Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2012, **3**, 188–194 doi: 10.1111/j.2041-210X.2011.00142.x

When can we ignore the problem of imperfect detection in comparative studies?

Frédéric Archaux[1]*, Pierre-Yves Henry[2] and Olivier Gimenez[3]

[1]*Cemagref, Domaine des Barres, F-45290 Nogent sur Vernisson, France;* [2]*UMR 7204 & UMR 7179 CNRS MNHN, Département Ecologie et Gestion de la Biodiversité, Muséum National d'Histoire Naturelle, 1 avenue du Petit Château, 91800 Brunoy, France; and* [3]*Centre d'Ecologie Fonctionnelle et Evolutive, UMR 5175, 1919 route de Mende, 34293 Montpellier Cedex 5, France*

Summary

1. Numbers of individuals or species are often recorded to test for variations in abundance or richness between treatments, habitat types, ecosystem management types, experimental treatments, time periods, etc. However, a difference in mean detectability among treatments is likely to lead to the erroneous conclusion that mean abundance differs among treatments. No guidelines exist to determine the maximum acceptable difference in detectability.

2. In this study, we simulated count data with imperfect detectability for two treatments with identical mean abundance (N) and number of plots (n_{plots}) but different mean detectability (p). We then estimated the risk of erroneously concluding that N differed between treatments because the difference in p was ignored. The magnitude of the risk depended on p, N and n_{plots}.

3. Our simulations showed that even small differences in p can dramatically increase this risk. A detectability difference as small as 4–8% can lead to a 50–90% risk of erroneously concluding that a significant difference in N exists among treatments with identical $N = 50$ and $n_{plots} = 50$. Yet, differences in p of this magnitude among treatments or along gradients are commonplace in ecological studies.

4. Fortunately, simple methods of accounting for imperfect detectability prove effective at removing detectability difference between treatments.

5. Considering the high sensitivity of statistical tests to detectability differences among treatments, we conclude that accounting for detectability by setting up a replicated design, applied to at least part of the design scheme and analysing data with appropriate statistical tools, is always worthwhile when comparing count data (abundance, richness).

Key-words: biodiversity monitoring, capture–mark–recapture, comparative studies, detection probability, nonparametric estimator, population size, sampling design, simulations, type I error

Introduction

Data on abundance and species richness of animals or plants are often collected in comparative studies to determine the variations in abundance and richness between treatments or habitats, areas that are managed differently, experimentally manipulated situations, monitoring over time, etc. Yet, the observed number of individuals or species is both a function of the true state of the system (i.e. the true number of individuals or species) and of the observation process. A significant proportion of the individuals or species are invariably missed during inventory observations. This has been demonstrated for a wide range of taxonomic groups, e.g. mammals (Baker 2004), birds (Boulinier *et al.* 1998), butterflies (Casula & Nichols 2003), spiders (Coddington, Young & Coyle 1996) and plants (Kéry *et al.* 2006). In this article, we will develop the case of abundance estimators, but our conclusions are equally relevant to species richness estimators as, from a methodological point of view, there is a parallel between species in a community and individuals in a population.

Some statistical approaches do account for differences in p; however, they usually require replicating counts in space or in time to be able to distinguish variations in the detection process

*Correspondence author. E-mail: frederic.archaux@cemagref.fr
Correspondence site: http://www.respond2articles.com/MEE/

from variations in abundance (but see Discussion for non-replicated methods). Replication greatly increases study costs in terms of manpower (time in the field), ecological disturbance (multiple sampling) and training requirements (statistical methodology). In practice, if manpower is limited, increasing the number of subsamples per plot is likely to necessitate reducing the number of plots proportionately. This would in turn reduce the statistical power of the study to detect differences among treatments. It is therefore tempting to ignore the detection difference issue with a view to maximising statistical power. This is probably one of the main reasons why imperfect detectability is still frequently ignored in species-monitoring schemes. For instance, 66% of 396 species-monitoring schemes in Europe do not control for detectability (EuMon 2006; consulted 18 October 2010).

A common belief among opponents of the systematic consideration of imperfect detectability (i.e. a detection probability of < 1) is that, when the degree of detectability is 'roughly' the same among the treatments, they should cancel out and tests for differences among treatments should not be affected. Therefore, any effort to account for imperfect detectability is a waste of manpower. However, Tyre *et al.* (2003) have already demonstrated that this simplification is unwarranted in the specific case of occurrence data analysis, where imperfect detectability can bias the estimator of the slope for the effect of covariates. The problem of detectability is generally acknowledged when the probability of detection varies with the factor of interest (e.g. experimental treatment, ecological or temporal gradient): differences in detectability may cause, accentuate or hide differences in the observed mean abundance between treatments and result in false interpretations. But, to our knowledge, no threshold values have been defined for the maximum difference in p that can be considered negligible among treatments and therefore can be neglected in the analyses (low risk of Type I error). In other words, when does the risk of drawing false conclusions become important enough to counterbalance the cost of replicating counts?

The detectability issue in comparative studies is not only a question of bias and precision in population size estimates (e.g. Burnham & Overton 1979; Chao 1987; Hellmann & Fowler 1999; Walther & Moore 2005; Xi, Watson & Yip 2008). Indeed, information about actual differences in N may be more reliable when detectability is low but varies little among treatments than it is when detectability is globally higher but varies more among treatments. Some specific case studies have emphasised the importance of accounting for detectability – i.e. accounting for p changed the conclusions of the study (Kéry, Gardner & Monnerat 2010a), while others have reached the opposite conclusion – i.e. accounting for p did not change the conclusions (Kéry & Schmid 2006; Bas *et al.* 2008). MacKenzie & Kendall (2002) discussed ways of incorporating detectability into abundance ratios (relative abundances) and recommended estimating detectability in all cases. Although they argued that there is generally no good reason to assume detectability to be constant (among treatments, along gradients), they did not provide a formal assessment of this statement.

The main goal of our study was therefore to estimate the minimum acceptable difference in detection probability that justifies estimating detectability when designing a comparative study and analysing data. We used simulations to explore the sensitivity of comparative tests of abundance data to among-treatment differences in mean individual detectability. We considered the additional influence of the mean number of individuals per plot and the number of plots. We then explored to what extent basic methods of accounting for detectability (using a nonparametric estimator) are effective at removing the detectability difference among treatments and limiting the type I error risk.

Materials and methods

Simulations mimicking real count data surveys were performed to compare two different treatments (e.g. two habitats, areas or experimental treatments) with different mean detectabilities ($p_1 \neq p_2$), but with identical mean population size (N), number of sampling plots (n_{plots}) and number of subsamples per plot (S). The simulations artificially created detection/non-detection histories, i.e. matrices indicating whether a particular individual was detected or not at a given subsampling occasion. The simulations were designed to estimate the risk of committing what is known in statistical terminology as a type I error, i.e. the null hypothesis H0, 'there is no difference in population size between the treatments', is rejected even though H0 is true ($N_1 = N_2 = N$ was fixed). The usual accepted risk is 5%, and a between-treatment difference is likely to inflate this risk.

Each run had four steps as described below (with i for individual, j for plot, t for treatment and s for subsample):

Step 1. Assign a population size N_{jt} to each of the n_{plots} of each treatment assuming a Poisson distribution: $N_{jt} \sim \text{Poisson}(N)$.

Step 2. Assign a detection probability p_{ijt} to each of the N_{jt} individuals present on each plot (jt) assuming a Beta distribution: $p_{ijt} \sim \text{Beta}(\text{beta1},\text{beta2}_t)$ (beta1 was identical for both treatments but not beta2; $p_t = \text{beta1}/(\text{beta1} + \text{beta2}_t)$), and determine whether the individual is detected on each of the S subsamples by performing S Bernoulli trials, $\text{Bernoulli}(p_{ijt})$.

Step 3. For each plot, count the number of individuals that were recorded in at least one subsample ($N.\text{raw}_{jt} \leq N_{jt}$) and calculate the corresponding Jackknife 2 estimate (Burnham & Overton 1979):

$$\hat{S}_{\text{Jack2}} = N.\text{raw} + n_1 \cdot \frac{2S-3}{3} - n_2 \cdot \frac{(S-2)^2}{S.(S-1)}$$

where n_1 is the number of individuals detected in only one subsample (singletons), and n_2 is the number of individuals detected in only two of the S subsamples (doubletons). We also calculated the corresponding Chao 2 estimate (see results in Appendix 2).

Step 4. Test whether the mean $N.\text{raw}_{jt}$ (and $\hat{S}_{\text{Jack2jt}}$) values statistically differ between the two treatments, assuming a Poisson (and respectively Gaussian) distribution; see below for the justification of using these distributions. Increment a counting variable if the 'treatment' P-value is significant (i.e. $< 5\%$).

For each combination of p_1, p_2, N and n_{plots}, steps one to four were repeated 5000 times. The proportion of runs with a P-value $< 0{\cdot}05$ (α_{sim}) was then used as an estimate of the type I error risk. As we randomly generated two sets of populations for the two treatments in step 1, setting the same p, N and n_{plots} values for the two treatments, by chance, significant between-treatment difference in mean N occurred in $c.$ 5% of the simulations. That is, estimated rejection rates are perfectly acceptable unless they are $> 5\%$.

To determine the maximum acceptable difference in p between the two treatments ($\Delta p_{max} = p_2 - p_1$) for which the type I error risk was still below the desired 5%, the value of p_2 was increased (from p_1) until the type I error risk was above the desired 5% threshold value (actually 5·6%, i.e. 0·05 + 1·96*√(0·05*0·95/5000)). This involved performing sets of 5000 runs without knowing a priori the number of sets needed to reach the p_{2max} value with the desired precision. To speed up the process, we incremented p_2 in steps of 0·1 among sets of runs, then in steps of 0·01 and finally in steps of 0·001, rather than incrementing p_2 directly in steps of 0·001.

For simplicity, we assumed that there was no variation in p_{ijt} among the S subsamples. Initially, we also explored the effect of the level of heterogeneity (var(p)) on the type I error risk by using combinations of beta1$_t$ and beta2$_t$ values and keeping the same mean detectability p_t. However, as the results varied only slightly with var(p) (at least in comparison with p), only the results for beta1 = 2 are presented herein.

We used a generalised linear model with a Poisson distribution for N.raw but a linear model with a Gaussian distribution for $\hat{S}_{Jack2}$ in step 4 because simulations with $N_1 = N_2$ and $p_1 = p_2$ (to check that the nominal rejection rate (α_{nom}) of the tests was effectively 5%) showed that α_{nom} was effectively 5% when assuming a Poisson distribution for N.raw but was generally above 5% for $\hat{S}_{Jack2}$. With a Gaussian distribution for $\hat{S}_{Jack2}$, the nominal rejection rate was 5%.

Obviously, we could not explore all possible values for all parameter combinations. Hence, we used values that are commonly found in biostatistics and ecology: N = 10, 50 or 100; n_{plots} = 10, 25, 50, 75 or 100; S = 1 (for raw counts), 3–10; p_1 = 0·1, 0·25, 0·5 and 0·8 (with beta1 = 2). We developed specific functions for the *R* software (R Development Core Team 2009), provided in Appendix 1, which will allow the reader to estimate the type I error risk in other situations (note that the functions can be easily modified as to draw a single set of populations for the two treatments, mimicking the case of a single set of plots visited by two different observers or sampled using two competing methods).

Results

RAW COUNTS OF INDIVIDUALS

As expected, the type I error risk α_{sim} gradually increased to above the 5% limit as the difference in p among treatments (Δp) increased. The increase was stronger for lower values of p.

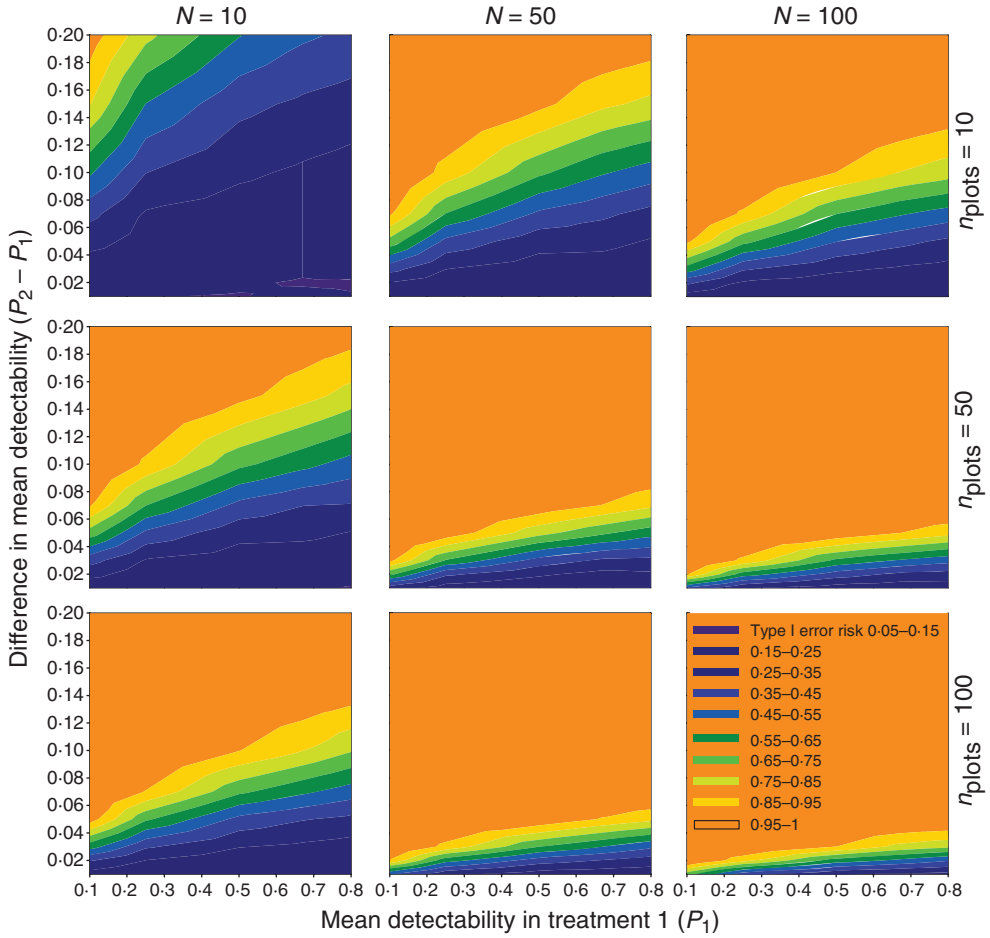

Fig. 1. Risk of concluding that the mean raw count of individuals/species per plot (N) is significantly different between the two treatments while actually it is not (type I error risk), depending on: (i) N (N = 10, 50 or 100; columns 1–3), (ii) the number of plots per treatment (n_{plots} = 10, 50 or 100 with n_{plots} being the same for the two treatments; rows 1–3), (iii) the mean detectability in treatment 1 (p_1; x-axis) and (iv) the difference in mean detectability between the two treatments ($p_2 - p_1$ with $p_2 > p_1$; y-axis). The hotter the colour, the greater the risk of committing a type I error.

Logically, α_{sim} also increased with both N and n_{plots}, as the two variables play symmetrical roles (Fig. 1).

More importantly, α_{sim} was far above the 5% threshold value, even for low Δp, in the majority of the situations considered. At the extreme, when $N = 100$, $n_{plots} = 100$ and $\Delta p \geq 4\%$, α was $> 0{\cdot}9$, for all values of p. The situations with a reasonably low ($< 0{\cdot}1$) risk of committing a type I error had very low statistical power (and therefore a high risk of committing a type II error), e.g. with $N = 10$, $n_{plots} = 10$, $p_1 = 0{\cdot}1$ and $\Delta p \leq 3\%$.

PERFORMANCE OF THE JACKKNIFE 2 ESTIMATOR

As expected, the factor that impacted the performance of the Jackknife 2 estimator the most was p: the higher the p value, the larger the Δp_{max} (shown by colder colours in Fig. 2). However, no simple monotonic relationship related Δp_{max} to the number of subsamples s. Indeed, when $p_1 = 0{\cdot}25$, we found an optimum number of subsamples below and above which Δp_{max} decreased ($s = 5$ or 6 for all N and n_{plots}; Fig. 2). For other combinations, the higher (at least 10 replicates per plot) or the

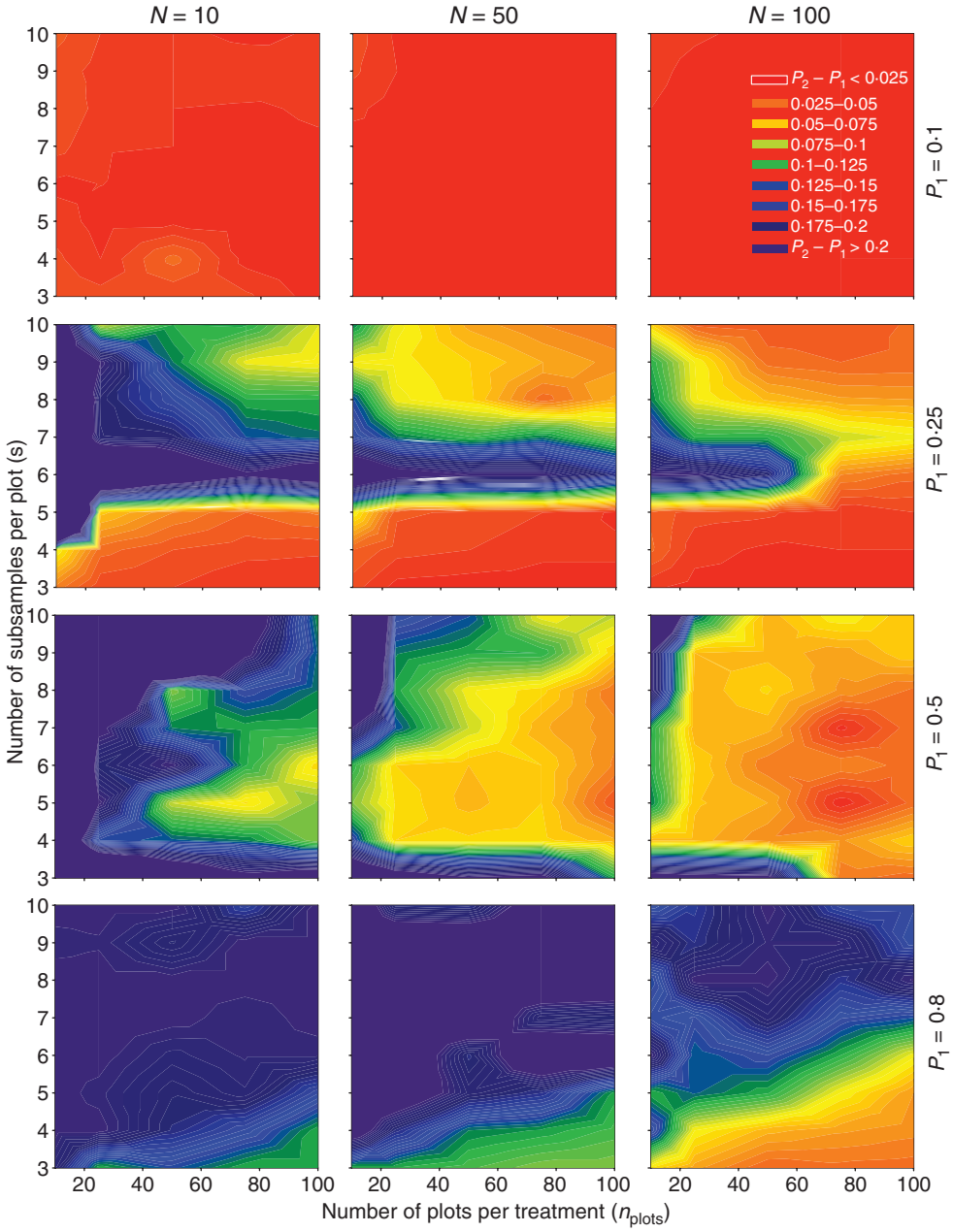

Fig. 2. Maximum acceptable difference in mean detectability between the two treatments (p_2–p_1) that ensures an acceptable nominal rejection rate ($\alpha_{sim} \leq 0{\cdot}056$) when a Jack2 estimator is used instead of raw data, depending on : (i) the mean number of individuals/species per plot ($N = 10$, 50 or 100; columns 1–3), (ii) the mean detectability in treatment 1 ($p_1 = 0{\cdot}1$, 0·25, 0·5 or 0·8; lines 1–3), (iii) the number of plots per treatment (n_{plots}; x-axis) and (iv) the number of replicates per plot (s; y-axis). In all simulations, $n_{runs} = 5000$ and $beta1_1 = beta1_2 = 2$. The hotter the colour, the smaller the differences in detectability between treatments that can be accounted for by the Jackknife estimator.

lower the *s* value, the higher the $\Delta p_{max.}$ In the remaining cases, s seemed to have little influence on Δp_{max}.

More quantitatively, when $p_1 = 0{\cdot}1$, the Jackknife failed to keep $\alpha_{sim} < 5\%$ when the difference in mean *p* between treatments was ≥2%, for all combinations of *N*, *S* and n_{plots} tested. When $p_1 = 0{\cdot}25$, the Jackknife performed better as it maintained $\alpha_{max} < 5\%$ for Δp_{max} values ranging from 5 to 40% (the % decreasing with increasing *N* and n_{plots}). Interestingly, Δp_{max} peaked when *s* equalled six replicates per plot. A more surprising pattern was observed when $p_1 = 0{\cdot}5$. Indeed, higher values for Δp_{max} were found when *s* was either low (3–4) or high (8–9) than when s was intermediate (especially at low n_{plots}, such as 20). When $p_1 = 0{\cdot}8$, the Jackknife estimator was able to preserve $\alpha_{sim} < 5\%$ for a Δp value of around 10% (i.e. p_2 around 0·9) when s was three for $N = 10$, 3–4 for $N = 50$ and 4–6 for $N = 100$ (depending on n_{plots}).

Discussion

DIFFERENCE IN *P* AMONG TREATMENTS AND TYPE I ERROR RISK

This study provides the first formal evidence supporting the often-repeated recommendation to systematically use statistical tools to account for detection probability when comparing count data among treatments. Our main result is that the risk of committing type I errors because of differences in mean detectability among treatments is high even when the difference is barely perceptible in the field. For instance, for two treatments with the same number of plots ($n_{plots} = 50$) and the same mean number of individuals/species per plot ($N = 50$), there is a 50–90% risk of erroneously declaring that the two treatments differ in their mean number of individuals/species per plot, when the mean probability of detection differs by only 4–8% among treatments (depending on the mean probability of detection, *p*). Table 1 provides a representative overview of the variability in *p* reported in ecological studies, including a wide variety of taxa and factors impacting *p* (e.g. species identity, observer, etc.). The range in detection probability (among species, observers, etc.) is almost invariably above 10%. We suspect that in many cases, if not most, assuming that the variation in *p* among treatments or over time is negligible could lead scientists or managers to draw misleading conclusions. We therefore agree with MacKenzie & Kendall (2002) who stated that '*a priori*, it is more likely that detection probabilities are actually different; hence, the burden of proof should be shifted, requiring evidence that detection probabilities are practically equivalent'. Small differences in detectability among treatments are likely to be commonplace in ecological studies. For example, the set of species may differ among treatments, different observers may participate in the surveys, individuals may be counted at different distances among treatments (e.g. because of variations in habitat proximity) or weather or seasonal conditions may not be the same.

Because the type I error risk decreases as the mean detection probability *p* increases, one way to reduce this risk is to increase *p*. This can be achieved simply by visiting the same plot a number of times or by installing several subplots (i.e. replicating counts) and calculating the total number of different species/individuals (or if individuals cannot be identified from one visit to the next, the highest number of individuals counted in one visit). Nonetheless, even though this strategy may help to limit type I errors, the risk still remains above the 5% threshold when raw counts are used. Therefore, as soon as counts are replicated, one should always use estimators of *N* accounting for *p*, rather than raw counts, because estimators always keep the type I error risk closer to the desired 5%.

ESTIMATORS OF *N* ACCOUNTING FOR *P*

The reliability of abundance/richness estimates (high precision, low bias) increases with the fraction of individuals or species that is recorded, i.e. with the mean detectability and the level of replication (Xi, Watson & Yip 2008). Our simulations point out that for very low detection probabilities (i.e. around 0·1), none of the strategies we explored (up to 10 subsamples) kept the rejection rate below 5% when the difference in mean detectability among treatments was >2%. In those cases, our simulations showed that the Chao2 estimator (Chao 1987) performed only marginally better than the Jack2 estimator (results

Table 1. Examples of factors impacting the probability of detection *p* in a variety of taxonomic groups (with mean value *p* and range of values)

Taxon	Cue	Factor of variation in *p*	Scale	*p* (%)	Range % (*p*)	Reference
Plants	Sightings	4 morphological types	Species	72·7	50·5–86·1	Archaux *et al.* (2009)
		4 cover classes		87	42·6–100	
		11 observers		81	67–90	
Butterflies	Sightings	150 species	Species	50	17–81	Kéry & Plattner (2007)
		40 observers		61	37–83	
		3 dates		50·3	42–65	
Anurans	Sightings/Songs	3 dates	Species	19·1	3·4–39·9	Royle (2004a)
Birds	Songs	128 species	Species	64	3–99	Kéry & Schmid (2008)
		8 observers		85·9	81–93	Nichols *et al.* (2000)
Deer	Sightings	Group size (1–3 ind)	Individuals	81·7	70–92	Cook & Jacobson (1979)
		2 observers		58·7	56·3–61	
Moose	Sightings	3 snow condition classes	Individuals	57	40–70	LeResche & Rausch (1974)
Dolphins	Sightings	Distance to vessel (1–5 nm)	Individuals	*c.* 80	*c.* 60–100	Marques & Buckland (2003)

shown in Appendix 2). Nonetheless, for greater detection probabilities (> 0·1), the risk of committing a type I error as a result of detectability difference was significantly reduced, sometimes to the 5% threshold, by replicating counts and using nonparametric estimators (Fig. 2). The Jackknife estimator was able to preserve a nominal rejection rate of 5% despite an among-treatment difference in p of 10%, with only three replicates per plots when $p \geq 0{\cdot}5$ and $N \leq 50$. A higher level of replication would be necessary for higher N values, especially for large sample sizes (large n_{plots}). Schmeller *et al.* (2009; see their Table 2) provided data on monitoring practices (mainly volunteer based) in five European countries: the median level of subsampling among 262 monitoring schemes was between 1 and 3·5, with the number of replicates being either the number of visits each year or the number of samples per visit. Our results show that such a limited level of subsampling may not be sufficient to adequately account for potential variations in detectability.

We focused on nonparametric estimators because they give sound results in many circumstances (e.g. Otis *et al.* 1978; Walther & Moore 2005) and are widely used; though, alternative, more flexible methods are now available to assess the size of a closed population or community. These methods are mostly based on generalised linear models and can incorporate a rich variety of factors possibly affecting N and p (see Royle & Dorazio 2008; King *et al.* 2009). It would be interesting to assess the minimum replication needed with these methods to ensure acceptable type 1 error risks as we have done for nonparametric estimators. We therefore suggest extending simulations to these recent approaches, including patch occupancy (MacKenzie *et al.* 2002; MacKenzie & Royle 2005), finite-mixture models (Pledger 2000, 2005) and N-mixture models (Royle 2004b). Finally, methods such as distance sampling (Buckland *et al.* 1993; Nichols *et al.* 2000) and spatial capture–recapture methods (Efford & Dawson 2009) that account for detectability without requiring replicating counts should also be considered.

Unless unreplicated methods can be implemented, a minimum requirement of doubling or tripling the number of subsamples per plot may be incompatible with the resources (manpower, funds) that can be devoted to many survey programmes. Furthermore, fieldworkers who have to repeat visits might become demotivated, thus potentially lowering the amount and quality of data (particularly when volunteers are concerned). If replication is traded off against the number of plots, it would also dangerously lower the ability to detect differences among treatments (Type II error). It is nevertheless crucial that biodiversity data be collected in ways that allow reliable inferences to be made (Yoccoz, Nichols & Boulinier 2001). MacKenzie & Kendall (2002) discussed various ways of incorporating detectability into abundance estimates (equivalence testing, model averaging). A double-sampling approach may be an interesting balance between limiting the cost of a study and improving its robustness. This approach relies on estimating the mean detectability and its standard deviation by replicating counts over a representative part of the study plots only (Bart & Earnst 2002; Pollock *et al.* 2002) and has been successfully used with patch occupancy models (Kéry *et al.* 2010b).

Conclusions

Statistical tests comparing mean plot population size or mean species richness between treatments are very sensitive to even small differences in mean probability of detection among treatments. As numerous factors are likely to significantly affect detectability in most biodiversity surveys, it is more reasonable to assume *a priori* that differences in detectability among treatments could bias the statistical comparison tests. Consequently, in line with MacKenzie & Kendall (2002), we recommend that scientists and managers always choose a robust sampling design that estimates and incorporates detection probability in the statistical analyses, before the routine sampling starts.

Acknowledgements

This work originated from a workshop held in 2005 at Obergurgl (Austria) jointly organised by the 6th European Funding Program Network of Excellence ALTER-Net (http://www.alter-net.info/) and the EuMon STREP EU-project (http://eumon.ckff.si; EU-Commission contract number 6463). We thank Frédéric Gosselin, Marc Kéry and one anonymous referee for constructive comments on earlier versions of the manuscript and Victoria Moore for English corrections.

References

Archaux, F., Camaret, S., Dupouey, J.-L., Ulrich, E., Corcket, E., Bourjot, L., Brêthes, A., Chevalier, R., Dobremez, J.-F., Dumas, Y., Dumé, G., Forêt, M., Forgeard, F., Lebret Gallet, M., Picard, J.-F., Richard, F., Savoie, J., Seytre, L., Timbal, J. & Touffet, J. (2009) Can we reliably estimate species richness with large plots? An assessment through calibration training. *Plant Ecology*, **203**, 303–315.

Baker, J.D. (2004) Evaluation of closed capture-recapture methods to estimate abundance of Hawaiian monk seals. *Ecological Applications*, **14**, 987–998.

Bart, J. & Earnst, S. (2002) Double sampling to estimate density and population trends in birds. *Auk*, **119**, 36–45.

Bas, Y., Devictor, V., Moussus, J.-P. & Jiguet, F. (2008) Accounting for weather and time-of-day parameters when analysing count data from monitoring programs. *Biodiversity and Conservation*, **17**, 3403–3416.

Boulinier, T., Nichols, J.D., Sauer, J.R., Hines, J.E. & Pollock, K.H. (1998) Estimating species richness: the importance of heterogeneity in species detectability. *Ecology*, **79**, 1018–1028.

Buckland, S.T., Anderson, D.R., Burnham, K.P. & Laake, J.L. (1993) *Distance Sampling: Estimating Abundance of Biological Populations*. Chapman, London.

Burnham, K.P. & Overton, W.S. (1979) Robust estimation of population size when capture probabilities vary among animals. *Ecology*, **60**, 927–936.

Casula, P. & Nichols, J.D. (2003) Temporal variability of local abundance, sex ratio and activity in the Sardinian chalk hill blue butterfly. *Oecologia*, **136**, 374–382.

Chao, A. (1987) Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, **43**, 783–791.

Coddington, J.A., Young, L.H. & Coyle, F.A. (1996) Estimating spider species richness in a southern Appalachian cove hardwood forest. *Journal of Arachnology*, **24**, 111–128.

Cook, R.D. & Jacobson, J.O. (1979) A design for estimating visibility bias in aerial surveys. *Biometrics*, **35**, 735–742.

Efford, M.G. & Dawson, D.K. (2009) Effect of distance-related heterogeneity on population size estimates from point counts. *Auk*, **126**, 100–111.

EuMon (2006) DaEuMon: a database of animals, plants and habitats monitoring schemes in Europe. URL http://eumon.ckff.si/about_daeumon.php [Accessed 18 October 2010].

Hellmann, J.J. & Fowler, G.W. (1999) Bias, precision, and accuracy of four measures of species richness. *Ecological Applications*, **9**, 824–834.
Kéry, M., Gardner, B. & Monnerat, C. (2010a) Predicting species distributions from checklist data using site-occupancy models. *Journal of Biogeography*, **37**, 1851–1862.
Kéry, M. & Plattner, M. (2007) Species richness estimation and determinants of species detectability in butterfly monitoring programmes. *Ecological Entomology*, **32**, 53–61.
Kéry, M. & Schmid, B. (2006) Estimating species richness: calibrating a large avian monitoring programme. *Journal of Applied Ecology*, **43**, 101–110.
Kéry, M. & Schmid, H. (2008) Imperfect detection and its consequences for monitoring for conservation. *Community Ecology*, **9**, 207–216.
Kéry, M., Spillmann, J.H., Truong, C. & Holderegger, R. (2006) How biased are estimates of extinction probability in revisitation studies? *Journal of Ecology*, **94**, 980–986.
Kéry, M., Royle, J.A., Schmid, H., Schaub, M., Volet, B., Häfliger, G. & Zbinden, N. (2010b) Site-occupancy distribution modeling to correct population-trend estimates derived from opportunistic observations. *Conservation Biology*, **24**, 1388–1397.
King, R., Morgan, B.J.T., Gimenez, O. & Brooks, S.P. (2009) *Bayesian Analysis for Population Ecology*. CRC Interdisciplinary Statistics Series, Chapman & Hall.
LeResche, R.E. & Rausch, R.A. (1974) Accuracy and precision of aerial moose censusing. *Journal of Wildlife Management*, **38**, 175–182.
MacKenzie, D.I. & Kendall, W.L. (2002) How should detection probability be incorporated into estimates of relative abundance? *Ecology*, **83**, 2387–2393.
MacKenzie, D.I. & Royle, J.A. (2005) Designing occupancy studies: general advice and allocating survey effort. *Journal of Applied Ecology*, **42**, 1105–1114.
MacKenzie, D.I., Nichols, J.D., Lachman, G.B., Droege, S., Royle, J.A. & Langtimm, C.A. (2002) Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, **83**, 2248–2255.
Marques, F.F.C. & Buckland, S.T. (2003) Incorporating covariates into standard line transect analyses. *Biometrics*, **59**, 924–935.
Nichols, J.D., Hines, J.E., Sauer, J.R., Fallon, F.W., Fallon, J.E. & Heglund, P.J. (2000) A double-observer approach for estimating detection probability and abundance from point counts. *Auk*, **117**, 393–408.
Otis, D.L., Burnham, K.P., White, G.C. & Anderson, D.R. (1978) Statistical inference from capture data on closed animal populations. *Wildlife Monographs*, **62**, 1–135.
Pledger, S. (2000) Unified maximum likelihood estimates for closed capture-recapture models using mixtures. *Biometrics*, **56**, 434–442.
Pledger, S. (2005) The performance of mixture models in heterogeneous closed population capture–recapture. *Biometrics*, **61**, 868–876.
Pollock, K.H., Nichols, J.D., Simons, T.R., Farnsworth, G.L., Bailey, L.L. & Sauer, J.R. (2002) Large scale wildlife monitoring studies: statistical methods for design and analysis. *Environmetrics*, **113**, 105–119.
R Development Core Team (2009) R: a language and environment for statistical computing. R foundation for statistical computing, Vienna. URL http://www.R-project.org [accessed 16 September 2009].
Royle, J.A. (2004a) Modeling abundance index data from Anuran Calling Surveys. *Conservation Biology*, **18**, 1378–1385.
Royle, J.A. (2004b) N-Mixture models for estimating population size from spatially replicated counts. *Biometrics*, **60**, 108–115.
Royle, J.A. & Dorazio, R.M. (2008) *Hierarchical Modeling and Inference in Ecology: The Analysis of Data from Populations, Metapopulations, and Communities*. Academic Press, San Diego, CA, 444 pp.
Schmeller, D.S., Henry, P.-Y., Julliard, R., Gruber, B., Clobert, J., Dziock, F., Lengyel, S., Nowicki, P., Déri, E., Budrys, E., Kull, T., Tali, K., Bauch, B., Settele, J., van Swaay, C.A.M., Kobler, A., Babij, V., Papastergiadou, E. & Henle, K. (2009) Advantages of volunteer-based biodiversity monitoring in Europe. *Conservation Biology*, **23**, 307–316.
Tyre, A.J., Tenhumberg, B., Field, S.A., Niejalke, D., Parris, K. & Possingham, H.P. (2003) Improving precision and reducing bias in biological surveys: estimating false-negative error rates. *Ecological Applications*, **13**, 1790–1801.
Walther, B.A. & Moore, J.L. (2005) The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. *Ecography*, **28**, 815–829.
Xi, L., Watson, R. & Yip, P.S.F. (2008) The minimum capture proportion for reliable estimation in capture-recapture models. *Biometrics*, **64**, 242–249.
Yoccoz, N.G., Nichols, J.D. & Boulinier, T. (2001) Monitoring of biological diversity in space and time. *Trends in Ecology and Evolution*, **16**, 446–453.

Received 29 October 2010; accepted 10 June 2011
Handling Editor: Robert Freckleton

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Appendix S1 R functions "TypeIerror.2treatments" and "Max.deltap.2treatments.Jack2".

Appendix S2 Maximum acceptable difference in mean detectability between the two treatments ($p2$-$p1$) ensuring nominal rejection rate ($\alpha_{sim} = 0{\cdot}056$) when Chao2 estimator is used instead of raw data.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials may be reorganized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.