special issue:
Advances in tools for bird population monitoring in New Zealand

A comparison of different approaches to monitoring bird density on New Zealand sheep and beef farms

Florian Weller

Centre for the Study of Agriculture, Food and Environment, University of Otago, PO Box 56, Dunedin 9054, New Zealand
(Email: florian.g.gweller@gmail.com)

Published on-line: 30 July 2012

Abstract: When designing large-scale bird monitoring schemes, financial constraints often require researchers to make trade-offs in the spatial resolution and precision of the density estimates by varying the number of sites monitored and the intensity of sampling effort per site. Here, I compare density estimates of four common farmland bird species (skylark, blackbird, song thrush, Australasian magpie) on South Island sheep & beef farms collected using a large-scale monitoring scheme with equivalent estimates from a more intensive survey, and discuss possible sources of imprecision and bias in each. Density estimates from the intensive survey were generally lower and more precise than the monitoring-scheme ones, and the surveys were susceptible to different types of bias. These effects were linked to combined differences in modelling methods and sampling effort distribution, and to observer-related issues. In long-term designs, estimates from data pooled over several annual surveys are likely to become accurate quickly, but an increase in monitoring effort per site may be required to increase the precision of individual-survey estimates. Species that pose challenges to observers may be hard to estimate accurately with the monitoring-scheme design explored here, and the use of pilot surveys is recommended. Raw count data from the same species were tested for their usefulness in the creation of reliable relative population indices. A sufficiently constant ratio of plot-count indices to absolute density estimates was found in skylarks and thrushes, while inclusion of a correction parameter to account for woody vegetation effects on detectability was necessary in blackbirds, and magpie estimates proved unreliable. Similar analyses are recommended for all monitored species when trend estimation using relative indices is intended.

Keywords: bias; distance sampling; population index; precision; survey design

Introduction

Designing a bird monitoring survey is an exercise in transforming available resources into the most appropriate data. The overall design of any large survey programme will be dictated by the primary aim of the project. This might be broadly categorised as interest in population distribution, indices of population change, or evaluating the effects of management actions, in ascending order of resolution and cost of result (Johnson 1999; Svensson 2000). Generating information on long-term population trends, achievable using index counts, has frequently been the foremost aim of large monitoring schemes (MacLeod et al. 2012a). In contrast, the answering of specific ecological questions or tracking of specific conservation and management actions is often relegated to more specialised projects that can expend the effort to produce estimates of absolute abundance (Johnson 1999). It is, however, often entirely feasible to structure a large longitudinal study in such a way that it yields accurate estimates of absolute abundance while still covering large areas (e.g. Newson et al. 2005, 2008).

Ideally, the results of a survey should be both precise (i.e. repeated measurements, taken under the same conditions, should show the same results to a high degree) and unbiased (i.e. no systematic errors should be present in the estimates compared with the true value). In practice, this will generally be hard to realise. Financial and practical constraints frequently make it necessary to trade off either or both of these qualities against other factors. Depending on the type of survey, this might concern the number of temporal replicates and the coverage of different times of the year, the spatial scope and spatial resolution of the survey (trade-offs in the application of available effort to smaller or larger areas or number of areas), and constraints on field methods, e.g. limitations on observer continuity, training, or equipment. Survey design to control for bias and imprecision thus has to be shaped by priorities. Lack of bias (i.e. accuracy) is important when interest lies in closely approximating a 'true' value and absolute population sizes are required, and for comparisons across survey types; fundamentally, the presence of different biases in different results makes them incompatible, and the only solution is recognition and removal. Precision is most important where

This special issue reviews the advances in tools for bird population monitoring in New Zealand. This issue is available at www.newzealandecology.org.nz/nzje/.

repeat measurements are concerned, and confidence in relationships within a series is needed. This is primarily the case in trend estimation.

The expected precision and susceptibility to bias of a survey may be difficult to assess before implementation, although approaches such as statistical power analyses can be of great benefit. Often, useful information may only be gained after some results are available for investigation. A comparison with estimates derived from an alternative source can be particularly useful at this stage. I therefore carried out a comparison of two bird surveys, undertaken on the same properties and with an overlapping time frame, to investigate likely sources of imprecision and bias in each.

The two survey designs discussed are the ongoing ARGOS farmland bird monitoring scheme and a short-term intensive bird survey carried out on a subset of the same area. These surveys represent two different design approaches. The ARGOS scheme is a longitudinal study covering many sites in an ongoing programme, and trades this off against limited data from each site visit. The intensive survey had a more limited temporal and spatial scope with monitoring effort applied to deliver higher replication at fewer farms. Distance sampling methods (Buckland et al. 2001, 2004) were used in both designs as a way to transform raw counts into absolute density estimates. This removes imprecision and bias generated by changes in detection probability, which, for example, may depend on species behaviour, habitat type, and observer identity (e.g. Newson et al. 2005, 2008; Alldredge et al. 2007; MacLeod et al. 2012b), with the aim of producing abundance estimates that are as unconditional as possible. Failure to account for variation in detectability will tend to underestimate true densities (Norvell et al. 2003; Newson et al. 2005, 2008; White 2005; Buckland 2006). However, if absolute estimates are not a high priority, simpler (uncorrected) counts might serve instead. In many situations uncorrected estimates are of use as relative indices, e.g. to track population trends on the same sites across time. As this is one of the objectives of the ARGOS survey, the use of uncorrected estimates produced from raw detection data to supplement the results is an interesting possibility. As a secondary focus, I therefore tested the reliability of raw count estimates derived from the intensive survey.

Methods

Survey design

This study assessed the effect of varying the intensity of sampling effort on individual farms on the precision and accuracy of density estimates for four focal species (skylark *Alauda arvensis*, blackbird *Turdus merula*, song thrush *T. philomelos*, Australasian magpie *Gymnorhina tibicen*), by comparing information collected by a long-term monitoring scheme with that collected using an intensive survey effort. The study focuses on 12 sheep & beef farms on the South Island of New Zealand, which were a subset of the 36 sites monitored by the ARGOS monitoring scheme (MacLeod et al. 2012b). The 12 farms were grouped into four clusters (of three farms each) located on Banks Peninsula (Canterbury) and near Oamaru (Otago), Outram (Otago) and Owaka (Southland).

The ARGOS monitoring scheme was set up in 2004, with three surveys carried out over a 6-year period (summers of 2004/05, 2007/08 and 2009/10). All surveys were undertaken in December–January, the core breeding season for most of the monitored species; each farm was surveyed once during each of the three rounds, and all bird sightings were recorded. Results from these three surveys were analysed separately at survey-level ('ARGOS 1–3') and pooled ('ARGOS Global') (MacLeod et al. 2012b), based on models fitted to the complete 36 farm dataset and evaluated for the 12 focus farms (which were a subset of those included in the ARGOS monitoring scheme).

The 'intensive' survey was limited to the 12 focus sheep & beef farms, with only the four focal species monitored. Surveys took place between November 2005 and August 2007, thus overlapping the initial two monitoring scheme surveys. Each farm was visited 9–10 times (circuits): five visits occurred during the breeding season of the focal species (September–January), two in winter (June–August), and three in between (February–May). Sightings for each farm were pooled over all visits for analysis (Weller 2009; Weller et al. 2012).

Line-transect distance sampling was used as a monitoring method (Buckland et al. 2001; Spurr et al. 2012), and a standard distance sampling protocol was followed (MacLeod et al. 2012b). In both surveys, 8–12 line transects of target length 500 m (randomly placed each time) were walked on each farm visit and sighting distances and angles recorded using a laser rangefinder and compass.

Estimating density

Detections were modelled in Distance 6.0 and 6.0 Beta 1 (Thomas et al. 2010) using multiple covariate modelling (Marques & Buckland 2003). However, the specific structure and treatment of the collected data differed by design between the two surveys. For the ARGOS datasets, where effort was split between many farms, there were generally too few detections of a given species on a given farm to fit reliable farm-level distance models (Table 1). Therefore, detections for each species were pooled across all 36 farms for each survey (survey-level) and across all surveys (global-level), with detection functions fitted independently to the survey-level and global-level pooled datasets (MacLeod et al. 2012b). Individual estimates for the 12 focus farms were then generated by post-stratifying the dataset by farm, i.e. the common detection function was applied to each farm-level dataset. This allowed the generation of estimates, using well-formed detection functions, even for farms with very few detections (Macleod et al. 2012b). There were sufficient data from the intensive survey, when detections were pooled across all visits, to fit farm-level detection functions (Table 1), although sample sizes were sometimes close to the lower recommended limit (Buckland et al. 2001; Weller 2009; Weller et al. 2012).

Estimate analysis

Differences in farm-level density estimates between survey types were explored using a mixed-effect general linear model (function *lme* in R 2.12.1; R Development Core Team 2010), with farm identity as random effect to account for the repeated samples taken from farms, and using pairwise post hoc comparisons (Tukey's HSD) to identify instances of significant difference. To control for environmental and observer differences in detectability between survey years that were not modelled in ARGOS Global, two averages of the survey-level ARGOS estimates were included in this analysis: an average of all three ARGOS survey years, and a 2-year average of ARGOS 1 and 2 only, to account for ARGOS 3 being performed later than the time frame covered by the intensive survey. A data series of estimates derived from the breeding season of the intensive survey only was also included to test for the influence of different sampling seasons.

Table 1. Average sample size (number of detections) for the focal species, both available for modelling and attributable to each individual farm (12 in the intensive survey, 36 in the ARGOS surveys), in each survey.

Species	Intensive		ARGOS Global		ARGOS 1		ARGOS 2		ARGOS 3	
	For modelling	Per farm	For modelling	Per farm	For modelling	Per farm	For modelling	Per farm	For modelling	Per farm
Skylark	151	151	3774	117.9	1352	36.5	1878	55.2	559	21.5
Blackbird	95	95	1087	34.0	325	8.8	405	11.9	296	11.4
Song thrush	46	46	715	22.3	225	6.1	274	8.1	230	8.8
Magpie	71	71	1061	33.2	460	12.4	452	13.3	189	7.3

Raw count analysis

For raw count data to be regarded as reliable indices of population size, the raw count estimate (i.e. the density estimate derived by dividing the unmodified number of detections by total surveyed area) should vary in proportion to the more accurate detectability-corrected estimate. To test this, I explored the relationship between raw count and distance-sampling estimates, using raw count data for each farm collected by the intensive survey. Ten artificial count sets per species were created by selecting all detections within an increasing maximum (cut-off) distance (25–250 m in 25 m increments), and density estimates computed by dividing this number by the area of inference (total transect lengths on farm × 2 × cut-off distance). A general linear model was then fitted at each cut-off distance with the farm-level distance sampling density estimate as a dependent variable, and the raw count estimate as the predictor. A second model was fitted with 'vegetation cover' (average farm-level woody vegetation cover percentage in the raw count area), derived from GIS maps prepared by ARGOS, as an additional predictor.

Results

Model parameters

Figure 1 shows a comparison between the surveys of the averages (across all farms) of several parameters related to distance model fitting. Detection probability (p), describing the average (av.) probability of detecting any given bird on a transect, was broadly similar across surveys and species, tending towards values around 0.5 (Fig. 1a). Variance of

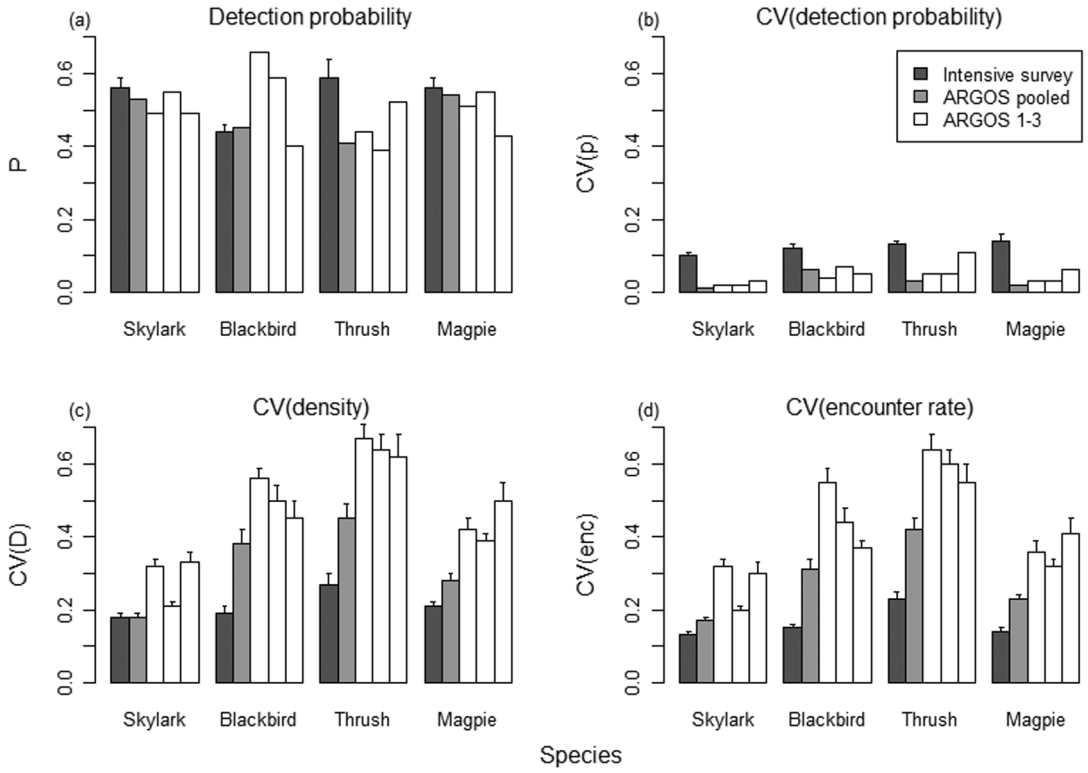

Figure 1. Comparisons between surveys (intensive survey, ARGOS Global (pooled), and ARGOS 1–3) of averages of (a) farm-level detection probability and (b) farm-level coefficients of variation in detection probability, (c) density, and (d) encounter rate, for the four focal species. Bars show averages of farm-level estimates + 1 SE.

detection probability, in the form of the coefficient of variation CV(*p*), was always largest in the intensive survey (av. 0.12 across species); variance tended to be lowest in ARGOS Global (av. 0.03; Fig. 1b). The variance of the density estimate CV(*D*) and the encounter rate estimate CV(*enc*) (the number of detections over total transect length) showed similar characteristics. Both measures were always lowest in the intensive survey (av. CV(*D*) 0.18 & av. CV(*enc*) 0.14), higher in ARGOS Global (0.32 & 0.28), and higher again in ARGOS 1–3 (0.41 & 0.42) (Fig. 1c,d). In contrast to CV(*D*) estimated at farm-level, this variance measure (unlike CV(*enc*)) was much lower in the ARGOS surveys (av. 0.06) when estimated across all farms, i.e. for globally pooled bird populations.

There were large differences between the surveys in both number of detections that the fitted models were based on (on average 17× and 7× as many for ARGOS Global and ARGOS 1–3 vs intensive survey, respectively) and number of detections that these models were applied to on a farm level (on average 1/2× and 1/7× as many for ARGOS Global and ARGOS 1–3 vs intensive survey, respectively; Table 1). Overall, there was a clear inverse relationship between farm-level sample size and magnitude of estimate variances.

Distance-sampling density estimates

To assess the similarity of farm-level density estimates between surveys, values for each species were plotted against the same axis, with farm ordered by increasing density in the intensive survey (Fig. 2). Magpie densities were very similar across all series, with only a few outliers. Skylark estimates also agreed relatively closely, but ARGOS 2 values tended to be larger than those of the other series. There was greater discrepancy in the blackbird and thrush datasets. ARGOS 3 showed generally elevated values in thrushes, while all three annual ARGOS series were outliers for blackbirds. In both species, intensive-survey estimates tended to be among the lowest or the lowest of the series, while the majority of ARGOS values were greater (sometimes by several hundred percent). Here, as in all species, ARGOS-Global estimates were closer to intensive-survey estimates than those for ARGOS 1–3 (but were still almost always higher), and tended to occupy a middle range between extremes in the individual years (as would be expected from a pooled dataset). There was no obvious dependence of the degree of discrepancy on position in the magnitude order of estimates (Fig. 2).

When ARGOS 1–3 were averaged across surveys for modelling in a GLM, no significant differences were found between this 3-year average and the global estimate in any species, and none were present between the intensive-survey and the breeding-season estimates. Intensive-survey estimates were not directly compared with ARGOS Global or the 3-year average, as these included the 2009/10 sampling season not covered by the former. However, a 2-year average of ARGOS 1 and 2, excluding the third ARGOS survey year, showed no significant differences to any of the other sets.

Raw count estimates

Raw count density estimates for the four farm clusters in the intensive survey, and for all clusters together, were constructed from an increasing series of cut-off distances and the sum of all detections out to that distance (Fig. 3). For all species, raw count density estimates were consistently lower (by 62% on average) than distance-sampling density estimates based on the same data. For skylarks, blackbirds and thrushes, graphs of raw count estimates versus cut-off distance showed steady drop-offs with increasing distance, with variations depending on the individual cluster populations. Cluster 2 series for these three species, and estimates for skylark in general, displayed a bell shape indicative of avoidance behaviour in the surveyed

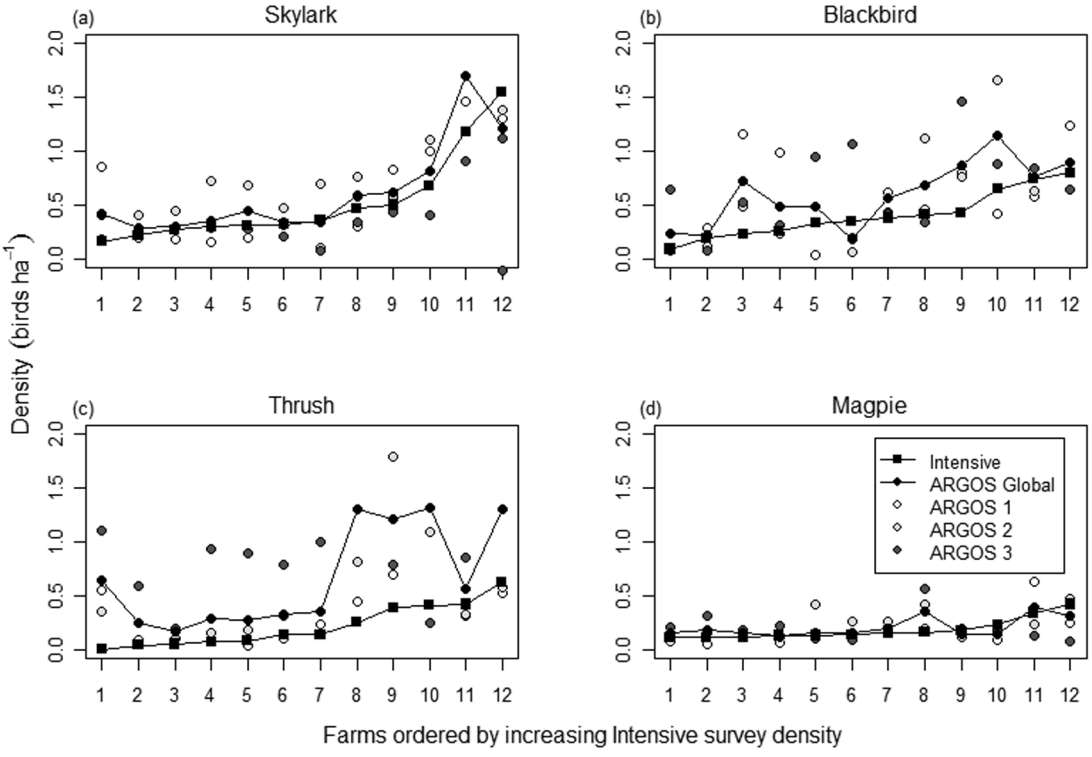

Figure 2. Population density estimates per farm for each species (a–d) for the three sets of survey analyses (intensive, ARGOS Global, and ARGOS 1–3). Farms were ordered by increasing 'intensive survey' density estimate for each species.

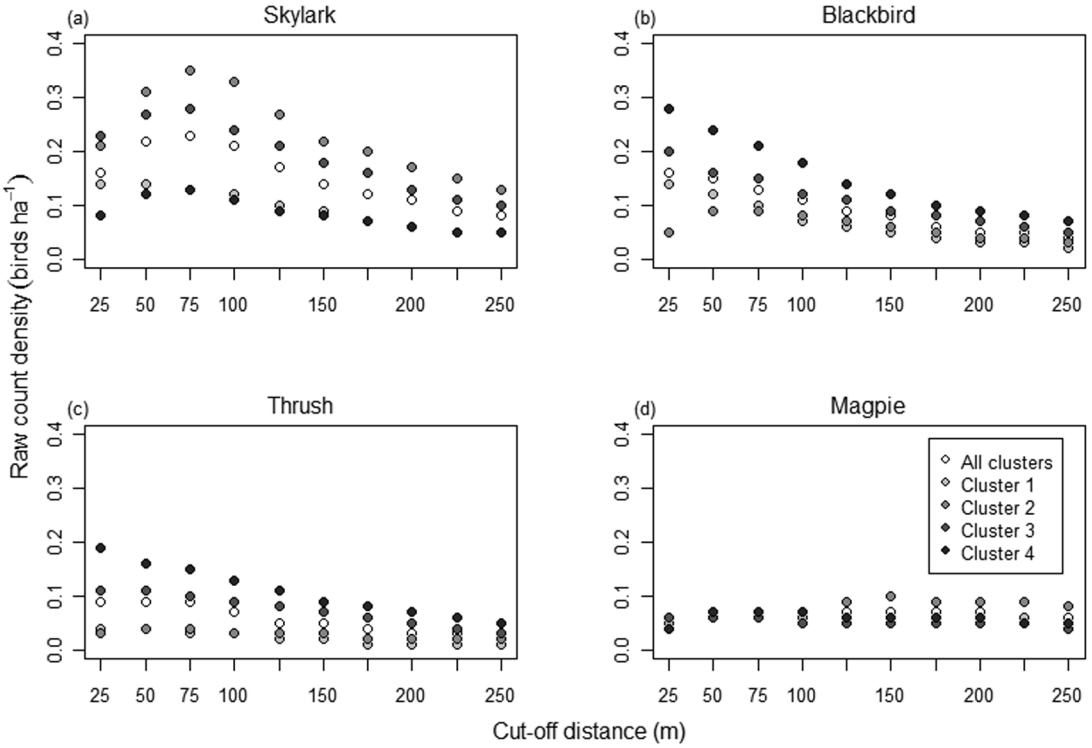

Figure 3. Raw count estimates per hectare for the four farm clusters and for all clusters by species, showing the dependence of estimated density on chosen truncation distance. Detections were selected from the 'intensive survey' database out to the chosen distance and divided by surveyed area. White circles denote results averaged over all farms, other data series are cluster-level.

animals (Buckland et al. 2001, 2004). In magpies, estimates remained at a nearly constant level or showed a very slight increase with increasing distance (Fig. 3).

Plotting the explained variance (R^2) of the linear regression of raw count density on distance-sampling density versus cut-off distance yielded similar results for skylarks and thrushes (Fig. 4); model fit was at its lowest ($R^2 \sim 0.8$) at 25 m, rose to a maximum (>0.95) at 50 m (skylark) or 100 m (thrush), and kept to a similar level after that. In blackbirds, model fit declined sharply (to a minimum of 0.62) after a distance of 50 m. In magpies, model fit was much lower (0.45–0.65) and more erratically related to cut-off distance (Fig. 4).

Average farm-level vegetation cover is a primary source of detectability differences in farmland bird monitoring (Weller et al. 2012). Adding this parameter produced virtually no improvement to the skylark and thrush models. For magpies, the explained variance percentage was still lower than for the other species, but increased by 5–7.5% compared with the base model. For blackbirds, model fit improved strongly with the introduction of the vegetation cover covariate, to an average R^2 of 0.86 and a distribution shape similar to that for skylarks and thrushes. A test for differences (two-sided *t*-test) in R^2 at 25 m, 50 m, and 200 m showed no significant differences at any of these levels for skylarks, thrushes and magpies, but significant differences for blackbirds at 50 m ($p = 0.051$) and 200 m ($p = 0.002$). The best model fit was found at 50 m for skylarks (either model, $R^2 = 0.96$) and blackbirds (with covariate vegetation cover, $R^2 = 0.93$), at 100 m for thrushes (either model, $R^2 = 0.98$), and at 175 m for magpies (with covariate vegetation cover, $R^2 = 0.66$; Fig. 4).

Discussion

Sources of uncertainty in the survey results

The two survey types discussed here represent two different approaches to constructing an image of bird populations on sheep & beef farms. The ARGOS monitoring scheme is an ongoing longitudinal programme that apportions its effort across a large number of sites, tracking long-term population dynamics. The intensive survey delivered one-off population estimates of high resolution at a few sites. Because these designs put different constraints on modelling procedure, replication, and treatment of field parameters, imprecision and bias arising in each of these areas were addressed with differing degrees of success in the two surveys.

Precision

The primary indicators of precision in distance-sampling-based density estimates derive from model fit, and are readily available from program Distance. Model-based precision was good in both surveys (Fig. 1a,b). Average detection probability *p* was similar across series and in a range frequently encountered in distance models of bird populations (Diefenbach et al. 2003; Norvell et al. 2003; Newson et al. 2005, 2008; Buckland 2006; Gale et al. 2009). Precision of detectability CV(*p*) was better for the ARGOS monitoring scheme than for the intensive survey, and better for ARGOS Global than ARGOS 1–3, showing the beneficial effects of increasing the sample size available for model fitting (which reduces variance in the detectability estimate).

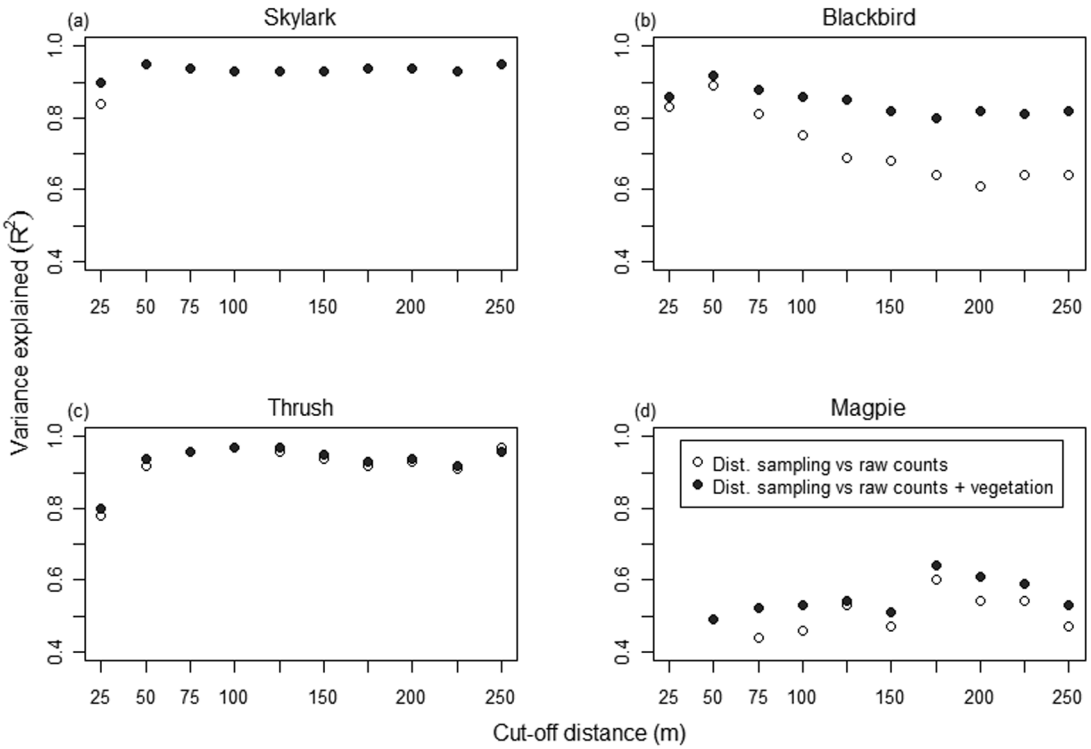

Figure 4. Variance explained (R^2) by regressing the constructed raw count estimates for each farm at different cut-off distances, on the equivalent set of distance-sampling density estimates, by species in a linear regression. White circles denote data series where average woody vegetation cover in surveyed (raw count) area was added as a covariate, filled circles denote data series without this covariate.

The variances of both detection probability and encounter rate (number of detections over total transect length) form components of the variance of the final density estimate (Buckland et al. 2001). It is noticeable that high encounter rate variance in the ARGOS surveys resulted in low precision in *D*, while the components were more balanced in the intensive survey (Fig. 1c,d). Hence, while CV(*D*) in the latter was well within the range for useful estimation (Norvell et al. 2003; Newson et al. 2005, 2008; Buckland 2006; Somershoe et al. 2006), this was questionable for the individual ARGOS years in particular.

Large sample sizes for model fitting in the ARGOS scheme were generated by pooling across all farms for ARGOS 1–3, and further across the three surveys (repeat visits) for ARGOS Global. The intensive survey, with only a third the number of farms, achieved suitable sample sizes by pooling across a larger number of repeat visits only. These two methods were not equivalent in their effects on estimate precision because they differed in the distribution of monitoring effort to individual farms. The total length of transects walked per farm during the intensive survey was seven to eight times larger than for each of the ARGOS survey years (c. 36 000 m vs c. 4600 m per farm respectively) (Weller 2009; MacLeod et al. 2012b). This translates into strong differences in the likelihood of encountering a representative sample of each site's bird population – unrealistically high or low density estimates would have resulted more frequently in ARGOS 1–3. A measure of this is variance in encounter rate, which was largest in the individual ARGOS years, reduced in the global model, and smallest in the intensive survey (Fig. 1d). Another effect of increased replication was better farm coverage, increasing the precision of farm-wide density estimates by sampling the area more completely. While encounter rate variance is incorporated into the variance of *D* and as such accounted for in the main precision criterion, a possible lack of area coverage must be realised by the user themselves. Even if transects are assigned at random, low replication might then inject a random bias into the farm-level estimate. The efficient yet thinly spread application of effort in the ARGOS scheme would be more susceptible to this than a design employing more repeat visits.

Bias

The existence of bias in the estimates is less readily assessed. In the absence of knowledge of true population densities, it cannot be said with certainty which survey design provides the more accurate results. However, comparisons between the different surveys' results can be made with care to allow the identification of several possible sources of bias.

When ARGOS estimates differed from intensive-survey values, they were generally larger. This was true for the individual surveys (ARGOS 1–3) for each species, and for the globally modelled result for blackbirds and thrushes (Fig. 2). This was probably not an effect of the restriction of ARGOS farm visits to the summer months while intensive-survey visits sampled all times of the year. In the GLM, intensive-survey results did not differ significantly from those of the same survey's breeding season only – on the surveyed farms, population densities of the focal species tended to be on a par with or slightly below average during December–January, with variably both higher and lower numbers during the rest of the year (Weller 2009).

Both survey methods are likely to have been affected by two main sources of bias: changes in conditions affecting detection probability over the course of the surveys, and model-based influences.

Changes in detectability

Differences in the bird detection and identification abilities of observers can have a pronounced influence on the number and attribution of detections (Sauer et al. 1994; Kendall et al. 1996; Buckland et al. 2001; Diefenbach et al. 2003; Jiguet 2009). Changes in field parameters that influence detectability were accommodated in both surveys by the inclusion of covariates denoting observers, vegetation cover, and other parameters (Marques & Buckland 2003; MacLeod et al. 2012b; Weller et al. 2012). While differences between individual observers can be mitigated by such methods, they cannot be accounted for entirely (additionally, any detection bias shared by *all* observers on a survey is not controlled by this approach). The ARGOS scheme is likely to be more prone to this bias than the intensive survey since it was possible to have c. 50% of transects performed by the same observer in the latter (Weller et al. 2012).

Modelling of detectability covariates is largely ineffective if detectability changes between individually analysed surveys. Survey-level results for ARGOS 1–3 must be expected to be conflated with inter-survey detectability changes (again mostly caused by changing observers) that were partly or wholly removed in ARGOS Global, which may account for some of the striking differences between these surveys and the global and intensive estimates (Fig. 2). However, the ARGOS 1–3 average was not significantly different from ARGOS Global for any species; as the difference between these sets is individual versus global model fitting, this indicates that no very large improvement was made to the estimates from being able to accommodate inter-survey detectability variations.

It is notable that among the four focal species there was a marked split regarding the amount of inter-survey discrepancy, with blackbirds and thrushes showing many more differences than skylarks, and there being hardly any for magpies (Fig. 2). This directly parallels the order of detection and identification difficulty among these species, with magpies being highly conspicuous, vocal, and found in open areas, skylarks being slightly harder to locate in similar habitats due to their smaller size, and blackbirds and thrushes being frequently hidden in vegetation, with an added possibility of confusion between males of the two species by song, and females by sight (Weller 2009). It is therefore apparent that the accuracy of density estimates also depends on the characteristics of the species monitored.

Pooling and model fitting

In the ARGOS monitoring scheme, models were fitted to a dataset pooled across all farms and then evaluated for individual farms, whereas in the intensive survey models were fitted at the farm level. This difference in treatment, born of the necessity to achieve satisfactory sample sizes, favoured the ARGOS scheme with a large dataset (Table 1) that enabled high precision in model fitting and hence estimation of detectability (Fig. 1b), while modelling for the intensive survey ran the risk of generating comparatively less appropriate models for the same farms because they were based on fewer total sightings. The process of model selection by parsimony criterion (AIC) employed in both surveys (MacLeod et al. 2012b; Weller et al. 2012) would be more likely to result in a simplified model the smaller the size of the modelled dataset, and the chance for the effect of a covariate to be correctly assessed and included in the model would be reduced. This might have led to a bias away from true densities in the intensive-survey results. Its strength is difficult to estimate; a reduction of modelled dataset size to a third of Global size seems to have been of little consequence (no significant differences between ARGOS Global and 3-year ARGOS average in the GLM), but intensive survey sample sizes were on average only a sixth of this and sometimes close to the recommended minimum, making a detrimental effect probable.

Sample sizes were not large enough in the ARGOS surveys to allow the fitting of models at the farm level, which was possible for all farms in the intensive survey due to the higher number of surveys per farm. These models had the benefit of generating farm-specific detection functions and covariates. Considering the differences in the habitat characteristics and vegetation make-up of the surveyed farms, and that the amount of farm-level woody vegetation is the primary influence on detectability in these surveys (Weller et al. 2012), this can be expected to have resulted in some improvement of the estimates compared with those derived from global models. However, as both the global and 3-year average estimates encompass a season (2009/10) that was not covered by the intensive survey, it is not possible to empirically determine whether observed differences between these surveys are due to methodology or genuine population changes (e.g. the thrush data for ARGOS 3 (Fig. 2c) might indicate a general increase in thrush numbers in that year). A comparison of the 2-year average (which excludes this season) to the intensive estimates is more valid. The lack of significant differences between that set and the intensive survey for any species in the GLM is a good indication that bias due to modelling method was, if present, not very noticeable.

Bias in raw counts

In tracking changes in population densities over time, the relative differences between successive estimates are of primary interest. Relative indices are not considered to be unbiased; rather the only requirement is that bias between surveys remains constant. For such a purpose, raw counts might represent an efficient way to evaluate the data generated in surveys such as those discussed here (Johnson 2008). But even when treating such estimates as purely relative indices, they come hedged with specific vulnerabilities. An obvious characteristic of raw counts is one of diminishing calculated density with increasing cut-off distance (farthest distance from which detections are taken). Increases in cut-off distance, and hence area of inference, equal reductions in estimates as areas with progressively fewer detections are being added in (Fig. 3). To achieve comparability between surveys, the same cut-off distance for detections would have to be used across the board.

Yet it would still be unwarranted to assume that such indices are necessarily in some constant ratio to the true density, i.e. that all biases, known or unknown, remain constant for the duration and extent of the survey. An empirical test of the existence of such a ratio is to try to establish a systematic relationship between a set of raw count estimates and an equivalent set of more accurate densities. For three of the four focal species discussed here, a very good linear relationship could be established between the raw count estimates taken from the intensive-survey database and the equivalent farm-level distance sampling estimates (Fig. 4). In skylarks and thrushes, and in blackbirds when including the additional

parameter of woody vegetation cover percentage, the variance explained by the predictor(s) never fell below 80% and was usually substantially higher. Beyond species-specific cut-off distances, the achievable fit (i.e. the suitability of a single equation to describe the ratio for all farms) remained roughly constant within 10% of the maximum-achieved R^2 (Fig. 4). This demonstrates a basic suitability of count indices for these species as relative density indicators, but shows the same approach to be questionable for magpies. Information from such regressions can be used to choose an optimal cut-off distance for index calculations, the main criterion being R^2 maximisation. For skylarks, blackbirds and thrushes, this suggests a common cut-off distance of 50 m. If magpies are included, the cut-off distance is somewhat larger (175 m), with presumably negative results for overall detection reliability due to observers dividing their attention over a wider area.

The results for blackbirds provide an example of the influence of additional predictors on computed density, as a satisfactory ratio between estimates could only be achieved when adding a parameter for average woody vegetation cover into the equation (Fig. 4). This factor has been shown to be of great importance for farm-level detectability (Weller et al. 2012) and thus might have been expected to be similarly important in thrushes, which for no clear reason was not the case. In magpies, models with the added predictor showed more consistency of fit than the base model, but the effect was irregular and fit was still lower than in the other species ($R^2 < 0.7$). A possible explanation for this is that at the larger spatial scales that came into play with magpies due to conspicuousness at long distances (Weller et al. 2012), topographic visual disturbances such as hills or depressions that were not recorded or modelled gained an impact. The need for additional detectability predictors should likely be taken to indicate that the species in question would be more efficiently monitored using methods like distance sampling that easily incorporate such parameters, as the main benefit of raw counts lies in their ease of implementation, a benefit which might quickly be nullified if additional data have to be collected and modelled.

Conclusion and recommendations

The trade-offs in the design of both the ARGOS monitoring scheme and the intensive survey resulted in differences in the precision of their density estimates and their susceptibility to bias. The ARGOS surveys generated less precise farm-level density estimates than the intensive survey, which also benefitted from higher replication of visits and more complete farm coverage. Farm-specific models might have provided a benefit to accuracy in the intensive survey, but smaller datasets meant that appropriate models could be fitted less often than in the ARGOS scheme. Both designs suffered from hard-to-quantify biases in detectability due to changing observers, which likely affected the ARGOS surveys somewhat more.

Overall there seems to be a higher vulnerability to bias in absolute farm-level estimates in the ARGOS surveys than in the intensive survey. This is not unexpected considering that approximately the same amount of effort (in transect length walked) was spent on 36 farms in the former as on 12 farms in the latter. The design of the ARGOS scheme as a longitudinal study on many properties thus comes at a cost to both precision and accuracy that are incurred to a lesser extent in more focused designs. However, the long-term nature of ARGOS makes it possible to identify potential problems and address them in the ongoing monitoring by updating the survey design (although this would necessarily trade off against data comparability across years).

The ARGOS scheme would undoubtedly gain from a higher number of visits to each farm in a given year. Better temporal replication would yield a further increase in sample sizes and decrease encounter-rate variance, and thus improved density estimate precision. As the precision of the model estimates seems to be directly linked to sample sizes, this is probably the single most effective improvement that can be made to the quality of estimates. The increased spatial coverage would also enhance farm coverage and lead to a more representative sampling of farm-level populations. The single visit per survey in the current design took place during the main breeding season of most bird species present on the farms, which is often recommended for bird surveys if not all seasons can be sampled (Svensson 2000; Newson et al. 2008). Sampling events bracketing the breeding season might provide better information about the causes of detected population changes, while the establishment of long-term trends would be best served by identifying and sampling the time of year with the least variability in bird numbers, so programme priorities should be considered.

The ARGOS scheme is well suited to the purpose of tracking global population trends, as it can capitalise on the substantial pooled sample sizes gained from surveying a large number of sites. The property of pooling robustness (Buckland et al. 2001, 2004) inherent in distance sampling models ensures that even though individual farm datasets might differ in their precision and detectability characteristics, a pooled estimate will essentially be unbiased by such influences (Buckland et al. 2004). However, as shown, employing survey-level models for the generation of annual estimates (as was done for ARGOS 1–3) is not without problems, as true variation in population densities and variation in detectability are conflated between surveys. This can be addressed by deriving survey-level estimates from a cross-survey pooled function (as generated for ARGOS Global). However, by applying a model to a survey that may be strongly influenced by circumstances only present in other surveys, such an approach runs the risk of introducing bias. Hence, it will likely be sounder to aim for annual estimates, while increasing survey effort and trying to control for or standardise as many detectability influences as possible.

The use of raw count data as relative population indices offers the possibility of more quickly attaining a dataset of sufficient size for trend estimation, by allowing simpler field methods and less effort spent on modelling. Of the four focal species, reasonable results using such an approach were obtained for skylarks and thrushes, and also for blackbirds after additional correction for habitat influences, while index estimates for magpies were unreliable (Fig. 4). These differences show a split in index reliability that seems counter-intuitive when the characteristics of the species' density estimates are considered (Fig. 2), and indicate the need for a similar comparative analysis to be carried out for any species intended to be monitored in this manner. This need for calibration estimates suggests that raw counts might be best employed as an extension for a well-established programme to add coverage for additional areas or times of year, rather than as a replacement. Regardless, their use would make it imperative to standardise all controllable detectability influences as far as possible, with observer identity being the main issue.

Observer error was a prominent problem in both surveys. Continuity of observers between surveys seems to be the most

desirable improvement in long-term designs, although likely difficult to implement. Among other benefits, the identification of genuine inter-survey population density changes would become easier if there was no possibility of changes in numbers across the board for a certain species being based on unrealised biases in detection or identification in a fresh group of observers. The results for the four focal species also demonstrate that the accuracy of density estimates will, among other things, depend on the degree of challenge that species pose to observers. It should be realised that a multi-species survey will likely deliver estimates of different quality for different species, and that this will not necessarily be apparent in measures of variance. The use of specialised pilot or complementary surveys of the same sites to identify such issues might be a sensible safeguard. Further surveys covering other species on the properties monitored by ARGOS are likely to become available in the future (e.g. Meadows et al. 2012) and will be useful for such purposes. The nature of ARGOS as a long-term project with a large permanent group of participating farms offers the opportunity of carrying out survey designs of unusual scope, and the availability of supplementary results from other studies taking place within the same programme will likely prove to be of continuing benefit.

Acknowledgements

Work on this project was made possible by a University of Otago Postgraduate Scholarship. I am grateful to Catriona MacLeod and Henrik Moller for helpful and incisive comments on this paper's first draft, and to Simon Butler and an anonymous reviewer for insightful notes on a later version. The ARGOS project work was funded by the Foundation for Research, Science and Technology (Contract No. AGRB0301) with financial assistance from the Certified Organic Kiwifruit Producers Association, Fonterra, Merino New Zealand Inc., a meat packing company, Te Rūnanga o Ngāi Tahu, and ZESPRI Innovation Company.

References

Alldredge MW, Simons TR, Pollock KH 2007. A field evaluation of distance measurement error in auditory avian point count surveys. Journal of Wildlife Management 71: 2759–2766.

Buckland ST 2006. Point-transect surveys for songbirds: robust methodologies. The Auk 123: 345–357.

Buckland ST, Anderson DR, Burnham KP, Laake JL, Borchers DL, Thomas L 2001. Introduction to distance sampling: estimating abundance of biological populations. Oxford, Oxford University Press. 432 p.

Buckland ST, Anderson DR, Burnham KP, Laake JL, Borchers DL, Thomas L 2004. Advanced distance sampling: estimating abundance of biological populations. Oxford, Oxford University Press. 416 p.

Diefenbach DR, Brauning DW, Mattice JA 2003. Variability in grassland bird counts related to observer differences and species detection rates. The Auk 120: 1168–1179.

Gale GA, Round PD, Pierce AJ, Nimnuan S, Pattanavibool A, Brockelman WY 2009. A field test of distance sampling methods for a tropical forest bird community. The Auk 126: 439–448.

Jiguet F 2009. Method learning caused a first-time observer effect in a newly started breeding bird survey. Bird Study 56: 253–258.

Johnson DH 1999. Statistical considerations in monitoring birds over large areas. In: Bonney R, Pashley DN, Cooper RJ, Niles L eds Strategies for bird conservation: The Partners in Flight planning process. NY, Cornell Lab of Ornithology.

Johnson DH 2008. In defense of indices: the case of bird surveys. Journal of Wildlife Management 72: 857–868.

Kendall WL, Peterjohn BG, Sauer JR 1996. First-time observer effects in the North American Breeding Bird Survey. The Auk 113: 823–829.

MacLeod CJ, Greene T, MacKenzie DI, Allen RB 2012a. Monitoring widespread and common bird species on New Zealand's conservation lands: a pilot study. New Zealand Journal of Ecology 36: 300-311.

MacLeod CM, Blackwell G, Weller F, Moller H 2012b. Designing a bird monitoring scheme for New Zealand's agricultural sectors. New Zealand Journal of Ecology 36: 312-323.

Marques FFC, Buckland ST 2003. Incorporating covariates into standard line transect analyses. Biometrics 59: 924–935.

Newson SE, Woodburn RJW, Noble DG, Baillie SR, Gregory RD 2005. Evaluating the Breeding Bird Survey for producing national population size and density estimates. Bird Study 52: 42–54.

Newson SE, Evans KL, Noble DG, Greenwood JJD, Gaston KJ 2008. Use of distance sampling to improve estimates of national population sizes for common and widespread breeding birds in the UK. Journal of Applied Ecology 45: 1330–1338.

Norvell RE, Howe FP, Parrish JR 2003. A seven-year comparison of relative-abundance and distance-sampling methods. The Auk 120: 1013–1028.

R Development Core Team 2010. R: a language and environment for statistical computing. Version 2.12.1. Vienna, Austria, R Foundation for Statistical Computing.

Sauer JR, Peterjohn BG, Link WA 1994. Observer differences in the North American Breeding Bird Survey. The Auk 111: 50–62.

Somershoe SG, Twedt DJ, Reid B 2006. Combining breeding bird survey and distance sampling to estimate density of migrant and breeding birds. The Condor 108: 691–699.

Spurr EB, Borkin KM, Drew KW 2012. Line-transect distance sampling compared with fixed-width strip-transect counts for assessing tomtit *(Petroica macrocephala)* population trends. New Zealand Journal of Ecology 36: 365-370.

Svensson SE 2000. European bird monitoring: Geographical scales and sampling strategies. The Ring 22(2): 3–23.

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.

Weller FG 2009. Monitoring bird abundance on New Zealand pastoral farms. Unpublished PhD thesis, University of Otago, Dunedin, New Zealand. 190 p.

Weller F, Blackwell G, Moller H 2012. Detection probability for estimating bird density on New Zealand sheep & beef farms. New Zealand Journal of Ecology 36: 371-381.

White GC 2005. Correcting wildlife counts using detection probabilities. Wildlife Research 32: 211–216.