Observer Error and Statistical Power: Evaluating Survey Reliability for Conservation Management

S.A. Field[a], **A.J. Tyre**[b], **S.C. Ball**[a] **and H.P. Possingham**[b]

[a]*Department of Applied and Molecular Ecology, University of Adelaide, Waite Campus, Glen Osmond SA 5064, Australia (scott.field@adelaide.edu.au)*

[b]*The Ecology Centre, Department of Zoology, University of Queensland, St Lucia Qld 4072, Australia*

Abstract: A major issue in ecological monitoring programs is the implementation of power analysis to evaluate the reliability of statistical analyses and to plan appropriate levels of survey effort. In particular, although the occurrence of false negative survey results (species present but not recorded) is more or less ubiquitous, it is seldom explicitly addressed in analyses. Using the example of a conservation manager monitoring bird species across a fragmented landscape, we run stochastic simulations that explore how statistical power to detect declines in these species varies as a function of 1) the allocation of survey effort (sites sampled relative to repeat visits to sites); 2) the initial prevalence of a species in the landscape; 3) the false negative error rate of observers conducting the surveys; and 4) the size of the decline. We construct artificial datasests in which declines occur and calculate power as the number of times a "virtual ecologist" successfully detects these declines. In sampling, the virtual ecologist is subject to observational constraints typically encountered in the field and uses a maximum likelihood method that estimates overall occupancy probability (p) across the landscape while accounting for observer error. We identify the key variables influencing power and consider how a manager may respond in attempting to developing an optimal monitoring design.

Keywords: Monitoring; Power analysis; Observer error; Biological survey; Maximum likelihood

1. INTRODUCTION

An effective survey and monitoring program should be a fundamental component of any wildlife conservation management strategy. However, any broadly applicable monitoring technique must overcome two key challenges: the high degree of variability inherent in large-scale systems; and the ubiquity of observer error in recording the presence and/or abundance of the species of interest.

The general solution to the first challenge is to increase sampling effort so that the statistical power of any significance tests applied to the data reaches an acceptable threshold. But how is statistical power affected by the inevitable occurrence of false negative survey results? These are errors that occur when a species is in fact present at a site but goes unrecorded. No matter how experienced or skilful the observer, there is always a finite chance of a species eluding detection. The issue of false negatives is therefore of critical importance in assessing the reliability of survey and monitoring programs, but is rarely dealt with in practice.

In this paper, we address the problem by conducting simulations that mimic the plight of a conservation manager faced with the task of monitoring declining birds across a fragmented landscape. Given a fixed annual budget, she must design a survey regime that maximizes her ability to detect any declines that have occurred, i.e. she is aiming to maximize *statistical power*. In our analysis we focus on four main factors that can influence power: 1) the allocation of survey effort, in terms of the number of sites sampled (n) relative to the number of repeat visits made to each site (m), subject to the constraint $nm \leq 500$; 2) the initial prevalence of a species in the landscape, measured by p, the proportion of patches occupied; 3) the error rate of observers conducting the surveys, specifically the false negative error rate, which is determined by the species-specific observability, q; and 4) how large a decline (d) must be before it is considered serious from a management point of view, i.e. the effect size we wish to be able to detect.

2. TWO COMMON PROBLEMS IN BIOLOGICAL MONITORING

We now formulate two key problems likely to be encountered by practitioners of biological surveys: 1) how to maximize statistical power when monitoring one or more species with unknown initial distribution(s); and 2) when a rare species is known to be limited to a certain number of sites, how to ensure that enough visits are made to those sites so that a substantial decline would be detected.

2.1 Problem 1: Monitoring an Assemblage

In the first case, we assume that the manager has a budget sufficient to conduct 500 surveys in a season. Each species can be surveyed simultaneously on a given survey visit, but with differing probabilities of detection (q). Her assessment of decline for each species is based on two rounds of surveying, an arbitrary length of time (e.g. ten years) apart, for each of which she estimates the overall occupancy ($\hat{p}$, proportion of patches occupied) across the landscape. Because the method for estimating $\hat{p}$ involves estimating two parameters ($\hat{p}$ and $\hat{q}$, see Section 3.1), the manager is constrained to conduct at least three repeat visits at each site.

The main variable under the manager's control is the allocation of survey effort between number of sites surveyed and number of repeat visits to those sites, i.e. the ratio n:m. The magnitude of a decline undergone by a given species obviously cannot be controlled, but power for a species can still be influenced indirectly by altering the threshold at which a decline is considered significant enough to trigger management actions. Species occupancy levels (p) and observability (q) are also out of the manager's control; all she can do is abstain from concluding anything about species that suffer from low power because p and/or q are low.

2.2 Problem 2: Monitoring a Threatened Species

In the second case, we assume that the manager is responsible for monitoring a threatened species that is known to be confined to only 15 sites. Given a species-typical observability (q) and a fixed threshold of a 50% decline for triggering management actions, she must calculate the number of repeat visits required to the 15 sites to ensure an adequate level of statistical power (e.g. the conventional level of 0.8) to detect declines is achieved. Here n, d and q are fixed and p is assumed to be 1, leaving her with only one variable, m, to vary.

3. PARAMETER ESTIMATION AND SIMULATIONS

The tools necessary to solve both of these problems are: 1) a method for estimating $\hat{p}$ when there is a chance of false negative survey results occurring; and 2) a simulation technique that mimics the process of data collection in biological surveys and thus allows statistical power to be calculated. These are described below.

3.1 Estimating $\hat{p}$ in the Presence of Observer Error

The method for estimating $\hat{p}$ in the presence of observer error is described in detail and its performance tested by Tyre et al. [in preparation] and is summarized briefly here. We assume that the result for any given survey (species present/not present) is the outcome of two binomial processes acting simultaneously: 1) the probability that the species is indeed present in the site (p); and 2) the probability that the species is observed in any given survey, given that it is present (q). Thus the survey results follow a 'finite mixture distribution' with a mixing probability (the probability a site is occupied), and two binomial components, one with a probability of success equal to zero. We use maximum likelihood methods to estimate the parameters [Hilborn and Mangel, 1997].

We assume that n sites have been visited m times each and that a species has a probability of occupying a site, p, that is constant for all sites throughout the landscape. A species also has a species-typical probability q of being observed during any one visit. After the m visits are complete, the number of observations of a species at a given site is o ($o \leq m$). If the species was observed at least once, then the likelihood of this observation is

$$L(o > 0) = p\binom{m}{0}q^{o}(1-q)^{m-o} \tag{1}$$

which is the binomial probability of o successes in m trials multiplied by the probability that the site was occupied. If the species was not observed at a site, the likelihood is:

$$L(o=0)=(1-p)+p(1-q)^m \qquad (2)$$

which is the probability that it was not there plus the probability that it was there but was not observed in m visits. Note that the second term in $L(o=0)$ is the probability of not seeing the species in m visits multiplied by the probability that it was there. The negative logarithms of these are summed over all sites, and this value minimised to find the maximum likelihood estimate for the two parameters, p and q. We implemented the code in C++ builder and used the Nelder-Mead simplex algorithm to find the best fit parameters.

3.2 Power Simulations: the "Virtual Ecologist"

We calculated statistical power (for problem 1) and number of required surveys (for problem 2) by employing a "virtual ecologist" " [Berger et al., 1999; Grimm et al., 1999; Tyre et al., in press] to sample artificial datasets in which we caused species' occupancy levels to decline by specified amounts.

When calculating power for problem 1, we varied the parameters $n{:}m$, p, q and d over ranges likely to be interesting to a manager (Table 1) and generated random datasets for each combination. The process was as follows. First, actual pre-decline patch occupancies for each of the n patches were generated as Bernouilli random variables with Pr(success) = p_1. The virtual ecologist recorded a given species as present on a single visit with probability p_1q. After m visits to n sites, the resulting dataset consisted of a vector of n binomially distributed random variables, $\tilde{o} \sim (m,p)$, representing the number of times the species was observed. Post-decline patch occupancies were then generated, again as Bernouilli random variables, but with Pr(success) = $p\text{-}pd$. We then estimated parameters for two models, one where Pr(success) = $\hat{p}_t$ in (1) and (2) is assumed to be constant across the two survey periods, and a second model where a separate $\hat{p}_t$ is fit for each period. We compared these models using a Likelihood Ratio test, and where the model with two separate $\hat{p}_t$ has a significantly lower likelihood (at $\alpha = 0.05$), we identify a negative trend.

For each of the 1362 parameter combinations listed in Table 1, we carried out 1000 repetitions of this process and recorded the number of repetitions in which $\hat{p}_1$ was significantly greater than $\hat{p}_2$.

Table 1. Parameter combinations used in power simulations.

Parameter	Values	Total
$n{:}m$	125:4, 100:5, ... 50:10	7
p_1	0.3, 0.4, ... 0.9.	7
d	0, 0.25, 0.5, 0.75	4
q	0.3, 0.4, ... 0.9	7
Total combinations		1362

This was a measure of statistical power, i.e. the probability of detecting a decline, given that a real decline had actually occurred.

When calculating the minimum number of surveys required for problem 2, we fixed n, (the number of sites that the threatened species is known to occupy from previous records) at 15, p_1 at 1 (assuming all sites are still currently occupied), q at 0.4 (representing a moderately observable species) and d at 0.5 (the management threshold for triggering recovery actions). We then ran simulations in which we varied m, the number of repeat visits, over the values three to ten inclusive and calculated the resultant statistical power.

4. RESULTS

4.1 Problem 1: Monitoring an assemblage

From the simulations designed to address the first problem, visual inspection indicated that statistical power increased monotonically with increasing levels of each of the parameters varied (Figure 1). Thus there appeared to be no optimal combination of parameters for which statistical power reached a distinct peak. A higher ratio of sites to visits ($n{:}m$), higher initial occupancies (p_1), and larger declines (d) all resulted in a greater chance that the manager was able to detect a decline (Figure 1). Power also increases with the observability q, but this effect is generally much smaller than all others, except when declines are small, and relatively few visits are made to each site. It is worth noting that the power of the highest sites:visits ratio is always higher than the lowest ratio, even at the lowest observability tested. When there is no decline, the method is making Type I errors (ie. detecting a trend when none is present) approximately 5% of the time.

4.2 Problem 2: Monitoring a threatened species

The simulations designed to address the second problem showed that statistical power rose sharply as a function of survey effort, and the conventionally acceptable level of 0.8 was attained

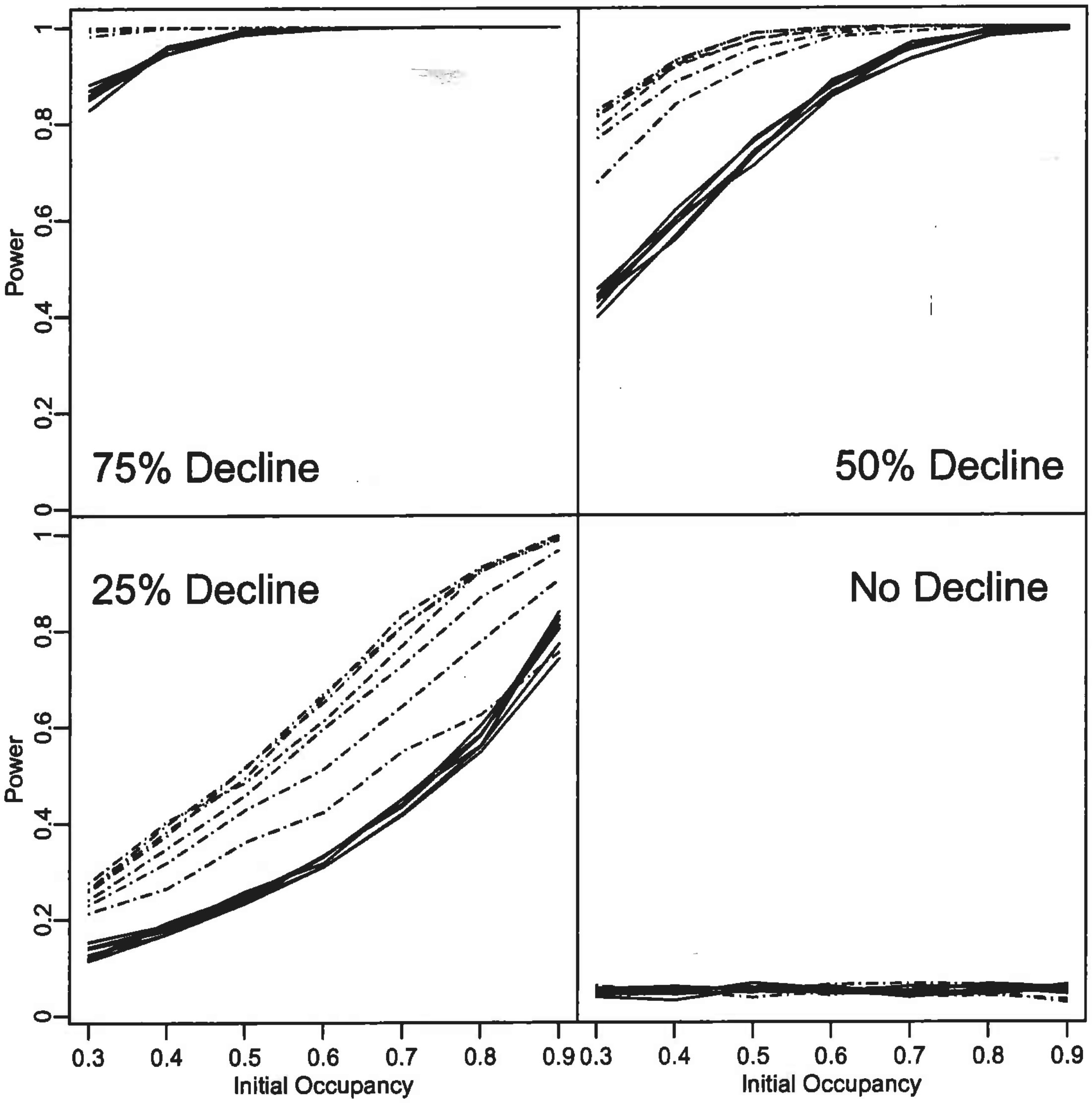

Figure 1. Power as a function of the % patches initially occupied (p_1) for each % decline (d) tested. For clarity, only the two most extreme sites to visits ratios (n:m) are shown: 125:4 (dot-dash lines), and 50:10 (solid lines). Each value of observability q is shown with a separate line; labels are omitted for clarity, but power generally increases with q.

after five repeat visits (Figure 2). From this point on, the gain in power from additional repeat visits reached a plateau. For a smaller decline of 25% even ten repeat visits only achieves a power of 0.7.

5. CONCLUSIONS

The main recommendation to arise from our analysis is that when total survey effort is fixed at 500 visits, the best way for a manager to increase statistical power is to allocate resources preferentially to obtaining more survey sites rather than revisiting the same sites many times. Power was consistently higher when the ratio of sites:visits was set at 125:4, for all combinations of the other parameters.

Predictably, the size of the decline and the proportion of sites in the landscape that were initially occupied also had strong effects on power. This means that small declines in species that are already rare will be very difficult for a manager to detect, even when many sites are visited. For such species, a manager may be forced to set a more generous decline threshold for triggering management actions. However, this strategy may result in actions being triggered only when the species has already declined to such low levels that recovery is difficult or impossible. The other option is to practice 'ecological triage' and

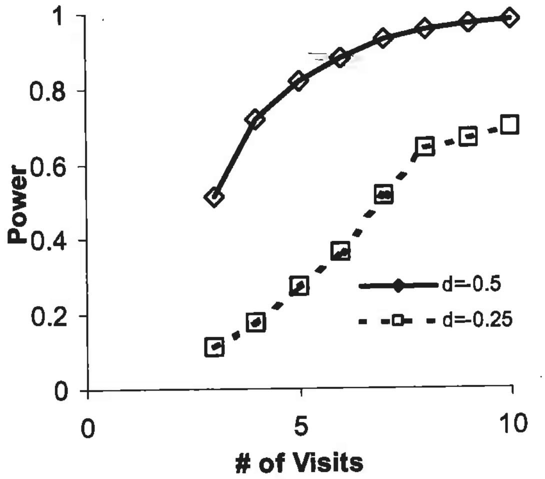

Figure 2. Statistical power as a function of the number of repeat visits to 15 sites that were initially occupied but suffered a 25% or 50% decline between surveys.

abandon surveying efforts for very rare species altogether, acknowledging that they are simply too difficult to monitor rigorously. On the other hand, one could argue that irretrievably low statistical power is evidence that a species is in need of immediate recovery action.

Our result that power was largely independent of species observability, q is an encouraging affirmation that Tyre et al.'s occupancy estimation method is functioning correctly. It shows that provided a species is well-represented in the landscape, a manager using this method should usually be able to detect substantial declines. However, some caution in accepting this interpretation may be warranted due to the fact that there are other realistic regions of parameter space yet to be explored. In particular, when q is very low, or when the overall survey budget is much more limited, the method may not be able to shield the user from the debilitating effect of low observability on statistical power. For example, the lowest number of repeat visits in our simulations was four. Even at the lowest observability we tested (0.3), there is still only a probability of $(0.7)^4 = 0.24$ of obtaining a false negative result at a given site. Reducing the number of visits to three would increase this probability to 0.34 and may lead to a substantial reduction in power.

5.1 Limitations and Future Applications

We can identify several other limitations of this study that suggest the need for further work. Firstly, we have not explicitly incorporated economic considerations into our analysis, the primary one being that acquiring new survey sites is likely to be substantially more expensive than performing repeat visits to existing ones. To allow for this, our constraint could change from $nm \leq B$ to $nm + cn \leq B$, where B is the total survey budget and c is the extra cost involved in setting up a new site. Inclusion of such a cost function might markedly alter the effect on power of the sites:visits ratio and may well produce an optimum rather than a monotonically increasing effect.

Secondly, in testing for declines, we have adhered to the traditional 5% risk of obtaining a false positive result. However, a conservation manager may be much more concerned about avoiding false negatives, as this kind of error may lead to the highly undesirable outcome of a species extinction. Managers may therefore be willing to accept an increased risk of false positives, if it means the ability to achiever greater power. Simulations trading off the risk of false positive and false negative errors might therefore be a worthwhile exercise.

Thirdly, we have only addressed the situation where the manager is making decisions based on two 'snapshots' of data. Although this is by no means unheard of, it would also be interesting to extend the method so it could be applied to time-series datasets collected over many years. This also raises the possibility of incorporating power simulations into an active adaptive monitoring framework, in which each years' updated estimates of occupancy are used to optimize the ensuing year's surveying, in terms of how much effort needs to be expended and where in the landscape it should be concentrated.

In summary, we have shown how estimates of patch occupancy across a landscape, once adjusted for observer errors, can be used as a management tool for deciding how to best allocate survey and monitoring efforts. Although we used declining birds as our example, the methods described could be equally well applied to a wide variety of biological systems. Thus we believe this analytic tool has great potential for use in solving common problems faced by conservation managers.

6. ACKNOWLEDGEMENTS

We would like to thank Max Possingham and David Paton for helpful discussions on bird survey techniques.

7. REFERENCES

Berger, U., G. Wagner and W.F. Wolff, Virtual biologists observe virtual grasshoppers: an assessment of different mobility parameters for the analysis of movement patterns, *Ecological Modelling* 115(2-3), 119-127, 1999.

Grimm, V., T. Wyszomirski, D. Aikman and J. Uchmanski, Individual-based modelling and ecological theory: synthesis of a workshop, *Ecological Modelling* 115(2-3), 275-282, 1999.

Hilborn, R. and M. Mangel, *The Ecological Detective*, Princeton University Press, pp., Princeton, New Jersey, 1997.

Tyre, A.J., H.P. Possingham and D.B. Lindenmayer, Matching observed pattern with ecological process: can territory occupancy provide information about life history parameters?, *Ecological Applications*, in press.

Tyre, A.J., B. Tenhumberg, S.A. Field and H.P. Possingham, Maximum likelihood estimates of false negative rates in biological survey data, in preparation.