Head to Head

On the design of monitoring programs and the use of population indices: a reply to Ellingson and Lukacs

By Richard L. Hutto and Jock S. Young

Key words  convenience sampling, distance sampling, index values, monitoring, point count, study design, survey methods

Theoretical and applied ecology represent large and complex disciplines, and it is easy to get lost in the details, particularly the analytical details (Anderson 2001:1294).

In our paper (Hutto and Young 2002), we outlined why regional monitoring programs are needed; why birds should be used as monitoring tools; why we designed our monitoring program (including the distribution of long-term monitoring points) the way we did; the benefits associated with monitoring from permanent points on a biennial rather than an annual basis; why we have gained support from management and may have been able to exert some influence on management by including a short-term, effects-oriented component within our monitoring program; and what we need to do to improve the effectiveness of adaptive management through monitoring results. Comments by Ellingson and Lukacs were focused on just two aspects of the larger program—placement of long-term monitoring points and use of indices of bird abundance in our data analyses. Given that the goal of our paper was "to focus discussion about the overall goals and methodological issues associated with monitoring," we welcome the chance to further highlight the strengths of our program and to defend our use of population indices.

With respect to the use of population indices, Ellingson and Lukacs have undoubtedly overstated the potential problem. By discrediting our results because they are based on indices of abundance, they discredit all other studies in avian ecology that have been based on indices of bird abundance (and that would certainly be the vast majority of them). We disagree that inferences drawn from such data fail "to provide a positive exemplar for future studies" (Ellingson and Lukacs 2003:896). Ultimately, it comes down to common-sense evaluation of the possible sources of bias in the process of science. As Steidl et al. (2000:518) recently stated, "recipes for quantitative analyses are no substitute for critical thinking in wildlife science." The idea that methodologies must follow an exact formula to be defensible is short-changing the process of science. No methods are perfect, and all involve assumptions. Indeed, we stand behind every general biological pattern that we have published. In no way does our particular choice of a method of analysis impugn the entire Northern Region Landbird Monitoring Program (NRLMP), as Ellingson and Lukacs imply. Instead of suggesting that our data are not credible, we would much rather have seen some empirical evidence to confirm that our conclusions were false or misleading.

With respect to the design of a research and monitoring program, the overall design will be largely, if not entirely, independent of whether one chooses to adjust the data to account for potential detectability problems. Let us first defend the NRLMP design, then move on to defend our use of population indices.

Program design and scientific methodology

The current NRLMP evolved through several years of pilot research dealing with the costs and
benefits of on- vs. off-road counts (Hutto et al. 1995), analyses of the tradeoff in statistical power associated with an annual vs. biennial long-term monitoring design (internal report to the United States Forest Service [USFS] Region 1, April 1997), the costs and benefits of purely random point placement for long-term monitoring (based on independent studies in USFS Regions 1 and 4), and through additional years of minor subsequent adjustments. Ellingson and Lukacs labeled our sampling design a form of "convenience sampling," a pejorative term that has been used to refer to any type of non-probability-based sampling (Peterson et al. 1999, Anderson 2001), and also a term that seems to belittle our efforts by equating our design with that of a study conducted on a street corner or in and around camp. The suggestion that we chose to abandon all probabilistic sampling in favor of pure convenience is simply wrong. Our long-term transects were selected by delineating strata, selecting points within strata using simple random sampling, and then designating the nearest point on a tertiary road or trail as the transect start point. Furthermore, all of our short-term, alternate-year studies have involved randomly selected, off-road treatment and control plots.

With respect to potential roadside bias, the real problem Ellingson and Lukacs have is with the sampling frame (tertiary roads and trails), not the sampling design. We devoted a section of our paper to the use of roads and trails (pp. 744–745), and we understand and have written about the potential problem of roadside bias (Hutto et al. 1995). There are numerous practical reasons, however, why a well-designed study need not be based on truly random sampling, not the least of which are sample size and safety. We conducted a pilot year-long study of this issue and found that even something as simple as orienting transects randomly from the start point resulted in visits to roughly half the number of points we could visit on roadside transects. Moreover, other struggles (including injuries) associated with a multi-year effort by USFS Region 4 to design a monitoring program based on truly random sampling resulted in an even smaller sample size. Given these practical considerations, we opted for a roadside design. Indeed, we have yet to see success with a truly random design for monitoring. We need to re-iterate two additional thoughts here: (1) while estimates of long-term trends in population size may be biased by restriction of the survey to roads and trails, we suspect that habitat relationships established for most species are less subject to bias by the use of roadside counts; (2) the solution to a potential roadside bias that some feel can be avoided by positioning points, say, 200 m off-road will not eliminate a roadside bias. It will merely introduce whatever bias might be associated with positioning points near (200 m from) roads. We acknowledged in our paper that the issue needs further study, but also noted (and re-emphasize here) that professional consensus favors the tradeoff of some bias for the gain in logistical simplicity associated with roadside counts (Ralph et al. 1995, Johnson 2000).

A realistic monitoring program is a program of compromises, and extensive experience from the field has made us aware of just how hard it is to achieve the kind of elegance and perfection in design that one can achieve by working entirely in the modeling world. All aspects of the sampling design and data collection associated with the NRLMP were acknowledged by the statisticians we consulted at the outset to be as strong as we could hope for, given the constraints we outlined earlier. Moreover, the program design would be identical whether our data analyses were to include distance estimation (see below) or not. We stand by our program design as solid, rigorous, and perhaps even enviable in the levels of true replication (see Sallabanks et al. 2000) that we are able to achieve in association with our management effects studies.

A final comment involving the design of most monitoring programs is that analyses are generally meant to be exploratory rather than confirmatory. As stated clearly in our paper, one of the strengths of the NRLMP is that it reveals issues in need of confirmatory study. As we write, there are now numerous ongoing graduate studies of issues that we raised on the basis of data gathered through the NRLMP. If the conclusions from any of our studies were off-track, we will find out in short order. Either way, we believe our approach to monitoring is an especially strong way to proceed toward solid scientific understanding.

Indices of abundance vs. density estimates

Ellingson and Lukacs note (correctly) that we use our criticisms of distance sampling as a rationalization for using index values, but they then go on to claim (incorrectly) that our concerns with the distance sampling and the violations of its assumptions are "misunderstandings of modern distance
sampling.” We suggest instead that their rejection of our concerns reflects their own misunderstanding of the realities associated with bird survey work. In this section we will make as clear as possible that our notions about distance sampling and the problems associated with its use are not at all mistaken. Below we outline our biggest concerns with distance sampling—that none of the important assumptions associated with its use are likely to be met.

Model assumption 1—all birds are detected at a distance=0. The basal level of detectability (the probability of detection at or near each point) may not equal 1.0 and may be unequal among sites or habitats. Even structurally simple vegetation types (e.g., grasslands) often reveal birds that have gone unnoticed until one practically steps on them, so this problem is probably not trivial. McShea and Rappole (1997) found that singing birds within 50 m of observers had an average song rate only half that of singing birds beyond 50 m, and Bye et al. (2001) found significant differences in the numbers of birds recorded with or without the presence of a human observer. Clearly, the conspicuousness of a bird is affected by an observer’s presence, especially those occurring close to an observer.

Model assumption 2—bird movement. Birds move routinely and frequently in the course of their daily activities. If they are moving constantly, or every one to two minutes (e.g., Wasserman 1980), then they may be singing from several different places during a single count. It is very difficult to determine whether songs in different places are from the same individual if the bird is never seen (as is usually the case in forested landscapes), unless it can be surmised from the chronology that there is only a single individual of the species in the area. Thus, over-estimating the number of birds detected is probably common. Random movements may not otherwise affect distance sampling results, but responsive movements clearly do. We have no way of knowing how much responsive movement there is, but it would be naïve to think that it is negligible. Birds clearly respond to the presence of observers, by either reduced singing or by movement (McShea and Rappole 1997, Bye et al. 2001). Since birds routinely fly across their territory, it does not take much effort to imagine them doing so to avoid a temporary disturbance. We may fool ourselves into thinking birds do not respond to our presence just because we continue to hear birds singing around us during a count.

Many detection functions for common passerine species are hump-shaped, even after controlling for area (see figure 2.2 in Buckland et al. 2001). Why are they hump-shaped? Are there really fewer birds near where researchers always happen to place their points, or are the birds nearer the observer less detectable, or do birds move before they are detected? Although bird movement is most assuredly part of the answer, what sort of movement? Is the hump present because some birds move away from the observer, or because some birds from afar move in for a better look? The answer influences one’s estimate of density, and the answer is unclear. Moreover, not only is a major assumption violated, it is uncertain what kind of movement is involved, so one does not even know the direction of bias. Also, if birds are more likely to move away from an observer in open habitats than in forests (where they may feel relatively safe, especially in the tree canopy), then the bias will differ among habitats. One might argue that the potential problem associated with bird movement should be trivial as long as birds behave similarly in all environments. The problem is that birds do not behave similarly in all situations (habitats, weather conditions, etc.). Thus, not only will density estimates be biased because of bird movement, but the estimates will probably be biased differently in different habitats.

Also critical is the issue of bird movement to within the range of detectability before the end of the count period. The tallying of detections during a point count is a cumulative process; observers add birds as they detect them, but do not subtract birds if they leave (Johnson 1995). Moreover, the birds added during the second half of a 10-minute count (a point-count duration common to most research and many monitoring studies) almost certainly include some birds that have moved from outside to within detection range. This will clearly bias any density estimate upward, with the amount depending entirely on how much and how far birds move. Distance sampling may be a great tool for animals that are essentially stationary during a transect count, but given the way birds dart from one end of a canyon to another in a matter of seconds, the accumulation of detections during passage along a transect or during a 10-minute count is probably not a trivial problem.

Ellingson and Lukacs suggest (2003:899) that responsive movement is no problem because all we have to do is “design a sampling protocol that
reduces movement to a tolerable level," as if the problem could be solved if only we were clever enough. The effectiveness of the "often suitable" approach that Ellingson and Lukacs recommend (wait at the point for a few minutes before starting the count, thereby allowing birds to settle down) has never been tested, and it fails to eliminate the potential biases due to bird movement anyway. Birds will still move toward or away from the observer (some even to beyond detection range) before they "settle down," and they will continue to accumulate from beyond to within detection range during the count. The birds that move toward or away from the observer but stay within detection range will create a humped detection function (or "donut effect") because of their responsive movement. All of this movement results in density estimates that are biased to an unknown extent.

**Model assumption 3—distance estimates are accurate.** We disagree strongly with Ellingson and Lukacs (2003:899) that "the problem of obtaining suitable distance data in the field is one that is easily dealt with in most situations." We and our trained observers have conducted point counts for many years, and from our experience we believe that the problem of distance estimation is much greater than most of us like to admit. We can train observers to estimate the distance to an object in view, or we can give them laser rangefinders, but the fact remains that most birds are detected by sound alone (87% in our program). Estimating distance to sounds is not only very imprecise but has biases of variable and unpredictable magnitude and direction. Because the only clue available for estimating distance is the volume of a song, we have found that observers actually tend to systematically overestimate distances in habitats with dense vegetation, perhaps because sound attenuation makes birds sound farther away. Thus, the nature and extent of bias may itself be a product of vegetation type.

We have tracked down many birds after estimating distances to them and are frequently in error by 30% or more. Estimates by observers with less experience may be even less accurate. Ellingson and Lukacs suggest that placement of detection distance estimates into groups or intervals should reduce error, but this practice simply produces larger (though less frequent) errors. Whether it reduces bias is unclear. Furthermore, the use of large distance intervals reduces one's ability to estimate a detection function accurately. If our ability to estimate distances is poor, and if errors tend to be in one direction for a given person, then it does not matter if one lumps data into categories or not—indices of abundance and density estimates will both lead to biased estimates.

Ellingson and Lukacs suggest several options (e.g., rangefinders, pacing, reference flagging) to aid with distance estimation while working with slow-moving animals in open habitats, but we have found these to be ineffective solutions for forest songbirds. For example, if an observer cannot even estimate which row of trees a bird is singing from, there is nothing a rangefinder can do to help. Pacing to verify distances would certainly help, but this cannot be done during a count, and cannot be accomplished for more than perhaps one or two visually located birds per point after a count.

**Model assumption 4—detection functions are based on independent samples.** The inclusion of data from the same point multiple times violates an assumption basic to all statistical sampling. The detection function is meant to represent an accurate picture of how detectability decreases with distance from the observer, and ideally that picture must be drawn from independently drawn random samples. Unfortunately, many biologists choose to revisit points or transects, and if, for example, the same bird sings from the same place each time a point is revisited, then the shape of the detection function will clearly be an artifact of non-independent distance estimates. Buckland et al. (2001:36) claim that distance sampling methods are "remarkably robust" to violations of this assumption (probably because most birds move to more or less random locations within their territories from day to day), but we worry that the non-independence assumption could still be problematic with the small sample sizes used in many studies. Moreover, when sample sizes are small, a researcher may be most tempted to treat multiple visits to a point as independent samples, precisely when doing so may exacerbate the problems of non-independence.

**A practical issue—sample sizes.** If one chooses to use distance sampling to solve the theoretical problem of detectability bias among, say, habitat types, one would need enough data to model the detection function for each habitat separately at the very least. Even better, one would like to model simultaneously the interactions of all factors (observers, weather, etc.) that might influence detectability (Thomas et al. 2002a). Indeed, we
have found that the average distance estimates for different observers vary over a two-fold range (i.e., the average distance estimates to a particular species for the observer at one extreme was only half that of the observer at the other extreme). Unfortunately, the numbers of detections of most bird species associated with the average scientific study probably precludes such analyses.

The sample size limitation is well illustrated by a tally of the actual distance data from a series of bird species that range from rare to common in surveys conducted in association with our monitoring program in 2000. In that year we had 13 observers, and each visited 236–330 points—the maximum sample size that can reasonably be expected by a single observer in a breeding season. What proportion of the bird species detected generated the recommended sample size of at least 80 detections (Buckland et al. 2001) needed to obtain a reliable picture of the relationship between number of detections and distance? Of 153 species we detected, 57 had greater than 80 detections. However, only 29 of these species (19%) were detected at least 80 times by any single observer, and only 13 species (8%) had enough detections for modeling by more than half of the 13 observers.

Clearly, even relatively common species yield a sample size that is far too small for constructing an accurate detection function, especially if it is necessary to subdivide data to account for habitat and observer differences. Moreover, because most research projects are conducted by individual investigators who do not have the luxury of being able to hire multiple field technicians to work for them, most projects will also fail to generate sample sizes needed to build reasonable models. We agree that sample size problems do not in any way detract from the theoretical validity of distance sampling; we only want to emphasize here that sample sizes are a real problem in practice. Ellingson and Lukacs acknowledge that sample sizes are often too small for inference from distance sampling, but they suggest that researchers go ahead and use program DISTANCE (Thomas et al. 2002b) anyway; and let confidence intervals reveal the uncertainty. For small samples, however, simple means have much smaller standard errors than distance-based estimates because of the extra parameters that must be estimated in the model-based approach. The suggestion to use distance-based methods despite small samples ignores this reality.

Another practical issue—results from program DISTANCE are sensitive to variations in model selection and data grouping. We are not confident that most practitioners have the biological sophistication needed to select the most appropriate curve-fitting model for a given species because the “correct” model is more a product of an understanding of the behavior of birds, and whether and how their movement can be expected to modify the detection function, than it is a purely statistical curve-fitting exercise. We also note that the results are very sensitive to the way one defines distance intervals for the purpose of grouping data, which is a common practice. Our point is that the use of distance sampling computer programs is tricky at best, and subject to a good deal of error in the face of improper application.

**Conclusion**

We have no problem with the distance sampling method in theory. As long as assumptions are met, distance sampling is an elegant and potentially powerful method for adjusting counts to control for detection probability. Distance sampling may be an appropriate method for, say, desert tortoises (Anderson et al. 2001) because all detections are visual and can be precisely mapped, and because the animals are essentially stationary during a transect count. However, neither of those assumptions is true for birds, and violation of these assumptions may cause serious errors in density estimation using distance sampling techniques. Indeed, if any of the four distance sampling model assumptions listed above are violated (and at least three of them are probably always violated), or if the practical issues of sample size and model application are problematic, then any attempt to calculate a density estimate may not yield an accurate result.

Note that we have not ignored the detectability problem in our own work. We address the problem by using only those detections that come from within a fixed radius, and this probably catches most of the error that might be introduced because of a detectability bias. We have used simple frequencies of occurrence within fixed-radius sample points to examine the habitat distribution of numerous landbird species (Hutto 1998, Hutto and Young 1999). From these data we have concluded, for example, that the black-backed woodpecker (Picoides arcticus) is relatively abundant in, and nearly restricted in its distribution to, burned
forests; that the olive-sided flycatcher (*Contopus cooperi*) is relatively abundant in both heavily cut and burned forests; and that the brown creeper (*Certhia americana*) is relatively restricted to old-growth forest conditions in our region. We have also concluded from unadjusted fixed-radius data that the distribution of brown-headed cowbird (*Molothrus ater*) is more closely tied to landscape than to local-scale conditions (Young and Hutto 1999), and that partial-cut timber harvesting negatively affects (among others) brown creeper and Townsend’s warbler (*Dendroica townsendi*) and positively affects mountain chickadee (*Poecile gambeli*) and chipping sparrow (*Spizella passerina*) (Young and Hutto 2002). Our chosen method of dealing with detectability bias may be crude, but corroboration of our conclusions with those from other studies or syntheses of the same issues (Hejl et al. 1995, 2002; Kotliar et al. 2002) suggest that our conclusions are on the right track. We also question whether more involved adjustments associated with program DISTANCE make any difference when it comes to the consideration of unambiguous patterns about which land management decisions are generally made. Indeed, we have found that when the patterns of abundance are unambiguous, results are similar whether one adjusts data with distance sampling or not.

Theoretically, our data could be biased, but so could density estimates that emerge from distance sampling methods. Distance sampling methods do not prevent one from being deceived by detectability bias if important assumptions associated with its use are not met, and they are not. We would readily adjust our data using program DISTANCE if we generated the sample sizes needed to produce reliable detection functions and if we were convinced that the precise shape of a detection function is not influenced significantly by the movement of birds. Because neither condition is met for most species, and because the authors of such methods (Buckland et al. 2001) warn that it is inappropriate to use their models when critical assumptions are not met, we choose, for the time being at least, to sit in the “agnostic” camp—we are not yet convinced that the use of distance sampling is always going to be better than not doing so. We need good empirical data to resolve the issue more fully. Moreover, our concerns will not be alleviated through further improvements or refinements in the computer programs that allow one to force a better fit to (often) scanty data and then apply an (always) assumption-violated conversion of the data to generate a density estimate. We realize that there are other approaches that one might use to deal with detectability problems (e.g., double-observer sampling [Nichols et al. 2000], double sampling [Bart and Earnst 2002, Pollock et al. 2002], or removal modeling [Farnsworth et al. 2002]), but they too need more work and empirical validation before they become generally accepted as superior survey methods.

If tools like program DISTANCE become widely used and accepted without empirical verification, biologists may become more inclined to report density estimates (rather than probability of occurrence or number of detections) in the results sections of their papers, and may be less inclined to take the time to consider that the known failure of several key assumptions involved in the method of data conversion might have muddied rather than clarified actual trends in the data. Indeed, there is something alarming about the current trend toward use of a method that diverts our attention away from the data actually collected and toward results that emerge from a complex model that we know leans heavily on erroneous assumptions. The most insidious aspect of converting data to better reflect what one really would like to have measured is that the scientist may begin to believe he or she has a measure of actual density, and may lose intimacy with data by entering a black box (in much the same way one loses touch with his or her data through the use of complex multivariate statistics).

We wish to re-emphasize in closing that we collect point-count data the same way that everyone else does, whether they adjust the data with program DISTANCE later or not, so “differences of opinion on this issue should not hamper the ability to share data” (Hutto and Young 2002:9). Indeed, other authors (e.g., White and Bennetts 1996) have used our data to test their own methods of analysis. By recording detections the same way they are recorded for distance sampling, we also retain the option to analyze our data using program DISTANCE at some point in the future, should we become convinced that the method will improve our scientific understanding. Perhaps the best solution to the disagreement that surrounds this issue (until some reliable empirical data emerge) would be to encourage scientists to report both the unadjusted or fixed-radius data along with density estimates so that readers can better judge for themselves the veracity of conclusions drawn from such count data.
Acknowledgments. Although we may disagree with Ellingson and Lukacs about how to best handle the problem of potential bias associated with bird detectability, and although we may feel that our own use of indices is not as problematic as they suggest, we thank them for taking the time needed to write a carefully and clearly written response. We also thank Warren Ballard for giving us the opportunity to organize our own thoughts and attempt to rebut the issues they raised. We hope the exchange serves to better clarify thoughts on some recurring issues in wildlife biology, and serves to better the science that depends on accurate assessments of bird abundance. Thanks also to Jon Bart, Amy Cilimburg, Doug Johnson, and Skip Kowalski for checking to see that our response adequately and clearly addressed the issues raised by Ellingson and Lukacs.

Literature cited


Authors' address: Division of Biological Sciences and Wildlife Biology, University of Montana, Missoula, MT 59812, USA; e-mail for Hutto: hutto@selway.umt.edu.

Richard Hutto (left) is a professor of biology and wildlife biology at the University of Montana, where he teaches ornithology, Montana wildlife, and field ecology. Hutto serves as advisor and supervisor of the Northern Region Landbird Monitoring Program and has conducted research on habitat selection in landbirds in the northern Rocky Mountains during the breeding season, in the southwestern U.S. during the migratory seasons, and in western Mexico during winter. He received his M.S. from Northern Arizona University and Ph.D. from UCLA. Jock Young (right) is currently the coordinator of the Northern Region Landbird Monitoring Program. He received his B.S. from Oregon State University and M.S. degrees from the University of California at San Diego and the University of Montana. His interests include bird monitoring and conservation, and the modeling of habitat relationships.