When and how to integrate probability and nonprobability samples to monitor biodiversity

[1,2]Robin J. Boyd, [1†]Oliver L. Pescott, [1†]Simon Rolph, [1]Diana E. Bowler, [1]Nick J. B. Isaac, [1]Francesca Mancini, [1,2]David B. Roy, [1]Gesa von Hirschheydt, [1]Michael J. O. Pocock

[†]these authors contributed equally

[1]UK Centre for Ecology and Hydrology, Maclean Building, Benson Ln, Crowmarsh Gifford, OX10 8BB

[2]Centre for Ecology and Conservation, University of Exeter, Falmouth, Cornwall, UK

Corresponding author email: r.j.boyd@exeter.ac.uk

Abstract

1. Species occurrence and abundance data are generated by probability or nonprobability sampling. Under probability sampling, sites are selected according to a known design (e.g. stratified random sampling). Under nonprobability sampling—for example, the collection of data by citizen scientists at sites of their own choosing—they are not. Whereas probability samples tend to be relatively small but representative of the wider landscape, nonprobability samples are often much larger but less representative.

2. We evaluate when and how probability and nonprobability samples can be integrated to estimate temporal change in species occupancy. Using computer simulation, we compare four estimators that are applied to a small probability sample, a larger nonprobability sample, or both.

3. Scenarios vary in terms of baseline occupancy and how it changes over time, the size of the nonprobability sample, and the dependence of site inclusion in the nonprobability sample on occupancy. To isolate the effects of sampling design, we assume occupancy is measured without error at sampled sites.

4. Estimators based solely on nonprobability samples were generally biased and did not achieve nominal confidence interval coverage. Model-based integration improved precision but suffered from bias unless the dependence of site inclusion in the nonprobability sample on occupancy was stable over time—a condition that is difficult to verify in practice.

5. A design-based integrated estimator, which has not previously been considered in ecology, consistently performed well. It was approximately unbiased and achieved nominal confidence interval coverage in all scenarios, and it had low mean squared error in most scenarios.

6. The desirable statistical properties of the design-based integrated estimator arise from the sampling design rather than the ecological process and are therefore expected to apply regardless of study variable (e.g. occupancy, abundance, species richness). Extensions that account for measurement error are discussed but not tested.

7. Design-based integration provides a more reliable basis for combining probability and nonprobability samples than the model-based estimators considered here, and we recommend it as the default approach. More elaborate model-based estimators may perform better in some settings, but their performance depends on assumptions that are difficult to verify in practice.

Key words: citizen science; integrated distribution model; joint likelihood; species abundance; data defect correlation

Introduction

Action is needed to reverse losses in species abundance and reduce extinction risk. This sentiment is reflected in Goal A of the Kunming-Montreal Global Biodiversity Framework, which was endorsed by 196 countries (Convention on Biological Diversity, n.d.). Whether conservation action has been successful is ultimately an empirical question that can only be addressed by monitoring species' abundances and geographic distributions.

Monitoring species' abundances and distributions is fundamentally a sampling problem. The relevant geographic area (e.g. a country or administrative unit) can be conceived as a set of sites at which data could be collected. A target parameter is then defined that summarises a study variable and how it has changed over time. Examples include mean occupancy and abundance across sites in one year relative to another. Data are collected periodically at a sample of sites, and a sample-based estimator is used to transform those data into an estimate of the parameter's value.

Estimators come in many shapes and sizes. A simple option is to take the same mathematical operations that define the target parameter and to apply them to the sample. For example, we could use mean occupancy across *sampled sites* in one year relative to another as an estimator of its population analogue. As we explain below, however, more sophisticated estimators that incorporate statistical models and/or information about the sampling process are often preferable in terms of bias and precision (Boyd et al., 2024; Fink et al., 2023; Van Swaay et al., 2002).

The bias and variance of an estimator depend on how sites were selected for inclusion in the sample (i.e. the sampling design). Many standard estimators are at least approximately unbiased[1] when sample inclusion does not depend on the study variable (Rubin, 1976; Schuessler & Selb, 2025), and they are generally more precise when the number of sites included in the sample is larger. Unfortunately, for reasons that we explain below, it is difficult to achieve both things simultaneously.

One way to ensure that sample inclusion does not depend on the study variable is probability sampling. Under probability sampling, sites are selected for inclusion using a randomisation mechanism that depends on known design features (e.g. habitat or geography) rather than the study variable (Lohr, 2022 p. 31; Pescott et al., 2019).[2] Conditional on these features, sample inclusion and the study variable are independent (Sugden & Smith, 1984).

The problem with probability sampling is that it can be expensive. Some randomly selected sites will inevitably fall in inaccessible or unattractive locations, and people must be paid or otherwise incentivised to sample them. For example, the UK Pollinator Monitoring Scheme pays professionals to collect data at sites that volunteers are not willing visit (UK Pollinator Monitoring Scheme, 2026). As the number of randomly selected sites grows, so does the cost. Under a fixed budget, this places an upper bound on achievable sample size and hence precision.

[1] Unless stated otherwise, we use the term "bias" to refer to the bias caused by site selection. Measurement error (e.g. imperfect detection) at sampled sites is a separate source of bias. We return to the implications of measurement error in the *Discussion*.

[2] Under simple random sampling, sample inclusion does not depend on anything (i.e. there are no design features).

We refer to anything that is not probability sampling as nonprobability sampling (Boyd, Powney, et al., 2023). This includes convenience sampling by citizen scientists at sites of their own choosing (e.g. Sullivan et al., 2014) and nominal probability sampling when an appreciable number of sites do not get sampled despite having been randomly selected for inclusion (e.g. Brereton et al., 2011; Pescott et al., 2019). Some distinguish between convenience sampling and nominal probability sampling with appreciable dropout. We find it simpler to put them in the same category, since neither produces a probability sample (see Lohr, 2022, pp. 539-542).

From a statistical perspective, nonprobability sampling is essentially the flip side of probability sampling. Giving participants the freedom to choose where to collect data often increases participation and hence the number of sites that end up in the sample (Broughton & Pocock, 2022). However, it also means accepting that sample inclusion may depend on the study variable. Hence, we might expect that an estimator applied to a nonprobability sample would be more precise but also more biased than if it were applied to a probability sample.

Since there are pros and cons to probability and nonprobability samples, it is not surprising that there has been interest in constructing estimators that integrate the two (Wiśniowski et al., 2020; see below for more on data integration in ecology). The hope is that the integrated estimator will strike a better balance between bias and precision than an estimator based solely on one sample. As will become clear in later sections, there are even circumstances in which it is possible to integrate the nonprobability sample without introducing any bias whatsoever (Rao, 2021).

In ecology, data integration has often been framed in terms of "structured" and "unstructured" data (Farr et al., 2021; Isaac et al., 2020; Miller et al., 2019; Pacifici et al., 2017; Zipkin et al., 2017). The degree of "structure" in a given dataset captures not only the sampling design, but also the degree to which data were sampled according to a consistent protocol and what is reported (e.g. presence-only versus presence-absence data). A canonical problem is the integration of probability samples of presence-absence data with non-probability samples of presence-only data (Dorazio, 2014; Fithian et al., 2015; Simmonds et al., 2020). Here, we develop theory for datasets that differ in sampling design but are consistent in other aspects.

Despite the intuitive appeal of integrating probability and nonprobability samples, several questions must be answered before the approach can be recommended for biodiversity monitoring. For example, which integrated estimator has the best statistical properties? Does the optimal estimator vary across plausible data generating processes? Are there estimators that consistently perform well, even if they are not optimal?

We conducted an *in silico* experiment to address these three questions. The experiment comprises 64 scenarios that span a range of plausible data generating processes. Factors that vary between scenarios include, for example, the focal species' occupancy, the dependence of the nonprobability sampling process on the study variable, and the quality of information available for bias correction. The target parameter is the change in mean occupancy between years (on the log odds scale), but we expect our results to generalise to other study variables and parameters (see the *Discussion*). Four estimators are considered: one each for the probability and nonprobability samples and two that combine the samples in different ways. The estimators' statistical properties are compared across scenarios to understand which performs best and when, and whether any are consistently competitive.

Methods

Overview

The basic premise of our *in silico* experiment is as follows. There is a virtual landscape (i.e. target population) monitored over two years. Some sites are occupied by a focal species each year, some are included in a relatively small stratified random sample (StRS; a form of probability sample), and some are included in a much larger nonprobability sample (NS). It is not unusual for both a probability and a nonprobability sample to be available (Table 1). A confounding variable determines whether sites are occupied by the focal species and included in the NS, so the two are dependent (Boyd et al., 2025). The dependence of site selection in the NS on occupancy causes site-selection bias. Site inclusion in the StRS is independent of occupancy.

There is a virtual analyst seeking to estimate the change in mean occupancy between years. The analyst considers four estimators: one based on the StRS, one based on the NS, and two that integrate both samples. Information on the confounding variable is incorporated into the estimators that leverage the NS to limit the site-selection bias. The statistical properties of the four estimators are compared across a range of scenarios that differ in terms of the focal species' mean occupancy, how occupancy changes between years, the dependence of site inclusion in the NS on occupancy, how this dependence changes over time (which is important for estimating a trend), and the quality of the information available to the analyst to correct it. More detail on each component of the simulation framework is provided in the next section.

Table 1. A non-exhaustive list of scenarios in which both a probability and nonprobability sample are available for biodiversity monitoring. Programmes that aim for a probability sample but do not achieve it due to incomplete uptake (i.e. some sites are randomly selected but not sampled) are listed as producing nonprobability samples.

Application	Probability sample	Nonprobability sample	Details of probability sample	Details of nonprobability sample
European Butterflies	European Pollinator Monitoring Scheme (EU-PoMS)	National Butterfly Monitoring Schemes; 15-minute counts	EU-PoMS is due to be implemented in 2027 as part of the EU Nature Restoration Regulation (Potts et al., 2021). The design will be a stratified random sample.	National butterfly monitoring schemes provide count data for every species, but recorders decided where data are collected. Likewise for 15-minute counts.
UK Pollinator Monitoring Scheme (O'Connor et al., 2019; UK Pollinator Monitoring Scheme, 2026)	Randomised survey of 1 km squares	Public FIT (Flower Insect Timed) counts	Stratified random sample of pan traps and FIT counts (O'Connor et al., 2019). Professionals are paid to sample sites that aren't taken up by volunteers (UK Pollinator Monitoring Scheme, 2026).	All pollinating insects are recorded but data collectors decide where to sample. Not resolved to species level.

Birds in in the Rocky Mountains, US	Integrated Monitoring in Bird Conservation Regions programme	eBird (Sullivan et al., 2014)	Spatially balanced design stratified by e.g. state boundaries and bird conservation regions (Pavlacky et al., 2017).	Recorders decide where to sample but are encouraged to submit complete lists.
Birds in Switzerland	Periodic Swiss Breeding Bird Atlases (Knaus et al., 2018)*	Ornitho.ch	Complete census of 10 x 10 km squares in each atlas period.	Official birding exchange programme in Switzerland. Houses data on ~300,000 complete lists from various sources.
Plants in the United Kingdom	Countryside Survey	[UK] National Plant Monitoring Scheme (NPMS; Pescott et al., 2019)	Stratified random sampling. Early surveys were decadal snapshots; nowadays there is a rotating panel design (i.e. different sites sampled in different years)	NPMS aims for a stratified random sample but does not achieve it due to dropout
Plants in Wales	ERAMMP	[UK] National Plant Monitoring Scheme	Stratified random sampling.	NPMS aims for a stratified random sample but does not achieve it due to dropout

* technically this is not a probability sample but a complete census of sites. We include it in the probability sample column because it shares the same useful property that sample inclusion does not depend on the study variable.

Simulation framework

Target population

The target population is a landscape comprising $i = 1, \ldots, 10{,}000$ sites monitored over two years $t \in \{0,1\}$. For any variable A that varies by both site and year, we denote its value for site i in year t by A_{it}. When referring to the vector of site-level values in a given year t, we suppress the site index and write $A_t = \{A_{1t}, A_{2t}, \ldots, A_{nt}\}$. Likewise, we suppress the year index when referring to the time series for site i, writing $A_i = \{A_{i0}, A_{i1}\}$.

Species occupancy

For simplicity, we simulate one focal species' occupancy and how it changes between years. Whether site i is occupied by the species in year t is a Bernoulli random variable with probability φ_{it}, which depends on an environmental variable X_{it}:

$$Y_{it} \sim Bernoulli(\varphi_{it}), \qquad logit(\varphi_{it}) = \alpha + \beta_X X_{it}. \tag{1}$$

The parameter α determines the baseline occupancy probability, and β_X is the effect of X_{it}.

To simulate a change in occupancy from one year to the next, we modify X_{it}. In year one, $X_{i0} \sim N(0,1)$. In year two, $X_{i1} = X_{i0} + \varepsilon_i$, where $\varepsilon_i \sim N(\delta, 0.1)$. Hence, the parameter δ determines the

shift in the distribution of X_t and therefore in occupancy. Change in the distribution of X_t might reflect, say, the urbanisation of sites or climate change.

Sample inclusion

Probability sample

The probability sample is a StRS with proportional allocation and a fixed panel design. That is, strata are sampled in proportion to their size, and the same sites are sampled in both years. Whether site i in stratum $h \in \{1, \dots, H\}$ is included in the StRS is a binary variable whose expected value is given by

$$E\left[R_i^{(StRS)}\right] = \pi_i^{(StRS)} = \frac{n_h}{N_h}, \tag{2}$$

where N_h is the number of sites in stratum h and, here, $n_h = n\left(\frac{N_h}{N}\right)$ is the target sample size in that stratum. Importantly, (2) implies that $R_i^{(StRS)}$ is independent of occupancy (within strata) by design. We assume for simplicity that there are $H = 10$ strata of equal size; more details on the stratification are provided below. In all scenarios, $E\left[R_i^{(StRS)}\right] = 0.01$, meaning 100 sites are included in the StRS.

StRS is more efficient (yields lower sampling variability) when the stratifying variables explain more of the variation in the study variable (here occupancy). In our simple simulation framework, it is only the variable X that affects occupancy, and this would be the optimal variable on which to stratify (noting that continuous variables would first have to be discretised).

While optimal, stratifying on X would not reflect real-world practice. Even for a well-studied species, the analyst is unlikely to have perfect knowledge of all variables explaining its occupancy. And even if they did, most monitoring programmes target multiple species whose occupancies are explained by different variables. It would not be sensible for a multi-species programme to stratify on variables that explain one species' occupancy.

Since stratifying on X would not be realistic, we assume that a suboptimal covariate $X_{t=0}^{(obs)}$ is used instead. $X_t^{(obs)}$ has a predefined, time-invariant correlation with X_t, and the strength of this correlation determines the effectiveness of the stratification. As the correlation approaches ± 1, the sampling design approaches maximum efficiency for the focal species.

Nonprobability sample

A more complex model than (2) is needed to simulate sample inclusion in the NS, which is collected by citizen scientists. The model must accommodate a (possibly time-varying) dependence of site inclusion on occupancy. It must also permit an increase in the number of sites that are sampled over time, which is a common feature of citizen science datasets (Di Cecco et al., 2021).

Site inclusion depends on occupancy when the two share a common cause (here X). (Site inclusion could also depend directly on occupancy, but we do not consider that situation here.) Without loss of generality, imagine that the environmental variable X_{it} is the degree to which site i is 'built-up', or urbanised, in year t. The species might be an 'urban avoider', in which case it is less likely to occupy sites with high levels of X_{it}. At the same time, highly urbanised sites are more accessible, so they might be preferentially sampled by citizen scientists. In this scenario, occupancy would be lower at sampled sites than elsewhere (although the sign and magnitude of the bias depend on which variable(s) X is supposed to represent).

If the composition of citizen scientists, or their motivations, changes over time, then the dependence of sample inclusion on occupancy might change too. Indeed, Bowler et al. (2022) showed that the proportion of urban sites in a German citizen science dataset increased over time. Hence, the model

for site inclusion in the NS must permit change over time in the effect of X on sample inclusion and thus on the degree to which it depends on occupancy.

A reasonable model for site inclusion in the NS that accommodates a time-varying dependence of sample inclusion on occupancy and sample size is

$$R_{it}^{(NS)} \sim Bernoulli\left(\pi_{it}^{(NS)}\right),$$
$$logit\left(\pi_{it}^{(NS)}\right) = \underbrace{\left(\alpha^{(NS)} + \beta_{year}^{(NS)}\ year_t\right)}_{intercept} + X_{it}\underbrace{\left(\beta_X^{(NS)} + \ year_t\ \ \beta_{X\cdot year}^{(NS)}\right)}_{slope\ for\ X}. \quad (3)$$

The parameter $\beta_X^{(NS)}$ defines recorders' preferences for sites with certain values of X_{it}. If $\beta_X^{(NS)} \neq 0$, then X_{it} is a common cause of occupancy and inclusion in the NS, so the two are dependent. $\beta_{X\cdot year}^{(NS)}$ determines whether the effect of X_{it}, and thus the dependence of site inclusion on occupancy, changes over time. $\beta_{year}^{(NS)}$ captures year-to-year shifts in the log-odds of sample inclusion, and hence in the expected NS sample size, independent of X.

The estimand

The target parameter—or estimand—is

$$\Delta = logit(\bar{Y}_1) - logit(\bar{Y}_0), \quad (4)$$

where $\bar{Y}_t$ is mean occupancy (i.e. the proportion of occupied sites) in year t. This quantity is equivalent to a log odds ratio comparing occupancy between years, is symmetric around zero and remains well behaved when occupancy in either year is close to 0 or 1. It is also equivalent to the value of β_{year} that would be obtained by fitting the model

$$Y_{it} \sim Bernoulli(\varphi_{it}), \qquad logit(\varphi_{it}) = \alpha + \beta_{year}\ year_t \quad (5)$$

to a census in which Y_{it} is measured without error for every site in both years.

The estimators

We consider four estimators of Δ. One is based on the probability sample, one is based on the nonprobability sample, and the remainder are based on a combination of the two. Our goal was to construct 'sensible' estimators that analysts might opt for in the real world.

Model-based probability sample estimator (StRS-MB)

Under StRS with proportional allocation, the Maximum Likelihood Estimator (MLE) of β_{year} in (5) is approximately design-unbiased for Δ and is therefore a natural estimator.

Model-based nonprobability sample estimator (NS-MB)

Under NS, the MLE of β_{year} is not necessarily unbiased for Δ unless appropriate weights are applied to the likelihood. The ideal weight for site i in year t is the reciprocal of the probability that it was included in the NS: that is, $1/P\left(R_{it}^{(NS)} = 1\right)$. Since these probabilities are not known to the analyst, they must be estimated.

The usual approach to estimating sampling probabilities is to fit a logistic regression (or some other model) to estimate $P\left(R_{it}^{(NS)} = 1|A\right)$, where A is a matrix of 'auxiliary variables' that are measured at every site in every year (e.g. surface temperature or land cover measured by satellite; Boyd et al., 2024; Elliott & Valliant, 2017; Johnston et al., 2020). If the auxiliary variables completely explain the dependence of site inclusion in the NS on occupancy, then the weighted MLE of β_{year} in (5) is

design-unbiased for Δ. In our simple simulation framework, the dependence is explained solely by X. Hence, the weighted MLE of β_{year}, where the weights are given by $1/P\left(R_{it}^{(NS)} = 1|X_{it}\right)$, would be design-unbiased for Δ.

As explained in the section *Sample inclusion*, the analyst is unlikely to have perfect knowledge of or data on the variable(s) X. Hence, in the same way that it would not have been realistic to stratify the StRS on X, it would not be realistic to estimate sampling probabilities conditional on X. Rather, sample inclusion probabilities are estimated conditional on the variable $X^{(obs)}$, which is correlated with X. The closer this correlation to ± 1, the closer the weighted MLE of β_{year} based on the NS is to being unbiased for Δ.

Integrated estimators

Model-based integrated estimator (I-MB)

The model-based estimator for the combined StRS-NS is the joint MLE of β_{year} [i.e. a joint likelihood integrated model (Miller et al., 2019)]. Since the StRS and NS are independent conditional on the model parameters, the joint likelihood factorises into the product of their marginal likelihoods (or, equivalently, the sum of their log-likelihoods). We impose a shared β_{year} across samples, but each sample has its own intercept. The weights $1/P\left(R_{it}^{(NS)} = 1|X_{it}\right)$ are applied to the NS component of the likelihood as described above.

To prevent double counting, site-year combinations that are included in both samples are split between the two likelihood components using a fixed fraction 0.5. We do not expect the precise choice of splitting fraction to have an appreciable effect on the results. Detection is assumed to be perfect, so overlapping site-year combinations contribute identical occupancy observations. Moreover, overlap between datasets is expected to be small. The StRS covers 1% of sites, and the NS covers up to approximately 10% of sites (Fig. 2). Since the StRS is an equal-probability design as implemented here, inclusion in the StRS is independent of inclusion in the NS. Hence, the maximum expected proportion of sites appearing in both samples is approximately the product of the sampling fractions ($0.01 \times 0.10 = 0.001$), corresponding to about 0.1% of sites. Even at maximum NS coverage, this implies that only around 10% of StRS observations are expected to also feature in the NS.

Design-based integrated estimator (I-DB)

The design-based estimator for the combined StRS-NS was proposed by Rao (2021). An accessible introduction to the approach can be found in Fig. 1. Here we provide the technical details.

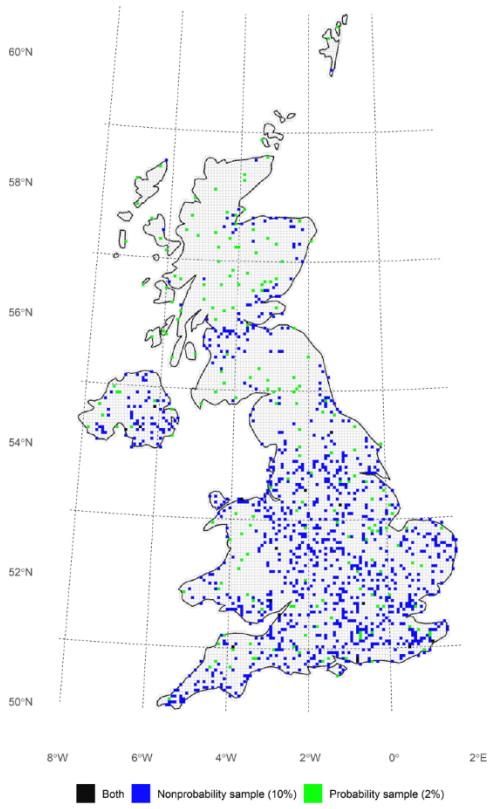
Let B_t be the set of sites included in the nonprobability sample in year t and C_t be its complement (the set not included). Recalling that we assume no measurement error, occupancy in B_t is known exactly from the nonprobability sample (since the nonprobability sample is a census of this stratum). We come back to potential extensions of the I-DB estimator that can accommodate measurement error in the *Discussion*.

For C_t, we have a probability sample (the StRS), so design-based inference applies. Under StRS with proportional allocation, each site has the same probability of being included in the sample, so the sample proportion of occupied sites is a design-unbiased estimator of stratum-wide occupancy in C_t. Data from the StRS collected at sites also included in B_t are ignored, although in practice one might use them instead of the data from NS at those sites if measurement error is suspected to be smaller in the StRS. Here the distinction does not matter because we assume no measurement error (see the *Discussion* for more on the implications of this assumption).

Population-level occupancy in year t can be estimated as a weighted mean of the known occupancy in B_t and the design-based estimator for C_t, with the weights being proportional to each stratum's size.

This integrated estimator is design-unbiased for the population occupancy, and its variance is smaller than a design-based estimator applied solely to the StRS (Rao, 2021; see *Discussion*).

The difference between the integrated estimators for years one and two is an unbiased estimator of the true difference in population-level occupancy between years. It is not strictly unbiased for the estimand Δ, which is defined on the log odds scale. However, it is approximately unbiased when occupancy in neither year is close to 0 or 1, and it is consistent for Δ regardless.



The design-based integrated estimator (I-DB) of Rao (2021) is implemented as follows:

1. Split the landscape into two strata: one for grid squares that are included in the nonprobability sample (blue and black squares on the map to the left), and one for the rest
2. Calculate mean occupancy across squares in nonprobability sample stratum. We have data for every square in this stratum, so no estimation is required*
3. Estimate mean occupancy for the remaining squares using the probability sample. Standard design-based estimators can be used
4. Take the weighted mean of the stratum means, where each stratum's weight is proportional to its size. This is the per time-period estimator
5. If we treat the nonprobability sample as fixed, the variance of the estimator is determined entirely by uncertainty in the probability-sampled stratum
6. Repeat for all time-periods of interest then summarise change over time using the chosen metric

*assuming no measurement error. This assumption can be relaxed by e.g. modelling detection probabilities

Figure 1. 'Recipe' for implementing the design-based integrated estimator (I-DB; Rao, 2021). The map depicts two hypothetical samples collected in the UK. One is a probability sample (a simple random sample) covering 2% of 5 km grid squares, and the other is a nonprobability sample covering 10% of grid squares. Squares with high population density were preferentially selected for inclusion in the nonprobability sample. These samples were produced for illustration are not used in the simulations.

Estimation error

We use the term "estimation error" to describe the discrepancy between a given estimator and the estimand (assuming no measurement error, a point we come back to in the *Discussion*). In the design-based framework, estimation error is defined with respect to the sampling distribution. Hence, for each scenario (see *Scenarios* below), we conduct $K = 1000$ replicate simulations with different realisations of the sampling design. Of course, occupancy is simulated using the model in equation 1, so it too varies between simulations. This is standard practice in simulation studies (Monte Carlo approximation) and means our results should not be interpreted as design-based statistical properties but rather approximations to them.

We consider the following four metrics of estimation error:

1. Mean Squared Error (MSE): $\frac{1}{K}\sum_{k=1}^{K}\left(\hat{\Delta}_k - \Delta_k\right)^2$, where $\hat{\Delta}_k$ is the estimate of Δ_k for replicate k.
2. Absolute bias: $\left|\frac{1}{K}\sum_{k=1}^{K}\left(\hat{\Delta}_k - \Delta_k\right)\right|$.
3. Variance: $\frac{1}{K}\sum_{k=1}^{K}\left(\hat{\Delta}_k - \underline{\hat{\Delta}}\right)^2$, where $\underline{\hat{\Delta}}$ is the mean of $\hat{\Delta}_k$ across simulations.
4. 95% confidence interval coverage: $1/K\ \sum_{k=1}^{K} I(L_k \leq \Delta_k \leq U_k)$, where $I(\cdot)$ is the indicator function taking the value 1 if the argument is true and zero otherwise, L_k is the lower bound of the confidence interval, and U_k is the upper bound. Confidence intervals were constructed using the standard approach for each estimator. The STRS and NS model-based estimators were fitted using the `glm()` function in R, and profile likelihood confidence intervals were obtained using `confint.glm`. For the I-DB estimator, the variance of the estimated occupancy was propagated to the logit scale using the delta method before constructing a Wald interval. For the I-MB estimator, Wald confidence intervals were constructed using standard errors obtained from the inverse observed Hessian.

Scenarios

To understand the estimation error of the four estimators in different situations, we constructed scenarios that differed in terms of six factors:

1. *The focal species' baseline mean occupancy in year 1*. Scarce species are less likely to picked up by smaller samples (Delgado, 2022; Hertzog et al., 2021). The risk is greater for probability samples, which tend to be smaller than nonprobability samples.
2. *The sign and magnitude of the change in occupancy between years*.
3. *The baseline size of the nonprobability sample* (the number of sites that were sampled in year 1).
4. *The baseline data defect (capturing the dependence of site inclusion in the NS on occupancy in year 1)*. The data defect is defined as the correlation between whether a site was included in the NS $R_1^{(NS)}$ and occupancy in year 1 (Meng, 2018). Its expectation across replicate samples is proportional to the bias of the sample mean as an estimator of the population mean (Aubry et al., 2024). Positive values indicate preferential sampling of occupied sites, whereas negative values indicate preferential sampling of unoccupied sites.
5. *The change in the data defect between years*. If the data defect changes over time, then estimates of temporal change become biased. If the data defect remains constant, then they do not (Bowler et al., 2022). This result is exact for linear estimands and holds approximately for non-linear estimands such as the difference in logits considered here.
6. *The correlation between the true biasing variable Z and the available covariate X*. The greater the correlation, the more efficient the STRS estimator and the less biased the NS-MB and NS-StRS estimators. We assume this correlation is time-invariant.

To construct the scenarios, we first identified the simulation parameters that control the 6 factors listed above and assigned each one a distribution (see supplementary material). These distributions were initially derived from first principles (e.g. by choosing logit-scale intercepts that produce realistic probabilities once back-transformed, reasonable correlations, etc.). From the initial distributions we randomly sampled 10,000 parameter sets, simulated a dataset for each one and inspected the resulting distributions of the six factors. This 'prior predictive check' prompted us to make some small refinements to the parameter distributions to obtain more realistic scenarios that more closely aligned with literature values. The final prior predictive distributions of the six factors are presented alongside literature values in Fig. 2, demonstrating that our simulated data are in accord with real studies where data are available. The final distribution for each of the 6 simulation parameters is listed in the supplementary material.

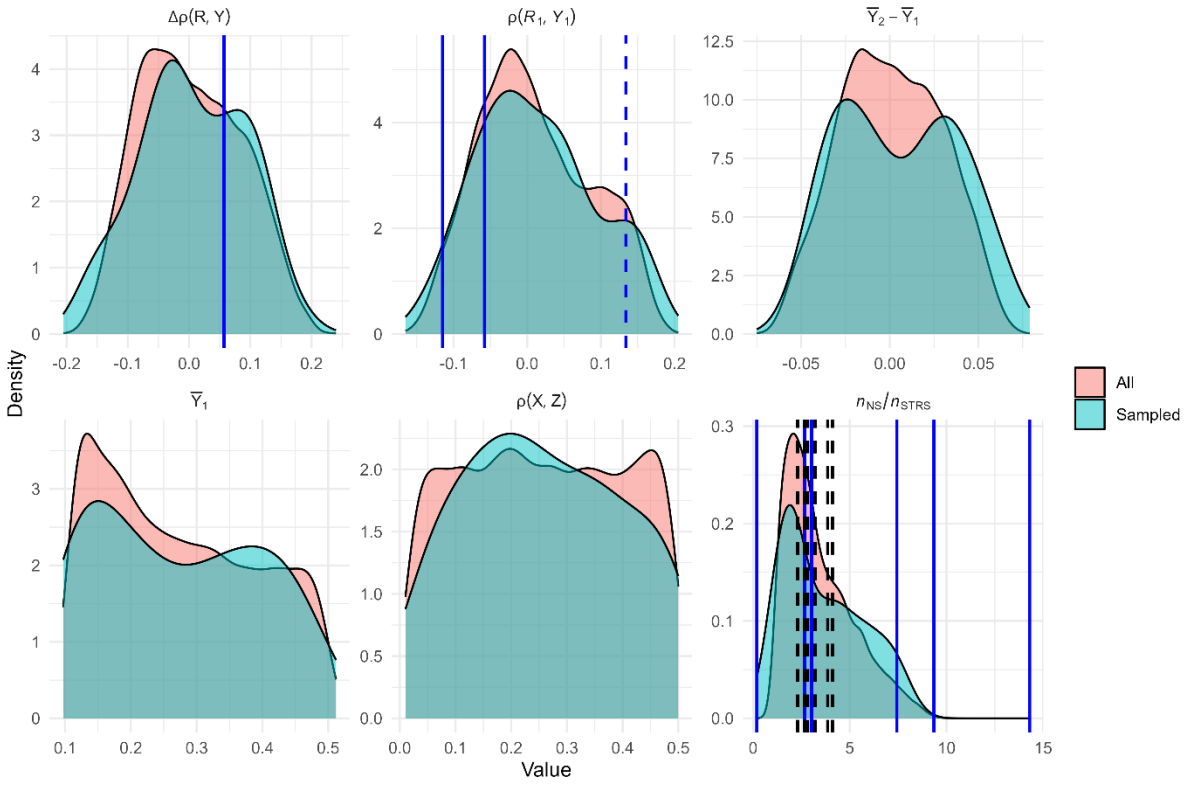


Figure 2. Close match between the distributions of the six factors that vary across scenarios generated by the 10,000 parameter sets (red) and the 64 sampled parameter sets (turquoise). The distributions are also generally in accordance with literature values (vertical lines) where data are available. Row-wise from the top left, the summary statistics are the change in the data defect between years, the baseline data defect in year 1, the difference in mean occupancy between years, the baseline occupancy in year 1, the correlation between the available covariate and the true biasing variable, and the ratio of the nonprobability sample size to the probability sample size in year 1. For the literature values, different combinations of line type (dashed versus solid) and colour distinguish studies. For the data defect and how it changes between years, values were taken from Boyd, Harvey, et al. (2023; see supplementary material for the derivation) and Boyd et al. (2024). The solid blue lines in the bottom right panel represent the ratio of the number of one-kilometre squares included in the UK Pollinator Monitoring Scheme randomised survey to the number of squares in which Flower-Insect Timed counts were conducted by the public. There is one value per year from 2017 to 2022. The black dashed lines represent the ratio of the number of one-kilometre squares surveyed by the National Plant Monitoring Scheme in Britain to the number sampled by the Countryside Survey. There is one value per year from 2019 to 2025.

While it was possible to conduct 10,000 prior predictive simulations, the same would not have been possible for the scenarios, since each one involves 1000 replicates. Running 10,000 scenarios, each with 1000 replicates, would require 10^7 simulations. Instead, we took a stratified random sample of 64 parameter sets according to the following algorithm. Each of the 6 factors was split at its midpoint to form two levels, and these were cross tabulated to define 64 strata. One parameter set was randomly sampled from each stratum. The resulting subset achieved excellent coverage of both the marginal and joint distributions of the 6 factors from the full set of 10,000 prior predictive simulations (Fig. 2 and Fig. S1). This algorithm is closely related to Latin hypercube sampling but enabled us to ensure good coverage of the 6 factors rather than the simulation parameters that generated them.

Estimator comparisons

When comparing estimators across scenarios, we ask three questions:

1. How often does each estimator perform 'best'?
2. Under what circumstances does it perform best?
3. Is it competitive across scenarios, even when it does not perform best?

To understand how frequently each estimator performed best, we tallied the number of scenarios in which it had the lowest MSE, lowest bias, lowest variance and highest confidence interval coverage. The overall winner for each metric is the estimator that performed best in the most scenarios. To assess Monte Carlo uncertainty, we bootstrapped the simulation replicates within each scenario, repeated the comparison, and calculated the proportion of bootstrap samples in which the original overall winner remained the overall winner.

Identifying the circumstances in which each estimator performs best is a classification problem. For each scenario, we designated the best estimator the one with the lowest MSE and encoded this as a categorical response variable. MSE is an appropriate criterion because it reflects both bias and variance. The 6 factors listed above (baseline occupancy, occupancy trend, baseline data defect, shift in data defect, baseline NS size and covariate quality) were used as predictors. We fitted a classification tree using the R package `rpart` with default setting and pruned it using the conventional one-standard-error rule (Breiman et al., 1984). To assess the stability of the inferred splitting rules, we generated 1000 bootstrap resamples, refitted the tree to each one, and tabulated how often the same variable appeared as the 'root' split (i.e. the most important predictor).

To establish whether each estimator was competitive across scenarios, even when it was not the best, we computed its *relative regret*. For a given estimator and scenario, the relative regret is the proportional increase in MSE (since it reflects both bias and variance) relative to the optimal estimator. The regret is zero for the optimal estimator itself. For each estimator, we summarised its performance using the median relative regret across scenarios.

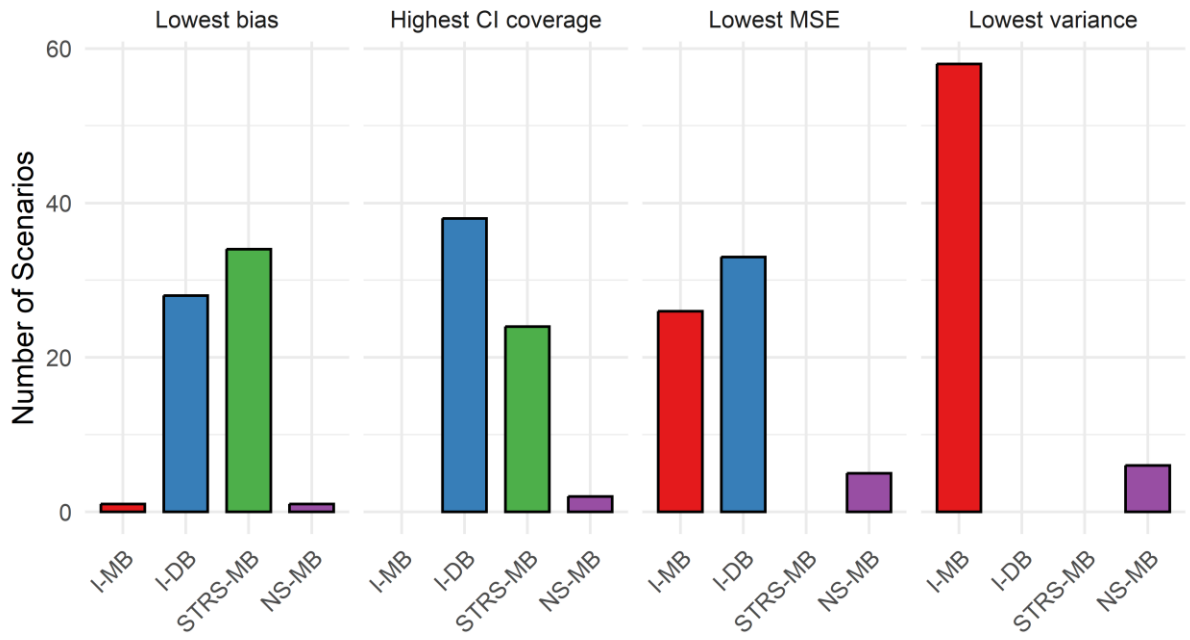
Results

Which estimator is the best?

Across the 64 scenarios, the I-DB (design-based integrated) estimator performed best overall. It was the winner in terms of confidence interval coverage (59% of scenarios) and MSE (52% scenarios), and it was a close second in terms of bias (44% of scenarios; Fig. 3). The I-MB (model-based integrated) estimator had the lowest variance in 91% of scenarios, but it ranked second for MSE (41% of scenarios) because it was almost always more biased than I-DB.

These results were robust to Monte Carlo error. The overall winning estimator remained unchanged in all 1000 bootstrap replicates for MSE, in 94% of replicates for confidence interval coverage, and in all 1000 replicates for variance (see the supplementary material). Absolute bias was the only exception: the overall winner alternated between I-DB and StRS-MB across bootstrap replicates, indicating that

406 there was little evidence that either estimator consistently outperformed the other for this metric.



407

408 *Figure 3. (Two columns.) "Best" performing estimator across the 64 scenarios by performance*
409 *metric. I-MB is the model-based integrated estimator, I-DB is the design-based integrated estimator,*
410 *NS-MB is the nonprobability sample model-based estimator and StRS-MB is the stratified random*
411 *sample model-based estimator.*

412 For an estimator to provide 'valid' inference, its confidence interval must achieve nominal coverage
413 (here 95%). With a tolerance of $\pm 5\%$, only the I-DB and StRS-MB (model-based stratified random
414 sample) estimators consistently achieved nominal (95%) confidence interval coverage (Fig. 4).

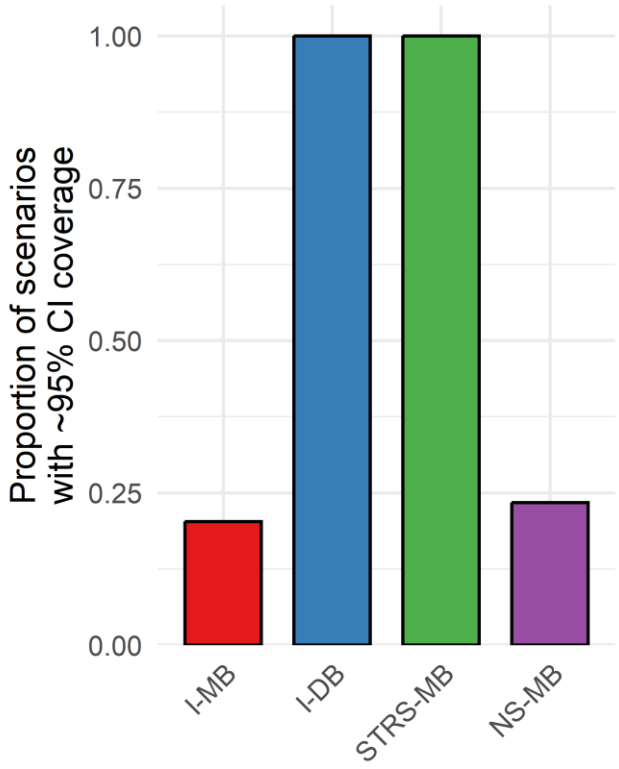


Figure 4. (One column.) Proportion of scenarios in which each estimator achieved nominal coverage with a tolerance of $\pm 5\%$*. I-MB is the model-based integrated estimator, I-DB is the design-based integrated estimator, NS-MB is the nonprobability sample model-based estimator and StRS-MB is the stratified random sample model-based estimator.*

What determines which estimator has the lowest MSE?

Recall that we fitted classification trees to understand which of the 6 factors (baseline occupancy, change in occupancy, baseline data defect, change in data defect, baseline NS sample size, quality of the covariate data) determined which estimator had the lowest MSE across scenarios. The absolute magnitude of the shift in the data defect between years $\Delta\rho(R,Y)$ (capturing the change in the dependence of site inclusion in the NS on occupancy between years) consistently emerged as the most important determinant. It was the 'root split' (i.e. the most important predictor) in 96% of the 1000 trees fitted to the bootstrap resamples. When $\Delta\rho(R,Y) < 0.028$, the shift in the dependence of site inclusion in the NS on occupancy is negligible, and the I-MB estimator wins (Fig. 4). This occurred in 14 of the 64 scenarios. Otherwise, the classification tree predicts that the I-DB estimator has the lowest MSE. Since we pruned the classification tree to avoid overfitting, neither class is pure in the sense that it had only one winner (Fig. 4).

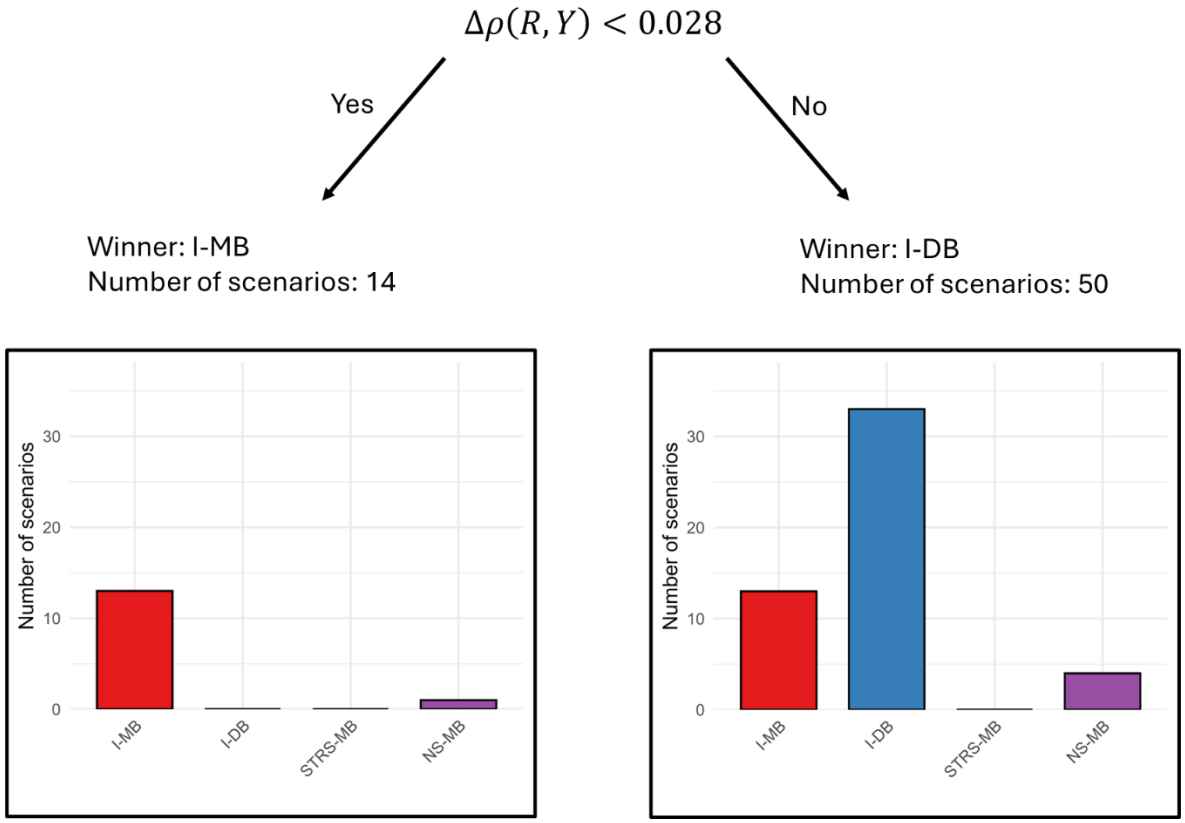


Figure 4. (Two columns.) Splitting rule from the final decision tree classifying which estimator had the lowest MSE. $\Delta\rho(R,Y)$ *is the change in the data defect over time. The barplots show the number of scenarios in which each estimator truly had the lowest MSE in each category (recalling that the decision tree does not split the data into 'pure' classes with only one winner due to pruning).*

Which estimators are competitive?

The I-DB and StRS-MB were consistently competitive, even when they did not perform best, whereas the others were not. For the I-DB estimator, the median increase in MSE relative to the optimal estimator (i.e. the regret) was 0% (Fig. 5), for StRS-MB it was 7%, for I-MB it was 40% and for the NS-MB it was 207%. These results imply that the I-DB and StRS-MB estimators are relatively 'safe' in the sense that they rarely perform disastrously relative to the optimal estimator.

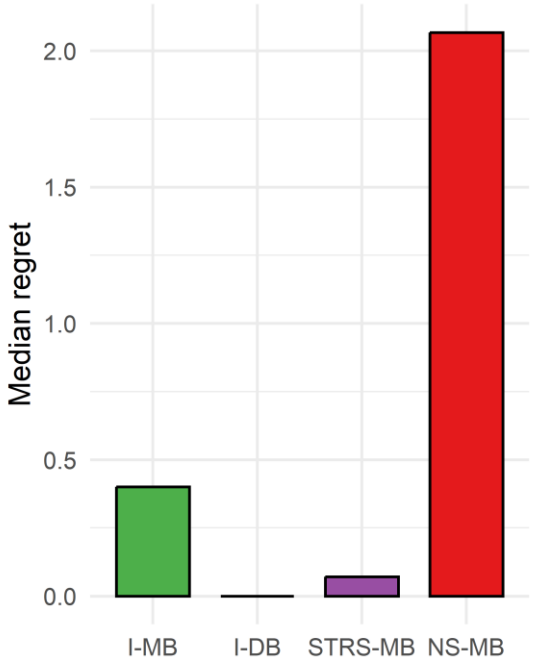


Figure 5. (One column.) Median regret across scenarios per estimator. For a given scenario and estimator, the regret is the proportional increase in MSE relative to the optimal estimator. I-MB is the model-based integrated estimator, I-DB is the design-based integrated estimator, NS-MB is the nonprobability sample model-based estimator and StRS-MB is the stratified random sample model-based estimator.

Discussion

There are many estimators available to analysts seeking to understand biodiversity change. Depending on what is available, each estimator may leverage one or more datasets. We proposed four candidate estimators that leverage a small probability sample, a large nonprobability sample, or both. We then conducted an *in silico* experiment to test the statistical properties of those estimators across a range of plausible scenarios.

Which of the tested estimators is best?

Of the four estimators that we considered, the model-based estimator that leverages the nonprobability sample, NS-MB, was by far the worst. It rarely performed best in terms of any performance metric, and when it lost, it lost by a lot. This is indicated by its median relative regret (i.e. proportional increase in MSE relative to the best estimator) of 207% (Fig. 5). Basing estimation solely on a nonprobability sample when a probability sample is available should be avoided.

At first glance the I-MB estimator appears to be a good candidate, since it had the lowest MSE in many scenarios. On closer inspection, however, it is not a safe option. It rarely achieved nominal confidence interval coverage, and it generally did not perform well in terms of MSE unless the dependence of site inclusion in the NS on occupancy was stable over time (Fig. 4). It is challenging to quantify the dependence of site inclusion on occupancy (or any other relevant study variable), let alone how it has changed over time (Manski, 2007, pp. 56-57; but see below). Hence, it is hard to see how the analyst could trust the I-MB's point estimate or confidence interval in any given situation.

By far the safest of the tested estimators are the design-based integrated estimator, I-DB, and the estimator based only on the probability sample, StRS-MB (i.e. discarding the NS data altogether). Both achieved nominal confidence interval coverage in all scenarios and were consistently competitive in terms of MSE (median regrets of 0 and 7%). The I-DB estimator slightly outperformed the StRS-MB in most respects, but usually not by much. Hence, both appear to be good options.

Since both estimators are (approximately) design-unbiased, the improvement in MSE achieved by the I-DB relative to the StRS-MB is driven by a reduction in variance. Let us set to one side for a moment the additional variance introduced by using Monte-Carlo simulation to approximate design-based properties. Rao (2021) shows that the *design variance* of the I-DB estimator, relative to an estimator based solely on the probability sample, decreases in proportion to the fraction of sites covered by the nonprobability sample but not the probability sample. That is, the more sites covered only by the nonprobability sample, the larger the expected gain in precision from adopting the I-DB estimator.

Alternative estimators

The four estimators tested here are not the only options. Perhaps the most obvious modification would be to incorporate covariates or site-specific random effects in the model-based estimators. Incorporating additional model terms that explain a portion of the variation in the outcome variable would improve the model's precision. In addition, incorporating variables that jointly affect site inclusion in the NS and the study variable (the environmental covariate X in our simulations) would reduce its bias (Boyd et al., 2025). Incorporating site-specific random effects will not necessarily reduce site-selection bias, because the estimated distribution of site effects might itself be biased due to preferential sampling.

For a model that includes additional terms, the MLE of its year effect is no longer an appropriate estimator of the descriptive estimand in equation (4). The problem is that by including additional terms, the year effect becomes a conditional quantity, which does not align with the marginal difference in the outcome variable between years. One solution in this case would be to use the model to predict the outcome for each combination of site and year and then apply the relevant summary (e.g. the difference in mean occupancy) to those predictions (Boyd et al., 2025; Elliott & Valliant, 2017).

In practice it would also be prudent to extend the estimators in such a way that they can deal with measurement error. The standard approach is to incorporate "observation" models that compensate for imperfect detection (Emmet et al., 2023; Pagel et al., 2014; Strebel et al., 2022). By modelling the observation process and weighting the "state" model's likelihood, it is possible in principle to construct a model-based estimator that simultaneously accounts for measurement error and the dependence of sample inclusion on the study variable (Irvine et al., 2018; Johnston et al., 2020).

The design-based integrated estimator, too, could be extended to account for measurement error. Recall that this estimator is constructed by first partitioning sites into those that are covered by the nonprobability sample and those that are not (Fig. 1). The estimate of the mean of the study variable in the first stratum is simply the mean of the observations, and for the second stratum standard design-based estimators are used. In principle, estimators that account for measurement error (i.e. that include observation models) could be used for both strata to create a model-assisted extension of the basic design-based estimator tested here.

In fact, one way to view Rao's (2021) integrated design-based estimator is as a special case of a broader class of estimators based on the decomposition of sites into those that are covered by the nonprobability sample and those that are not. Any estimators of the stratum means that account for the relevant observation processes can then be applied. This class of estimators has the extremely desirable property that, under standard design-based assumptions and appropriate within-stratum

estimators, it unbiased with respect to site selection for linear estimands and approximately so for nonlinear ones. Measurement error may still cause bias, but the stratum mean estimators can be set up to address this.

Each of the estimators tested here requires datasets that are reasonably consistent in terms of sampling protocol. For example, the two datasets might both yield counts or detections and non-detections. More elaborate estimators are required when the two datasets are not directly comparable (Dorazio, 2014; Farr et al., 2021; Fithian et al., 2015; Isaac et al., 2020; Mäkinen et al., 2023; Mostert & O'Hara, 2023).

Importance of probability sampling

Our results emphasise the fact that probability sampling is needed to guarantee that any given estimator will have good statistical properties. Without it, strong and untestable assumptions are required. If those assumptions fail, bias is to be expected, and confidence intervals cannot be trusted. It is important to be clear about limitations of nonprobability sampling; otherwise, funders might default to cheaper and lower quality forms of data collection (Boyd, Powney, et al., 2023).

We are not the first to find that integrating multiple datasets for parameter estimation increases precision (Hertzog et al., 2021; Mancini et al., 2022). Indeed, this result could be derived from first principles. Given an unbiased estimator, precision is clearly desirable for detecting change (Leung & Gonzalez, 2024; Valdez et al., 2023). However, we should remember that a precise and biased estimator merely gives us false confidence in the wrong answer (Aubry et al., 2026; Meng, 2018).

What if there is no probability sample?

Of course, there are many circumstances in which a probability sample is not available. In these cases, we must do the best that we can. An example of good practice might be to conduct a risk of bias assessment (Boyd et al., 2022), which can be used to qualify results (Powney et al., 2025) or better still to inform quantitative sensitivity analyses (Hartman & Huang, 2024; Little et al., 2020).

It is also important to recognise that not all nonprobability samples are equal. An estimator applied to a nominal probability sample with a small amount of dropout is likely to be less biased than the same estimator applied to an opportunistic dataset with no formal sampling design (Bailey, 2023). Defining a threshold level of uptake below which a probability sample should instead be regarded as a nonprobability will require individual judgement based on the potential level of bias that is considered acceptable.

Faced with multiple nonprobability samples, a Rao-type integrated estimator might still be a good option. One pragmatic approach would be to treat the sample that most closely resembles a probability sample as the "reference sample" and the other as the nonprobability sample. The strata would then be defined as i) sites included in the nonprobability sample and ii) the rest. Set up like this, the integrated estimator would inherit the bias of the estimator for the uncovered stratum based on the reference sample. If the reference-sample estimator is less biased than an estimator applied directly to the nonprobability sample, and the nonprobability sample covers a substantial fraction of the population, the resulting integrated estimator may exhibit substantially less bias while also benefiting from increased precision.

If neither sample resembles a probability sample, then alternative approaches will be required. One option would be to estimate sample inclusion probabilities for *both* datasets and use their inverses to weight the likelihood of a model-based estimator. Whetten et al. (2026) proposed an "R learner" integrated estimator that essentially operates on this premise.

Applicability to alternative study variables and estimands

The advantage of Rao-type estimators is not limited to estimating change over time in species occupancy. Under mild assumptions, this class of estimators is unbiased for the population mean, regardless of the study variable (Rao, 2021; e.g. occupancy, abundance or species richness). Consequently, change over time in the population mean can also be estimated without site-selection bias by differencing estimates from successive years. These guarantees do not generally extend to nonlinear functionals of the study variable (e.g. the difference in the logit of mean occupancy between years). Nevertheless, for smooth transformations the resulting site-selection bias should be small except in extreme cases, and the estimators remain consistent. Measurement error may still introduce bias and should be addressed separately using appropriate observation models.

Recommendations

Based on our results, we recommend that analysts of biodiversity monitoring data adopt the following three principles:

1. Always use a probability sample for estimation if one is available
2. If a nonprobability sample is also available, and it covers many sites not included in the probability sample, consider integrating the two
3. When integrating probability and nonprobability samples, consider a Rao-type estimator (what we refer to as I-DB).

By following these principles, the analyst may be able to benefit from the best of both worlds—the bias properties of a design-based estimator applied to the probability sample, and an increase in precision due to incorporating the larger nonprobability sample. This may be contrasted with model-based integration, where strong assumptions are needed to claim unbiasedness. Measurement error may still cause bias but can be addressed using standard measurement models already familiar to ecologists.

Acknowledgements

RJB, MJP and SR were supported by the MAMBO (Modern Approaches to Monitoring Biodiversity) project, which receives funding from the European Union's Horizon Europe research and innovation programme project, which receives funding from the European Union's Horizon Europe research and innovation programme under grant agreement no.101060639. RJB was also supported by a NERC Independent Research Fellowship: UKRI4865: A new paradigm for biodiversity monitoring. OLP and DBR were partially supported by NERC, through the UKCEH National Capability for UK Challenges Programme NE/Y006208/1. OLP was also supported by the Joint Nature Conservation Committee funding to the UK National Plant Monitoring Scheme. GH was supported by the Swiss National Science Foundation (Postdoc.Mobility grant number P500PB_230797).

References

Aubry, P., Francesiaz, C., & Guillemain, M. (2024). On the impact of preferential sampling on ecological status and trend assessment. *Ecological Modelling*, *492*. https://doi.org/10.1016/j.ecolmodel.2024.110707

Aubry, P., Quaintenne, G., Guillemain, M., & Matutini, F. (2026). When n is not enough: Assessing the interplay of spatial sampling error and sample selection in abundance estimation. *Ecological Modelling*. https://doi.org/10.5281/zenodo.2

Bailey, M. A. (2023). A New Paradigm for Polling. *Harvard Data Science Review*, *5*(3). https://doi.org/10.1162/99608f92.9898eede

Bowler, D. E., Callaghan, C. T., Bhandari, N., Henle, K., Barth, M. B., Koppitz, C., Klenke, R., Winter, M., Jansen, F., Bruelheide, H., & Bonn, A. (2022). Temporal trends in the spatial bias of species occurrence records. *Ecography*. https://doi.org/10.1111/ecog.06219

Boyd, R. J., Botham, M., Dennis, E., Fox, R., Harrower, C., Middlebrook, I., Roy, D. B., & Pescott, O. L. (2025). Using causal diagrams and superpopulation models to correct geographic biases in biodiversity monitoring data. *Methods in Ecology and Evolution*, *16*(2), 332–344. https://doi.org/10.1111/2041-210X.14492

Boyd, R. J., Harvey, M., Roy, D., Barber, T., Haysom, K., & ... (2023). Causal inference and large-scale expert validation shed light on the drivers of SDM accuracy and variance. *Diversity and Distributions*, *28*, 774–784. https://doi.org/10.1111/ddi.13698

Boyd, R. J., Powney, G. D., Burns, F., Danet, A., Duchenne, F., Grainger, M. J., Jarvis, S. G., Martin, G., Nilsen, E. B., Porcher, E., Stewart, G. B., Wilson, O. J., & Pescott, O. L. (2022). ROBITT: A tool for assessing the risk-of-bias in studies of temporal trends in ecology. *Methods in Ecology and Evolution*, *13*(7), 1497– 1507. https://doi.org/10.1111/2041-210X.13857

Boyd, R. J., Powney, G. D., & Pescott, O. L. (2023). We need to talk about nonprobability samples. *Trends in Ecology & Evolution*, *38*(6), 521–531. https://doi.org/10.1016/j.tree.2023.01.001

Boyd, R. J., Stewart, G. B., & Pescott, O. L. (2024). Descriptive Inference using large, unrepresentative nonprobability samples: An introduction for ecologists. *Ecology*, *105*(2), e4214.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Wadsworth inc.

Brereton, T. M., Cruickshanks, K. L., Risely, K., Noble, D. G., & Roy, D. B. (2011). Developing and launching a wider countryside butterfly survey across the United Kingdom. *Journal of Insect Conservation*, *15*(1), 279–290. https://doi.org/10.1007/s10841-010-9345-8

Broughton, R. K., & Pocock, M. J. O. (2022). *The opportunities for semi-structured and effort recording to enhance the value of biological recording by volunteers*. https://jncc.gov.uk/about-jncc/corporate-information/evidence-quality-assurance/

Convention on Biological Diversity. (n.d.). *2050 Goals*. Retrieved October 20, 2024, from https://www.cbd.int/gbf/goals

Delgado, R. (2022). Detecting target species: With how many samples? *Royal Society Open Science*, *9*(8). https://doi.org/10.1098/rsos.220046

Di Cecco, G. J., Barve, V., Belitz, M. W., Stucky, B. J., Guralnick, R. P., & Hurlbert, A. H. (2021). Observing the Observers: How Participants Contribute Data to iNaturalist and Implications for Biodiversity Science. *BioScience*, *71*(11), 1179–1188. https://doi.org/10.1093/biosci/biab093

Dorazio, R. M. (2014). Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. *Global Ecology and Biogeography*, *23*(12), 1472–1484. https://doi.org/10.1111/geb.12216

Elliott, M. R., & Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, *32*(2), 249–264. https://doi.org/10.1214/16-STS598

639 Emmet, R. L., Benson, T. J., Allen, M. L., & Stodola, K. W. (2023). Integrating multiple data sources
640 improves prediction and inference for upland game bird occupancy models. *Ornithological*
641 *Applications*, *125*(2). https://doi.org/10.1093/ornithapp/duad005

642 Farr, M. T., Green, D. S., Holekamp, K. E., & Zipkin, E. F. (2021). Integrating distance sampling and
643 presence-only data to estimate species abundance. *Ecology*, *102*(1).
644 https://doi.org/10.1002/ecy.3204

645 Fink, D., Johnston, A., Auer, M. T., Hochachka, W. M., Ligocki, S., Oldham, L., Robinson, O.,
646 Wood, C., Kelling, S., Rodewald, A. D., & Fink, D. (2023). A Double machine learning trend
647 model for citizen science data. *Methods in Ecology and Evolution*, *2023*(June), 1–14.
648 https://doi.org/10.1111/2041-210X.14186

649 Fithian, W., Elith, J., Hastie, T., & Keith, D. A. (2015). Bias correction in species distribution models:
650 Pooling survey and collection data for multiple species. *Methods in Ecology and Evolution*, *6*(4),
651 424–438. https://doi.org/10.1111/2041-210X.12242

652 Hartman, E., & Huang, M. (2024). Sensitivity Analysis for Survey Weights. *Political Analysis*, *32*(1),
653 1–16. https://doi.org/10.1017/pan.2023.12

654 Hertzog, L. R., Frank, C., Klimek, S., Röder, N., Böhner, H. G. S., & Kamp, J. (2021). Model-based
655 integration of citizen science data from disparate sources increases the precision of bird
656 population trends. *Diversity and Distributions*, *27*(6), 1106–1119.
657 https://doi.org/10.1111/ddi.13259

658 Irvine, K. M., Rodhouse, T. J., Wright, W. J., & Olsen, A. R. (2018). Occupancy modeling species–
659 environment relationships with non-ignorable survey designs. *Ecological Applications*, *28*(6),
660 1616–1625. https://doi.org/10.1002/eap.1754

661 Isaac, N. J. B., Jarzyna, M. A., Keil, P., Dambly, L. I., Boersch-Supan, P. H., Browning, E., Freeman,
662 S. N., Golding, N., Guillera-Arroita, G., Henrys, P. A., Jarvis, S., Lahoz-Monfort, J., Pagel, J.,
663 Pescott, O. L., Schmucki, R., Simmonds, E. G., & O'Hara, R. B. (2020). Data Integration for
664 Large-Scale Models of Species Distributions. *Trends in Ecology and Evolution*, *35*(1), 56–67.
665 https://doi.org/10.1016/j.tree.2019.08.006

666 Johnston, A., Moran, N., Musgrove, A., Fink, D., & Baillie, S. R. (2020). Estimating species
667 distributions from spatially biased citizen science data. *Ecological Modelling*, *422*(December
668 2019), 108927. https://doi.org/10.1016/j.ecolmodel.2019.108927

669 Knaus, P., Antoniazza, S., Wechsler, S., Guélat, K., Kéry, M., Strebel, N., & Sattler, T. (2018).
670 *Distribution and population trends of birds in Switzerland and Liechtenstein*.

671 Leung, B., & Gonzalez, A. (2024). Global monitoring for biodiversity: Uncertainty, risk, and power
672 analyses to support trend change detection. *Science Advances*, *10*, 1448. https://www.science.org

673 Little, R. J. A., West, B. T., Boonstra, P. S., & Hu, J. (2020). Measures of the Degree of Departure
674 from Ignorable Sample Selection. In *Journal of Survey Statistics and Methodology* (Vol. 8,
675 Number 5, pp. 932–964). American Association for Public Opinion Research.
676 https://doi.org/10.1093/jssam/smz023

677 Lohr, S. (2022). *Sampling: Design and analysis* (3rd ed.). CRC Press.

678 Mäkinen, J., Merow, C., & Jetz, W. (2023). *Integrated species distribution models to account for*
679 *sampling biases and improve range-wide occurrence predictions*. (October), 1–15.
680 https://doi.org/10.1111/geb.13792

Mancini, F., Boersch-Supan, P. H., Robinson, R. A., Harris, M., & Pocock, M. J. O. (2022). *An introduction to model-based data integration for biodiversity assessments*.

Manski, C. (2007). *Identification for prediction and decision* (1st ed., Vol. 1). Cambridge, Massachussets.

Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 us presidential election. *Annals of Applied Statistics*, *12*(2), 685–726. https://doi.org/10.1214/18-AOAS1161SF

Meng, X.-L. (2022). Comments on the Wu (2022) paper by Xiao-Li Meng: Miniaturizing data defect correlation: A versatile strategy for handling non-probability samples. *Survey Methodology*, *48*(2). http://www.statcan.gc.ca/pub/12-001-x/2022002/article/00006-eng.htm

Miller, D. A. W., Pacifici, K., Sanderlin, J. S., & Reich, B. J. (2019). The recent past and promising future for data integration methods to estimate species' distributions. *Methods in Ecology and Evolution*, *10*(1), 22–37. https://doi.org/10.1111/2041-210X.13110

Mostert, P. S., & O'Hara, R. B. (2023). PointedSDMs: An R package to help facilitate the construction of integrated species distribution models. *Methods in Ecology and Evolution*, *14*(5), 1200–1207. https://doi.org/10.1111/2041-210X.14091

O'Connor, R. S., Kunin, W. E., Garratt, M. P. D., Potts, S. G., Roy, H. E., Andrews, C., Jones, C. M., Peyton, J. M., Savage, J., Harvey, M. C., Morris, R. K. A., Roberts, S. P. M., Wright, I., Vanbergen, A. J., & Carvell, C. (2019). Monitoring insect pollinators and flower visitation: The effectiveness and feasibility of different survey methods. *Methods in Ecology and Evolution*, *10*(12), 2129–2140. https://doi.org/10.1111/2041-210X.13292

Pacifici, K., Reich, B. J., Miller, D. A. W., Gardner, B., Stauffer, G., Singh, S., McKerrow, A., & Collazo, J. A. (2017). Integrating multiple data sources in species distribution modeling: A framework for data fusion. *Ecology*, *98*(3), 840–850. https://doi.org/10.1002/ecy.1710

Pagel, J., Anderson, B. J., O'Hara, R. B., Cramer, W., Fox, R., Jeltsch, F., Roy, D. B., Thomas, C. D., & Schurr, F. M. (2014). Quantifying range-wide variation in population trends from local abundance surveys and widespread opportunistic occurrence records. *Methods in Ecology and Evolution*, *5*(8), 751–760. https://doi.org/10.1111/2041-210X.12221

Pavlacky, D. C., Lukacs, P. M., Blakesley, J. A., Skorkowsky, R. C., Klute, D. S., Hahn, B. A., Dreitz, V. J., George, T. L., & Hanni, D. J. (2017). A statistically rigorous sampling design to integrate avian monitoring and management within Bird Conservation Regions. *PLoS ONE*, *12*(10). https://doi.org/10.1371/journal.pone.0185924

Pescott, O. L., Walker, K. J., Harris, F., New, H., Cheffings, C. M., Newton, N., Jitlal, M., Redhead, J., Smart, S. M., & Roy, D. B. (2019). The design, launch and assessment of a new volunteer-based plant monitoring scheme for the United Kingdom. *PLoS ONE*, *14*(4), 1–30. https://doi.org/10.1371/journal.pone.0215891

Potts, S. G., DAUBER, J., HOCHKIRCH, A., OTEMAN, B., Roy, D. B., AHNRE, K., BIESMEIJER, K., BREEZE, T., CARVELL, C., FERREIRA, C., FITZPATRICK, Ú., Isaac, N. J. B., KUUSSAARI, M., LJUBOMIROV, T., MAES, J., NGO, H., PARDO, A., POLCE, C., Marino, Q., … VUJIC, A. (2021). *Proposal for an EU Pollinator Monitoring Scheme*. https://doi.org/10.2760/881843

Rao, J. N. K. (2021). On Making Valid Inferences by Integrating Data from Surveys and Other Sources. *Sankhya B*, *83*(1934), 242–272. https://doi.org/https://doi.org/10.1007/s13571-020-00227-w

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. https://doi.org/https://doi.org/10.1093/biomet/63.3.581

Schuessler, J., & Selb, P. (2025). Graphical Causal Models for Survey Inference. *Sociological Methods and Research*, *54*(1), 74–105. https://doi.org/10.1177/00491241231176851

Simmonds, E. G., Jarvis, S. G., Henrys, P. A., Isaac, N. J. B., & Hara, R. B. O. (2020). Is more data always better? A simulation study of benefits and limitations of integrated distribution models. *Ecography*, *43*(10), 1413–1422. https://doi.org/10.1111/ecog.05146

Strebel, N., Kéry, M., Guélat, J., & Sattler, T. (2022). Spatiotemporal modelling of abundance from multiple data sources in an integrated spatial distribution model. *Journal of Biogeography*, *49*(3), 563–575. https://doi.org/10.1111/jbi.14335

Sugden, R. A., & Smith, T. M. F. (1984). *Ignorable and Informative Designs in Survey Sampling Inference* (Vol. 71, Number 3).

Sullivan, B. L., Aycrigg, J. L., Barry, J. H., Bonney, R. E., Bruns, N., Cooper, C. B., Damoulas, T., Dhondt, A. A., Dietterich, T., Farnsworth, A., Fink, D., Fitzpatrick, J. W., Fredericks, T., Gerbracht, J., Gomes, C., Hochachka, W. M., Iliff, M. J., Lagoze, C., La Sorte, F. A., … Kelling, S. (2014). The eBird enterprise: An integrated approach to development and application of citizen science. *Biological Conservation*, *169*, 31–40. https://doi.org/10.1016/j.biocon.2013.11.003

UK Pollinator Monitoring Scheme. (2026). *The UK PoMS Annual report 2025.* .

Valdez, J. W., Callaghan, C. T., Junker, J., Purvis, A., Hill, S. L. L., & Pereira, H. M. (2023). The undetectability of global biodiversity trends using local species richness. *Ecography*, *2023*(3). https://doi.org/10.1111/ecog.06604

Van Swaay, C. A. M., Plate, C. L., & Van Strien, A. J. (2002). Monitoring butterflies in the Netherlands: how to get unbiased indices. *Proceedings of the Section Experimental and Applied Entomology of the Netherlands Entomological Society*, *13*, 21–27.

Whetten, A. B., Diaz, R. M., Hochachka, W. M., Stillman, A., Davis, C. L., Auer, T., Strimas-Mackey, M., Ruiz-Gutierrez, V., Howell, P. E., Kershner, E., & Fink, D. (2026). R-Learner Data Integration for Estimating Population Trends From Opportunistic and Structured Survey Data. *Environmetrics*, *37*(6). https://doi.org/10.1002/env.70116

Wiśniowski, A., Sakshaug, J. W., Perez Ruiz, D. A., & Blom, A. G. (2020). Integrating probability and nonprobability samples for survey inference. In *Journal of Survey Statistics and Methodology* (Vol. 8, Number 1, pp. 120–147). American Association for Public Opinion Research. https://doi.org/10.1093/jssam/smz051

Zipkin, E. F., Rossman, S., Yackulic, C. B., Wiens, J. D., Thorson, J. T., Davis, R. J., & Grant, E. H. C. (2017). Integrating count and detection–nondetection data to model population dynamics. *Ecology*, *98*(6), 1640–1650. https://doi.org/10.1002/ecy.1831