estimated density using true distances although obviously these distances would not be known in a real field study. The bias in density estimates, using true distance, decreased greatly but not completely. The remaining bias was probably caused by other individual variations in the detection probabilities, such as singing orientation, singing rate, or other unexplained environmental sources of variation.

Buckland *et al.* (2001) showed many examples of the successful application of distance sampling, including terrestrial and marine mammals in aerial surveys, etc. We believe that distance sampling is likely to work much better for species in habitats like open forest, savanna, and grassland where visual cues are available and localization errors are much smaller.

In our study, the multiple observer method worked better, and was less-biased, than the distance method. The main source of bias for the multiple-observer method was caused by double-counting errors. Double-counting errors and bias were reduced in the simplified experiments. Double-counting will inflate abundance estimates because the counts are inflated and because the detection probabilities are underestimated. Extra capture histories where only one observer supposedly heard the bird (10 or 1) occur as a result of these double-counting errors. Imagined birds have a similar effect on abundance estimates.

A very conservative approach to counting birds can lead to a reduction in counting errors. However, this approach may underestimate the population size, because birds that are hard to detect, perhaps due to low singing rates, will appear less frequently in counts than expected. Such an effect was demonstrated by pair C, which generally had lower estimates than the other pairs because they tended to count birds singing toward them or birds singing at a high rate, but they tended to miss birds with low singing rates or birds oriented away from them.

Matching errors are a critical source of bias in multiple-observer methods (Alldredge *et al.* 2006). In the field, our observers were allowed to match their observations based on any agreed upon criterion. In some cases, observers matched birds because they mapped them in the same location on their

data sheets. In other cases, observers matched birds based on information about the timing and frequency of songs as well as the perceived location of the song. Observers understood the problems with signal bounce and directionality, and thus on occasion, observers matched birds that were mapped in very different locations.

In a statistical sense, matching rules should be stringent and repeatable. For example, observations are matches if they occur within a specified distance of each other. Based on past experience, we realized that observer ability to accurately locate auditory cues was very limited (Allredge *et al.* 2007c), and we thought that a 45-degree matching rule would be reasonable. However, observers were not able to accurately map birds within 45 degrees of their true location, and the 45-degree rule tended to under-match observations, creating too many single observer (10 or 01) detection histories. In the simplified experiment, when observers had more time to accurately map the location of each song, the 45-degree rule performed better.

In general, our simulated field evaluations have demonstrated the difficulty of accurately estimating population size when observations are limited to aural detections of bird songs. Problems are related to limitations in the ability of observers to localize sound, estimate distance (Allredge *et al.* 2007c), and accurately identify birds during a count. Other sources of estimation error identified through this work are the effects of observers, singing rate, and singing orientation (Allredge *et al.* 2007b), and other factors such as background noise (Simons *et al.* 2007).

These findings clearly demonstrate some of the limitations of estimating bird populations when distance sampling and double-observer methods are applied to auditory cues. Nevertheless, we believe approaches that estimate detection probabilities directly will provide better inference than index counts that require even stronger assumptions. We encourage field ecologists to use the best science available while acknowledging the limitations and possible biases of the methods used to make inferences about population size, or trends. Finally, we encourage ecologists to take advantage of technical innovations for testing assumptions behind sampling and field methods, understanding patterns of animal movement (Holland *et al.* 2006) and testing hypotheses about basic biological processes (Halloy *et al.* 2007).

## Acknowledgements

We are very grateful to the volunteers who participated in these field experiments: Jerome Brewster, Adam Efir, Mark Johns, Shiloh Schulte, Clyde Sorenson, and Nathan Tarr. John Wettroth designed our song simulation system. Electrical engineering students at NCSU: John Marsh, Marc Williams, and Michael Foster and Wendy Moore provided valuable technical assistance. Funding for this research was provided by the USGS Status and Trends Program, the US Forest Service, the US National Park Service, and the North Carolina Wildlife Resources Commission.

## References

Allredge, M.W. (2004) *Avian point-count surveys: estimating components of the detection process*. PhD Dissertation, North Carolina State University, Raleigh, NC.

- Allredge, M.W., Pollock, K.H. & Simons, T.R. (2006) Estimating detection probabilities from multiple observer point counts. *Auk*, **123**, 1172–1182.
- Allredge, M.W., Pollock, K.H., Simons, T.R., Collazo, J.A. & Shriner, S.A. (2007a) Time of detection method for estimating abundance from point count surveys. *Auk*, **124**, 653–664.
- Allredge, M.W., Simons, T.R. & Pollock, K.H. (2007b) Factors affecting aural detections of songbirds. *Ecological Applications*, **17**, 948–955.
- Allredge, M.W., Simons, T.R. & Pollock, K.H. (2007c) An experimental evaluation of distance measurement error in avian point count surveys. *Journal of Wildlife Management*, **71**, 2759–2766.
- Allredge, M.W., Simons, T.R., Pollock, K.H. & Pacifici, K. (2007d) A field evaluation of the time-of-detection method to estimate population size and density for aural avian point counts. *Avian Ecology and Conservation*, **2** (2), 13. Available from <http://www.ace-eco.org/vol2/iss2/art13/>.
- Bart, J. (2005) Monitoring the abundance of bird populations. *Auk*, **122**, 15–25.
- Bart, J., Droege, S., Geissler, P., Peterohn, B. & Ralph, C.J. (2004) Density estimation in wildlife surveys. *Wildlife Society Bulletin*, **32**, 1242–1247.
- Bergstedt, R.A. & Anderson, D.R. (1990) Evaluation of line transect sampling based on remotely sensed data from underwater video. *Transactions of the American Fisheries Society*, **119**, 86–91.
- BTO (2006) *The Breeding Bird Survey 2006*. British Trust for Ornithology, Research report no. 471. ISSN 1368–9932. Thetford, UK.
- Buckland, S.T., Anderson, D.R., Burnham, K.P., Laake, J.L., Borchers, D.L. & Thomas, L. (2001) *Introduction to Distance Sampling*. Oxford University Press, New York.
- Burnham, K.P. & Anderson, D.R. (2002) *Model Selection and Multimodel Inference: A Practical Information Theoretic Approach*. Springer-Verlag, New York.
- Carrothers, A.D. (1973) Capture–recapture methods applied to a population with known parameters. *Journal of Animal Ecology*, **42**, 125–146.
- Edwards, W.R. & Eberhardt, L.L. (1967) Estimating cottontail abundance from live trapping data. *Journal of Wildlife Management*, **31**, 87–96.
- Fancy, S.G. (1997) A new approach for analyzing bird densities from variable circular-plot counts. *Pacific Science*, **51**, 107–114.
- Farnsworth, G.L., Pollock, K.H., Nichols, J.D., Simons, T.R., Hines, J.E. & Sauer, J.R. (2002) A removal model for estimating detection probabilities from point-count surveys. *Auk*, **119**, 414–425.
- Halloy, J., Sempo, G., Caprari, G., Rivault, C., Asadpour, M., Tache, F., Said, I., Durier, V., Canonge, S., Ame, J.M., Detrain, C., Corell, N., Martinoli, A., Mondada, F., Siegwart, R. & Deneubourg, J.L. (2007) Social integration of robots into groups of cockroaches to control self-organised choices. *Science*, **318**, 1155–1158.
- Holland, R.A., Thorup, K., Vonhof, M.J., Cochran, W.W. & Wikelski, M. (2006) Bat orientation using Earth's magnetic field. *Nature*, **445**, 702.
- Huggins, R.M. (1989) On the statistical analysis of capture experiments. *Biometrika*, **76**, 133–140.
- Huggins, R.M. (1991) Some practical aspects of a conditional likelihood approach to capture experiments. *Biometrics*, **47**, 725–732.
- Kery, M., Royle, J.A. & Schmid, H. (2005) Modeling avian abundance from replicated counts using binomial mixture models. *Ecological Applications*, **15**, 1450–1461.
- Laake, J.L. (1978) *Line transect estimators robust to animal movement*. MS Thesis, Utah State University, Logan, UT.
- Mares, M.A., Streilein, K.E. & Willig, M.R. (1981) Experimental assessment of several population estimation techniques on an introduced population of eastern chipmunks. *Journal of Mammalogy*, **62**, 315–328.
- Marsh, H. & Sinclair, D.F. (1989) Correcting for visibility bias in strip transect aerial surveys of aquatic fauna. *Journal of Wildlife Management*, **53**, 1017–1024.
- Nichols, J.D., Hines, J.E., Sauer, J.R., Fallon, F.W., Fallon, J.E. & Heglund, P.J. (2000) A double-observer approach for estimating detection probability and abundance from point counts. *Auk*, **117**, 393–408.
- Otis, D.L., Burnham, K.P., White, G.C. & Anderson, D.R. (1978) Statistical inference from capture data on closed populations. *Wildlife Monographs*, **62**, 135.
- Pollock, K.H., Marsh, H., Bailey, L.L., Farnsworth, G.L., Simons, T.R. & Allredge, M.W. (2004) Separating components of detection probability in abundance estimation: an overview with diverse examples. *Sampling Rare and Elusive Species: Concepts, Designs and Techniques for Estimating Population Parameters* (ed. W.L. Thompson), pp. 43–58. Island Press, Washington, D.C.
- Ralph, J.C., Droege, S. & Sauer, J.R. (1995) *Monitoring Bird Populations by Point Counts*. PSW-GTR-149. US Forest Service General Technical Report, Albany, CA.
- Rosenstock, S.S., Anderson, D.R., Giesen, K.M., Leukering, T. & Carter, M.F. (2002) Landbird counting techniques: current practices and an alternative. *Auk*, **119**, 46–53.
- Royle, J.A. & Nichols, J.D. (2003) Estimating abundance from repeated presence-absence data or point counts. *Ecology*, **84**, 777–790.

- Sauer, J.R., Hines, J.E. & Fallon, J. (2005) *The North American Breeding Bird Survey*. Results and analysis 1966–2004, version 2005.2. US Geological Survey Patuxent Wildlife Research Center, Laurel, MD. Available at [www.mbrpwrc.usgs.gov/bbs/bbs.html](http://www.mbrpwrc.usgs.gov/bbs/bbs.html).
- Simons, T.R., Allredge, M.W., Pollock, K.H. & Wettröth, J.M. (2007) Experimental analysis of the auditory detection process on avian point counts. *Auk*, **124**, 986–999.
- Thompson, W.L. (2002) Towards reliable bird surveys: accounting for individuals present but not detected. *Auk*, **119**, 18–25.
- White, G.C. & Burnham, K.P. (1999) Program MARK: survival estimation from populations of marked animals. *Bird Study*, **46**, S120–S139.
- Williams, B.K., Nichols, J.D. & Conroy, M.J. (2002) *Analysis and Management of Animal Populations*. Academic Press, San Diego, CA.
- Witzig, J.F. (1988) *The visual assessment of reef fish communities*. Ph.D thesis, North Carolina State University, Raleigh, NC.

Received 4 May 2007; accepted 22 May 2008

Handling Editor: Chris Elphick