Descriptive inference using large, unrepresentative nonprobability samples: An introduction for ecologists

[*1]Robin J. Boyd, [2]Gavin B. Stewart and [1]Oliver L. Pescott

[1]UK Centre for Ecology & Hydrology, Benson Lane, Wallingford, OX108BB

[2] Evidence Synthesis Lab, School of Natural and Environmental Science, University of Newcastle, Newcastle-upon-Tyne, NE1 7RU

*corresponding author email: robboy@ceh.ac.uk

Abstract

Biodiversity monitoring usually involves drawing inferences about some variable of interest across a defined landscape from observations made at a sample of locations within that landscape. If the variable of interest differs between sampled and non-sampled locations, and no mitigating action is taken, then the sample is unrepresentative and inferences drawn from it will be biased. It is possible to adjust unrepresentative samples so that they more closely resemble the wider landscape in terms of "auxiliary variables". A good auxiliary variable is a common cause of sample inclusion and the variable of interest, and if it explains an appreciable portion of the variance in both, then inferences drawn from the adjusted sample will be closer to the truth. We applied six types of survey sample adjustment—subsampling, quasi-randomisation, poststratification, superpopulation modelling, a "doubly robust" procedure, and multilevel regression and poststratification—to a simple two-part biodiversity monitoring problem. The first part was to estimate mean occupancy of the plant *Calluna vulgaris* in Great Britain in two time-periods (1987-1999 and 2010-2019); the second was to estimate the difference between the two (i.e. the trend). We estimated the means and trend using large, but (originally) unrepresentative, samples from a citizen science dataset. Compared to the unadjusted estimates, the means and trends estimated using most adjustment methods were more accurate, although standard uncertainty intervals generally did not cover the true values. Completely unbiased inference is not possible from an unrepresentative sample without knowing and having data on all relevant auxiliary variables. Adjustments can reduce the bias if auxiliary variables are available and selected carefully, but the potential for residual bias should be acknowledged and reported.

Introduction

As the data revolution gathers pace, it is not surprising to see "big data" being used to monitor biodiversity. Examples include observations submitted to mobile phone apps by amateur naturalists (Johnston et al., 2022) and digitised specimens from museums and herbaria (Nelson & Ellis, 2019). Such data become bigger still when combined in data aggregators such as the Global Biodiversity Information Facility (GBIF; https://www.gbif.org/) or metadatabases such as PREDICTS (Hudson et al., 2014). Unfortunately, quantity of data does not necessarily imply quality of insight.

Monitoring biodiversity is typically a matter of descriptive statistical inference. It is inferential in that the goal is to infer something about a target population from a sample of that population (Boyd, Powney, et al., 2023). The population might comprise, say, all areal units across some landscape ("sites"), in which case the sample would be a subset of those sites. The inference is descriptive in that the aim is to describe (rather than explain) a variable of interest in the population. A common example is the proportion of sites occupied by some species (Bowler et al., 2021; Outhwaite et al., 2020; Powney et al., 2019; Stroh et al., 2023; van Strien & van Grunsven, 2023), but there are many others.

Of more importance than the size of a sample for descriptive inference is whether it is representative of the population about which inferences are to be drawn (Meng, 2018). In a representative sample, the distribution of the variable of interest is similar to its distribution in the population (Bethlehem et al., 2008). An equivalent definition is that there is little to no correlation between inclusion in the sample and the variable of interest—the "data defect correlation", or ddc (Meng, 2018). Intuitively, statistics derived from a representative sample, such as means and proportions, will be similar to their population equivalents.

Unfortunately, ddc's are likely to be appreciable in big biodiversity datasets. For one, naturalists preferentially visit and collect data at sites where they are likely to see species that interest them (Bowler et al., 2022; Forister et al., 2023). Where those species' abundances or distributions are the variables of analytic interest, preferential sampling naturally results in a positive ddc (McClure & Rolek, 2023). On the other hand, naturalists might be constrained to visiting and collecting data in, say, built up areas, which are easier to access than remote locations (Geldmann et al., 2016; Hughes et al., 2020; Mandeville et al., 2022). Built-up areas generally have low quality habitat, meaning that species are less likely to occupy them in large numbers and that the ddc might be negative.

Inferences from unrepresentative samples, with appreciable ddc's, are likely to be misleading. Imagine a researcher who wants to estimate the average abundance of some species across a landscape. An obvious (but naïve) approach would be to calculate its mean abundance across sampled sites and assume that this is similar to its average abundance across the wider landscape. However, if the locations at which the species is most abundant were preferentially sampled, then the sample-based estimate of its mean abundance will be upwardly biased. To use the analogy of Forister et al. (2023), sampled locations would be life rafts; non-sampled locations would be the sinking ship.

It is simple to counteract the biasing effect of the ddc if the probability that each site was included in the sample is known; that is, if a *probability sample* is available. In this case, more weight can be placed on the data from sites that were less likely to be included. The effect of this type of weighting is easiest to explain heuristically: the sample is augmented with "copies" of the data from sites that were less likely to be sampled, effectively bringing sample inclusion probabilities across sites to parity. Two variables cannot be correlated if one of them is constant, which means there can be no correlation between the weighted sample inclusion probabilities and the variable of interest across sites. It follows that the ddc, which is the correlation between actual (weighted) sample inclusion and the variable of interest, is zero in expectation (Meng, 2022), and the sample can be considered representative (Lohr, 2022). Weighting of this type is known as "design-based" inference, because the inclusion probabilities are a feature of the sampling design.

Design-based inference is not applicable for the types of big biodiversity datasets we consider here, because they were not collected according to a probabilistic sampling design. We do not know the probabilities that sites were visited by the collectors of specimens now held in museums and on GBIF. Nor do we know the probabilities that citizen scientists visited and collected data at each site across most landscapes. Matters are simpler when using data from structured monitoring schemes, which often aim for a probability sample (e.g PoMS). However, incomplete uptake of sites that were selected for inclusion (Pescott et al., 2015, 2019) means that, in practice, these samples too are non-probabilistic. [Incomplete uptake in biodiversity monitoring is analogous to the issue of non-response in survey sampling (e.g. Bethlehem et al., 2008).] Where sample inclusion probabilities are not known, an alternative to design-based inference is needed.

Most approaches to inference from nonprobability samples involve *estimating* the inclusion probabilities. A relatively simple example is poststratification, where the observations (for each site) are split into strata based on covariates, and sites in strata that are underrepresented in the population (based on the population totals of the covariates) are given more weight (Valliant et al., 2018). Using covariates to estimate sample inclusion probabilities is equivalent to adjusting the samples in such a way that the distributions of those covariates in the sample more closely resemble their distributions in the population (i.e. across all sites in the wider landscape). If the covariates affect *both* the variable of interest and sample inclusion, then inferences drawn from the adjusted sample will be closer to the truth than those from the original (naïve) sample. In the context of inference from nonprobability samples, covariates affecting both sample inclusion and the variable of interest, which are not of direct analytic interest themselves, are known as "auxiliary variables" (Thoemmes & Mohan, 2015; Thoemmes & Rose, 2014).

Before going further, it is important to note that most approaches to inference from nonprobability samples rest on the bold assumption that the variable of interest is independent of sample inclusion after accounting for the auxiliary variables (Bailey, 2022); that is, non-sampled sites are "Missing At Random" (MAR; Rubin, 1976). If the MAR assumption holds, then unbiased inference is possible. In reality, the MAR assumption is likely to be violated, because data are not available on all relevant auxiliary variables, so the best we can hope for is a reduction in bias relative to naïve inferences drawn from the unadjusted sample.

Use of sample adjustments in biodiversity monitoring is variable. It is common for monitoring schemes to weight samples in such a way that the relative frequencies of habitats or geographic areas in the sample are similar to those in the population (Gregory et al., 2005; Van Swaay et al., 2002, 2008; Weiser et al., 2020). But it is also common to see samples treated as though they are representative despite clear evidence to the contrary. For example, Vellend et al. (2013) and Dornelas et al. (2014) purported to document globally representative time trends in species richness, but Gonzalez et al. (2016) showed that their samples were highly unrepresentative with respect drivers of biodiversity change and species richness itself. (See Boyd, Powney, et al. (2023) for a review of this debate and others like it.) We suspect that many of those who do not deal with issues of sample representativeness are not familiar with the gravity of the problem or the relevant theory and adjustment methods.

In this paper, we introduce six approaches to descriptive inference using unrepresentative nonprobability samples and demonstrate how they relate to each other (conceptually and mathematically). We apply each approach to a simple two-part biodiversity monitoring problem. The first part is to estimate mean occupancy of the plant *C. vulgaris* across 1 km grid squares in Britain in two time-periods; the second is to estimate the difference between the two (i.e. the time trend). *Calluna vulgaris* is an attractive case study because we have good estimates of its true geographic distribution in both periods from satellite (amongst other sources). The approaches to inference that we demonstrate are subsampling, quasi-randomisation (Elliott and Valliant, 2017), poststratification (Little, 1993), superpopulation modelling (Valliant, 2009), a "doubly robust" estimator (Chen et al., 2020), and Multilevel Regression and Poststratification (MRP; Gelman, 2007; Gelman and Little, 1997). Each can be (MRP more loosely than the rest) interpreted as an attempt to weight the sample in such a way that it more closely resembles the population, in the hope that this results in more accurate descriptive inferences. We demonstrate the effects of each approach on the distributions of auxiliary variables in the sample, as well as on the resulting estimates of mean occupancy in each period and the time trend between the two. Applying the adjustment

methods to a real-world example reveals challenges that ecologists are likely to face, and we discuss these in detail.

Methods

True distribution of *Calluna vulgaris*

We approximated the true distribution of the dwarf shrub vascular plant *Calluna vulgaris* (Heather) in two time periods: 1987–1999 and 2010–2019. For the first period, we used the 1990 UKCEH land cover map (Rowland et al., 2020); for the second, we used the 2018 version (Morton et al., 2022). The land cover maps are derived from satellite, which means that they provide information for every 1 km grid square. From these maps, we identified 1 km grid squares (British National Grid, EPSG:27700) with >0% heather or heather grassland cover. To these, we added 1 km squares in which *C. vulgaris* was recorded in each time period by the Botanical Society of Britain and Ireland (BSBI). The time periods considered cover the main periods of recording for two national distribution atlases, which involved a concerted effort by volunteers (citizen scientists) to document vascular plants across the United Kingdom (Preston, C.D., Pearman, D.A. & Dines, 2002; Stroh et al., 2023). Acknowledging that some 1 km squares may have been erroneously classed as having some heather or heather grassland coverage by the land cover maps, we removed any 1 km squares that fell within 10 km grid squares in which *C. vulgaris* had not been recorded by the BSBI in the period 1950–2019. Given that this period includes recording for three national distribution atlases (the two cited above plus Perring & Walters, 1962), we assume that the union of all 10 km occurrences within this period encompasses all known populations irrespective of finer scale changes. Figure 1 maps the resulting estimates of the true 1 km distributions of *C. vulgaris* in both time-periods.

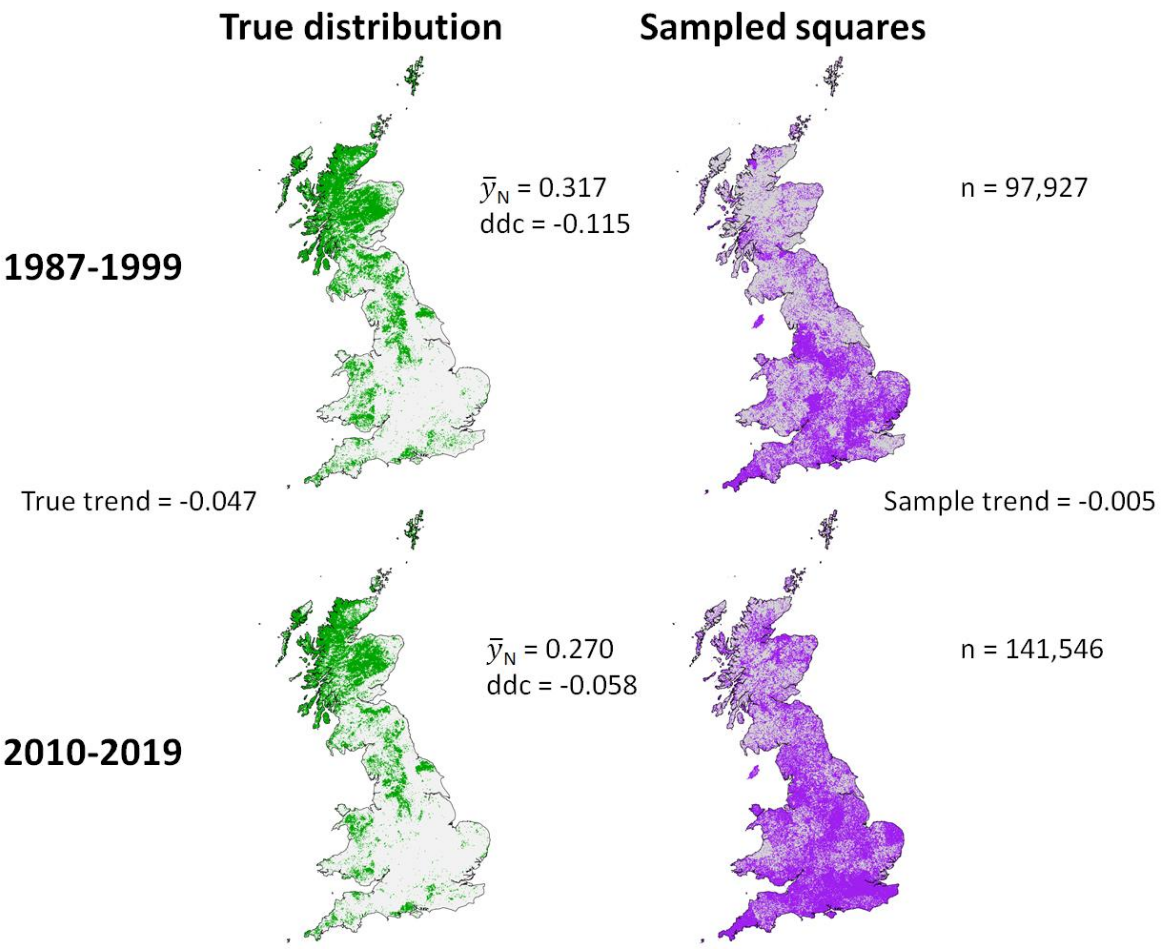

Figure 1. Left column: the distribution of *Calluna vulgaris* in both time-periods. Green squares are occupied and grey squares are not. $\bar{y}_N$ is mean occupancy or, equivalently, the proportion of squares occupied. The ddc's are the correlations between sample inclusion (1 if the square is in the sample and 0 otherwise) and occupancy. Right column: the nonprobability 1 km samples for each time-period. Purple squares were sampled and grey squares were not. n is the number of squares sampled. We assume that *C. vulgaris* was recorded in all sampled grid squares that it occupied in the relevant time-period. The true trend is the difference between population means, and the sample trend is the difference between sample means (i.e. mean occupancy across purple squares).

Sample data on *Calluna vulgaris* occupancy

The 1 km samples for both time periods ("Sampled squares in Fig. 1) encompass any vascular plant data for which the date of collection is known (i.e. the record is resolved to the day), either at the 1 km scale or finer, collected by the BSBI for the national distribution atlases of Preston et al. (2002) and Stroh et al. (2023). Having been collected by volunteers, the data come under the banner of citizen science.

Auxiliary data

We used two auxiliary variables for which data are available for all 1 km grid squares in Great Britain: the proportion of each 1 km grid square that falls within some form of protected area (including everything from SSSI's to local nature reserves; UNEP-WCMC & IUCN, 2020) and the average elevation of each 1 km square (Intermap, 2009). New protected areas are designated periodically, so we used the set that were designated prior to 1987 for the first time-period and prior to 2010 for the second (i.e. the beginning of each period). We

suspect that 1 km squares with more protected area coverage are more likely to be visited by naturalists (Girardello et al., 2019) and, because protected areas tend to have higher quality habitat, are also more likely to be occupied by *C. vulgaris*. Likewise, elevation should affect both sample inclusion and *C. vulgaris* occupancy. Sites at higher elevations are harder to access on account of their relatively harsh terrain and remoteness, and elevation is a known predictor of *C. vulgaris* occupancy (Stroh et al., 2023).

One of the adjustment methods that we describe below, quasi-randomisation, requires additional covariates (we use the term "covariate" to distinguish these from the auxiliary variables as defined earlier). The method involves the estimation of sample inclusion probabilities for every 1 km grid square in Britain. This is a matter of prediction rather than inference, because we know whether each 1 km square was sampled (i.e. there is no missing data), so it was sensible to use a wider range of covariates. See Table 1 in Boyd, Stewart, et al. (2023) for a list of the additional covariates used in this model.

Estimating the per-period population mean

The first step in our biodiversity monitoring problem is to estimate mean occupancy of *C. vulgaris* in each time-period. Although not usually written this way, it is helpful for what comes later to re-express the population mean as a weighted sum

$$\bar{y}_N = \frac{1}{N}\sum_{i=1}^{N} y_i = \sum_{i=1}^{N} \frac{y_i}{N} = \sum_{i=1}^{N} \frac{y_i\, w_i}{\sum_N w_i}, \qquad \text{equation 1}$$

where y is occupancy (1 = occupied and 0 = unoccupied), N is the population size, i indexes 1 km grid squares and $w_i = 1/N$ (N is the same in both time-periods). The denominator in the rightmost expression might seem unnecessary, because it equals one. We have retained it to illustrate the similarity between this expression and the sample-based estimators below, which have a similar form but whose sampling weights w do not necessarily sum to one. (We use the term "estimator" to describe a rule for estimating some quantity from a sample; here, that quantity is the population mean.) For notational simplicity, we do not index the time-period, and the reader should remember that $\bar{y}_N$ is time-period specific. In practice, y is not known for all i in the population, so sample-based estimators of $\bar{y}_N$ are needed.

The design-based estimator

The design-based estimator of the population mean, which is applicable only where a probability sample of some sort is available (Lohr, 2022), has a similar form to eq. 1

$$\overline{y}_{db} = \sum_{i=1}^{n} \frac{y_i\, w_i}{\sum_n w_i}. \qquad \text{equation 2}$$

The differences are that the sums are over the sample size n rather than N and that the weights w_i are not necessarily constant. Rather, the weight for unit i, w_i, is equal to the reciprocal of the probability that it was included in the sample $= 1/p_i$.

Sample inclusion probabilities are, by definition, not known for nonprobability samples, so alternative estimators are required. We present six such estimators below, three of which–quasi-randomisation, poststratification and superpopulation modelling–are explicit attempts to come up with a set of weights w_i that produce a reasonable estimate of $\bar{y}_N$ under eq. 2. The other three—a "doubly robust" estimator, subsampling and MRP—are not, but they are conceptually similar.

Estimators for nonprobability samples

The following estimators are used in survey sampling to estimate population means from nonprobability samples. More detail on each can be found in Valliant et al. (2018), Lumley (2010) and Lohr (2022). See supplementary material 1 for an R Markdown document containing the code to implement each of the adjustment methods.

Naïve sample mean

Where sample inclusion probabilities are unavailable, a simple option is to assume that $w_i = 1/n$ for all i. In this case, eq. 2 gives the (naïve) sample mean. As the weights are constant, the sample mean does not adjust for differences in y between the sampled and non-sampled population units. It is nevertheless widely used in biodiversity monitoring.

Quasi-randomisation

An alternative approach is to imagine that the nonprobability sample was selected probabilistically and to estimate the implied inclusion probabilities. Any binary model and covariates can be used. Once inclusion probabilities p_i have been estimated, the weights $w_i = 1/p_i$ (as in the design-based estimator). In our example, we used random forests and several covariates (including the auxiliaries) to estimate pseudo-inclusion probabilities. More complex appraoches are possible and have been used to map species distributions (Johnston et al., 2020).

Poststratification

Another approach to estimating sampling weights is poststratification. Poststratification requires categorical auxiliary data, so continuous variables must be discretized prior to analysis (Valliant, 2020). The auxiliary variables are crossed (think contingency tables) to create poststrata. Each poststratum j has a sample size n_j and population size N_j. The sampling weight w_i for population unit i in poststratum j is given by N_j/n_j.

In our example, we split elevation into ten categories using its deciles (i.e. cut points at the 10th and 20th percentiles, etc.). This did not make sense for the variables denoting the proportion of each grid square that falls within a protected area, because most squares took the value one or zero. We split this variables into two categories, 0 and >0, i.e. whether or not there is some protected land in the grid square. Discretization gave $10 \times 2 = 20$ poststrata.

It is sensible to discretize the auxiliary variables in such a way that the variable of interest varies among categories. Otherwise, the adjustment from poststratifying will be minor (or unnecessary!). Fig. 2 shows that mean occupancy of *C. vulgaris* in the samples differs appreciably among levels of the auxiliary variables.

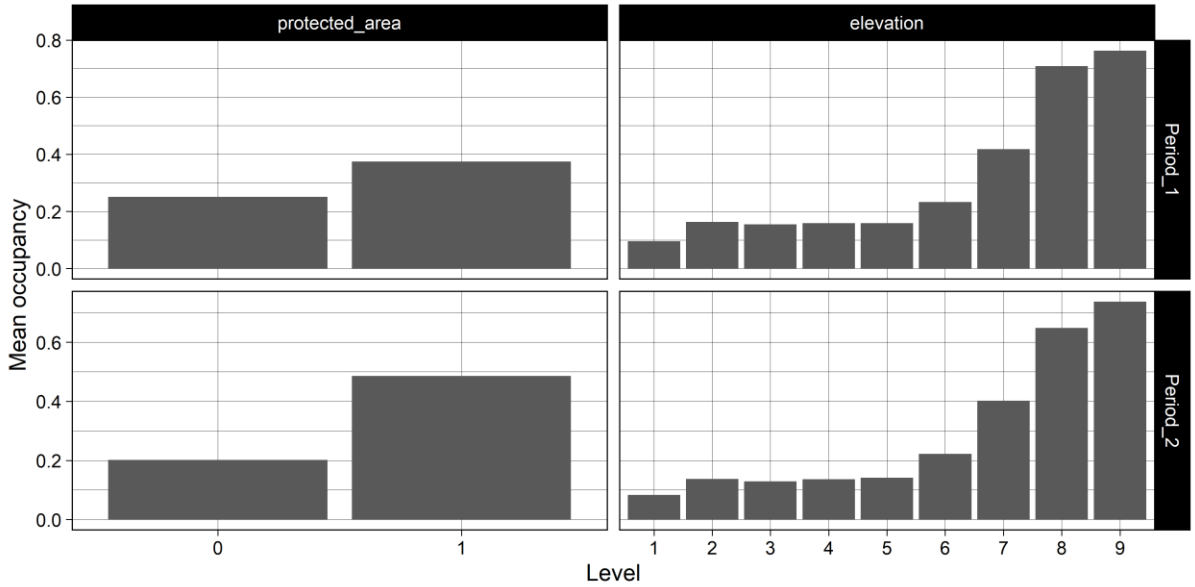

Figure 2. Mean occupancy of *Calluna vulgaris* for each level of the auxiliary variables in each time-period. The auxiliary variables were originally on a continuous scale, but we discretized them to enable poststratification. See the main text for details.

Superpopulation modelling

Superpopulation modelling is conceptually different to the adjustment methods described above. The premise is that there exists some model that describes the variable of interest in the population. If this model can be recovered from the sampled outcome variable y and the auxiliary data, it can be used to predict the variable of interest in non-sampled units. Given a prediction for each non-sampled i, it is then simple to estimate the population mean.

A general (i.e. multiple) linear regression model of y has the form

$$E_M(y_i) = \boldsymbol{x}_i^T \, \beta, \qquad \text{equation 3}$$

where the subscript M indicates that the expectation (mean) is with respect to the model, $\boldsymbol{x}_i$ is the vector of auxiliary variables for unit i, the superscript T indicates that the vector $\boldsymbol{x}_i$ has been transposed (to a row vector) and β is a column vector of parameters. A prediction of y for unit i is

$$\hat{y}_i = x_i^T \, \hat{\beta}. \qquad \text{equation 4}$$

The accent on β indicates that it is an estimate (the least squares estimate in this case). If $\bar{s}$ is the set of non-sampled population units, the superpopulation model prediction of the population mean is

$$\bar{y}_{sp} = \frac{\sum_{i \in s} y + \sum_{i \in \bar{s}} \hat{y}}{N}. \qquad \text{equation 5}$$

That is, it is the sum of the known outcome values in the sample and those predicted by the model for the remainder of the population divided by the population total.

A feature of $\bar{y}_{sp}$ is that it can be expressed in the same form as the design-based estimator in eq. 2, with the weights w_i being a function of the auxiliary variables in sampled and non-sampled population units (Elliott & Valliant, 2017). (Code to verify this numerically is available at https://github.com/robboyd/selectionBiasEffects/tree/master/R.) Like the other

adjustment models, then, the superpopulation estimator is an approach to estimating the sampling weights w_i.

Linear regression might seem like an unusual choice of model for a binary outcome (occupancy), but we felt that it was the best option here. One reason is that the implied model is actually linear for an estimator with the form of eq. 2 (Valliant, 2020). Most important, however, is that the use of a linear model enables the estimation of sampling weights (Valliant et al., 2018; supplementary material 1; https://github.com/robboyd/biasAdjustments). This is helpful, because those weights can be used to show the effects of superpopulation modelling on the distributions of the auxiliary variables in the sample (see "Evaluating the effects of the adjustments" below). Alternative models can be used where weights are not required (e.g. Wu and Sitter, 2001). In our example superpopulation model, we used the auxiliary variables as predictors.

Doubly robust estimator

The doubly robust estimator combines the superpopulation model and the sample inclusion model from the quasi-randomisation procedure in such a way that if either is correct, and the sample size is large, then the estimate of the population mean unbiased (Valliant, 2020). It has the general form (Wu, 2022)

$$\bar{y}_{dr} = \frac{1}{N}\sum_{i \in s}\frac{r_i}{p_i} + \frac{1}{N}\sum_{i=1}^{N}\hat{y}_i, \qquad \text{equation 6}$$

where $r_i = y_i - \hat{y}_i$ (i.e. the residuals of superpopulation model). The second term on the right is the superpopulation model prediction of $\bar{Y}_N$. If the superpopulation model is correctly specified, then it is an unbiased estimate of $\bar{Y}_N$. However, if the superpopulation model is misspecified, then the second term needs to be corrected, which is where the first term comes in. If the quasi-randomisation sample inclusion model is correctly specified, the first term corrects the second by adding the residuals of the superpopulation model divided by the (correctly) estimated pseudo inclusion probabilities. This is sufficient to produce an unbiased estimate of $\bar{Y}_N$ even where the superpopulation model is wrong. Where the superpopulation model is correct, the first term is 0, because $r_i = 0$. Where neither model is correct, $\bar{y}_{dr}$ is a biased estimator of $\bar{Y}_N$. See Chen et al. (2020), who combined probability and nonprobability samples, for a similar approach.

Subsampling

Perhaps more familiar to ecologists than the above approaches is subsampling (Beck et al., 2014; Steen et al., 2020). The idea is to create a representative "miniature" of the population out of the sample (Meng, 2022) and to calculate the quantity of interest (mean occupancy) from this subsample. Subsampling trades sample size for representativeness.

Our approach was to draw stratified random samples of size $N/10 = 22{,}958$ with replacement from the original samples. We used the same strata as described above (under Poststratification). The decision to set $n = N/10$ was somewhat arbitrary, but changing the subsample size makes little difference to the point estimates of the population means (although they become more precise with increasing subsample size; supplementary material 1). The subsample mean is the estimator of the population mean.

Multilevel regression and poststratification (MRP)

MRP is an extension of poststratification and a variation of superpopulation modelling (Gelman, 2007; Gelman & Little, 1997; Valliant et al., 2018). A hierarchical model is used to estimate mean occupancy in each poststratum. The advantage of using a hierarchical model is that cells with few or no data borrow information from cells with more data (i.e. partial

pooling or shrinkage is exploited). The population mean is the weighted mean of the stratum means, where the weights are equal to the proportion of the population in each stratum.

Our hierarchical model is a binomial GLM with a logit link function, a fixed intercept and random intercepts for the auxiliary variables and their interaction (see https://mc-stan.org/rstanarm/articles/mrp.html for a similar formulation). We fitted the model in a Bayesian framework using 5 Markov Chain Monte Carlo (MCMC) chains, each with 1000 iterations. This was sufficient to achieve convergence on all parameters in both time-periods.

Confidence intervals

We present 95% confidence/credible intervals for all estimates of mean occupancy (credible intervals for MRP, which we implemented in a Bayesian framework). For most methods—superpopulation modelling, quasi-randomisation, subsampling and the doubly robust estimator—we constructed bootstrap confidence intervals. Resampling the original data with replacement, we created 1000 bootstrap samples, from which we obtained a distribution of estimates from each method and calculated percentile intervals. For MRP, we extracted credible intervals from the posterior distributions of mean occupancy. We used the confidence intervals provided by the *survey* package (Lumley, 2010) for the poststratified and naïve (i.e. unadjusted) estimates.

Estimating the trend in mean occupancy

Having estimated mean occupancy in each time-period, the next step was to estimate the difference between the two $= \bar{y}_2 - \bar{y}_1$ (i.e. the trend). We constructed a confidence interval for the trend estimated using each method in one of two ways depending on whether the method produced one estimate or a distribution. The methods that produced a distribution of $\bar{y}_2 - \bar{y}_1$ include those that we bootstrapped and MRP, which we fitted in a Bayesian framework (meaning we have a posterior distribution). For these methods, we extracted percentile confidence intervals (95%) from the distributions of estimated trends. For the others, poststratification and the naïve estimator (the sample mean), we used the normal approximation of the 95% confidence interval, given by $\pm$ 1.96 $\times$ the standard errors, where the standard errors are $\sqrt{var(\bar{y}_2) + var(\bar{y}_1)}$ (Gelman, 2007).

Evaluating the effects of the adjustments

We used relative frequency plots (Cf. Makela et al., 2014) to assess whether the adjustments brought the distributions of the auxiliary variables in the samples closer to their distributions in the population. The first step was to split each auxiliary variable into fifty bins of equal width spanning its range. The relative frequency of grid squares (the i's) in each bin k is $N_{i,k}/N$, where $N_{i,k}$ is the number of grid squares in each bin k in the population and N is the population size (we use k to index the bins to distinguish them from the strata described earlier). Similarly, the relative frequency of sampled grid squares in each k is $n_{i,k}/n$, where $n_{i,k}$ is the number of sampled grid squares in bin k and n is the total sample size. In the adjusted samples, the equivalent relative frequency is $\frac{\sum_{i \in k} w_i}{\sum_N w_i}$ (slightly different for subsampling; see below). We compared the original and adjusted samples' deviations from the population using the Mean Absolute Error (MAE) of the relative frequencies across all k. If the MAE from the adjusted sample is smaller than the original sample, then the adjustment brought the distribution of the auxiliary variable closer to its population distribution.

We were not able to produce adjusted relative frequency plots based on the doubly robust estimator or MRP. The problem was that could not estimate reasonable sampling weights from either method, which are needed to adjust the relative frequencies of the auxiliaries. Whilst it has been shown how to derive unit-level sampling weights where the MRP

multilevel model is linear (Gelman, 2007), no formula has yet been derived for the case of the binomial GLM (Valliant et al., 2018). As for the doubly robust estimator, Valliant (2020) showed how to derive "model-assisted" weights. Unfortunately, in our case, many of the model-assisted weights were very large and negative. The extreme weights appear to be caused by the pattern of residuals from the superpopulation model (recalling that we used a linear regression despite the fact that occupancy is binary), but it is beyond the scope of this paper to definitively diagnose the problem. There is no obvious way to derive weights from the subsampling estimator either. However, for this estimator, the adjusted relative frequencies of the auxiliaries are simply their distributions in the subsamples so are simple to obtain.

Assessing whether the estimates of mean occupancy in each period and the trend were improved by each adjustment method was simpler. We measured the difference between the point estimates of mean occupancy and the truth using the absolute error = $|\bar{y}_N - \bar{y}_{est}|$, where $\bar{y}_{est}$ is the estimate. For the trends, whose signs are of interest, we simply used the differences between the estimates and the truth. We also assessed whether the confidence/credible intervals produced by each method covered the true means and trend. We did not consider the power to detect the trend—that is, whether the methods' uncertainty intervals span zero at some percentile—because many biodiversity applications are descriptive-inferential rather than decision-theoretic.

Results

Per-period sample representativeness and estimated mean occupancy

The samples are large but somewhat unrepresentative (Fig. 1). Forty-three percent of grid squares were sampled in period one, and the ddc is -0.115; in period two, 62% of grid squares were sampled, and the ddc is -0.057. A consequence of these ddc's is that the naive sample means underestimate the population means, especially in period one where the magnitude of the ddc is greater (Fig. 3).

With the exception of the doubly robust estimate in period two, the estimates of mean occupancy from all adjustment methods in both time-periods had lower absolute errors than the naïve sample mean (Fig. 3; mean absolute errors are provided in supplementary material 2). The confidence intervals for the poststratified, subsample and quasi-randomisation estimates covered the true population mean in period one. In period two, no method's confidence/credible interval covered the population mean.

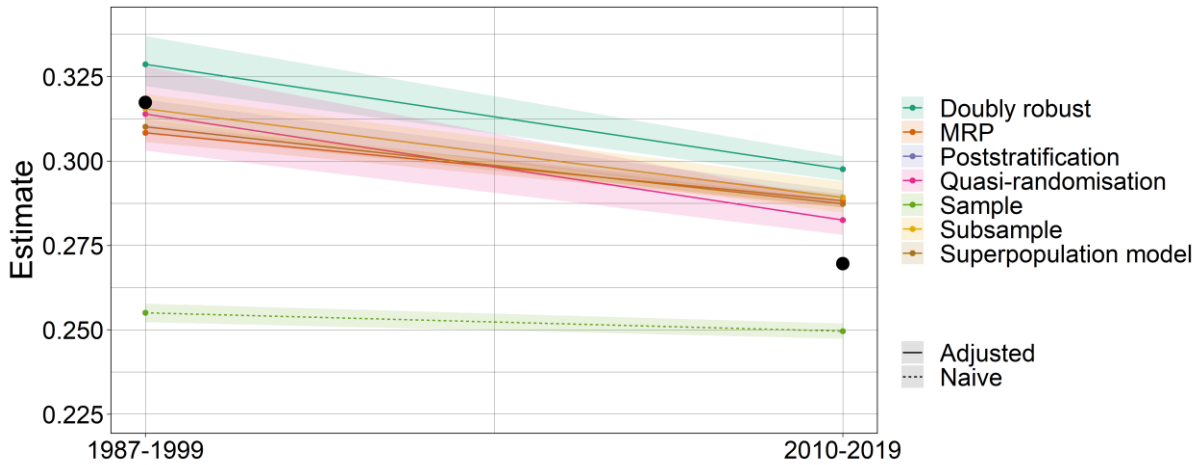

Figure 3. Naive (i.e. unadjusted) and adjusted sample-based estimates of mean occupancy in each time-period. The shaded regions are 95% confidence/credible intervals (see the main text for information on we constructed these for each method). The large black circles are the true population means in each time-period.

Estimated trend in mean occupancy

Estimates of the trend in mean occupancy from all adjustment methods were more accurate than the difference in sample means (i.e. the naïve estimate; Fig. 4). However, no method's point estimate came close to the true trend of -0.047, and their confidence/credible intervals did not cover it.

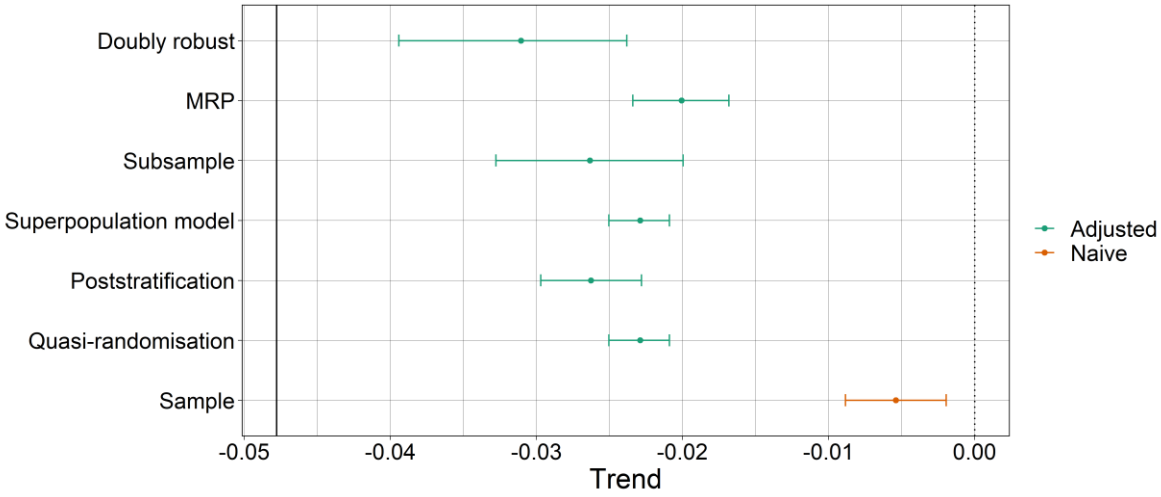

Figure 4. Trends in mean occupancy between periods one and two produced by the estimator from each adjustment method, in addition to the naive sample estimate. Error bars delimit 95% confidence/credible intervals. The solid vertical black line denotes the true population trend (-0.047).

Distributions of auxiliary variables

As measured using Mean Absolute Errors (MAEs), the adjustment methods were generally very good at bringing the distributions of the auxiliaries in the samples closer to those in the population. Fig 5 shows the sample and population distributions of elevation, but the MAEs for this and the proportion of each grid square that falls within a protected area can be found in supplementary material 2.

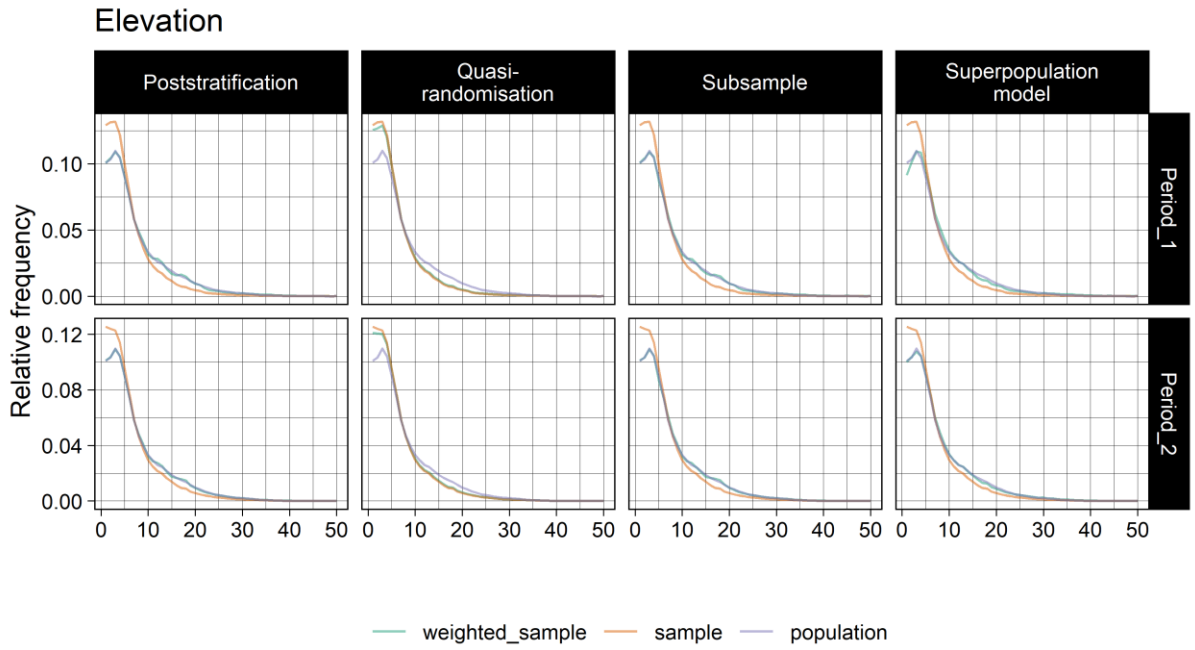

Figure 5. Sample, population and weighted sample distributions of the auxiliary variable road length (Table 1) in periods one and two.

Discussion

We applied six approaches to descriptive inference from nonprobability samples to a simple biodiversity monitoring problem: the estimation of mean occupancy of the plant *C. vulgaris* in two time-periods and the trend between the two. The methods generally worked well in the sense that they brought the distributions of auxiliary variables in the samples closer to their distributions in the population (all 1 km grid squares in Britain). Successful redistribution of the auxiliaries translated into improvements of the estimates of mean occupancy in both time-periods and the trend between the two. Importantly, however, no method was completely unbiased, and their uncertainty intervals did not cover the true values of occupancy in the second period or the trend. An abatement rather than an elimination of bias is probably the best outcome that can be expected, because most adjustment methods rest on the untenable assumption that non-sampled locations are "Missing At Random" (MAR); that is, the variable of interest is completely independent of sample inclusion given the auxiliary variables.

Unlike most practical situations, we were able to test the MAR assumption, because we know the true distribution of *C. vulgaris* in Britain. In the first time-period, the partial correlation between sample inclusion and occupancy, conditional on elevation and protected area coverage, is -0.018; in period two, it is 0.035 (supplementary material 1). These "adjusted" ddc's are lower in magnitude than the original ddc's, -0.115 and -0.058, which means that accounting for elevation and protected area coverage increased the representativeness of the samples (recalling that a smaller ddc means a more representative sample). That is not to say that the samples became *fully* representative, which would be the case in expectation in a MAR scenario. The usual yardstick for a representative sample is the simple random sample, whose ddc is of the order $N^{-1/2}$ (Meng, 2018). In our example, $N^{-1/2} = 2.2^{-6}$, which is several orders of magnitude smaller than the "adjusted" ddc's. This goes to show that without a truly miniscule ddc, which would only be induced (in expectation) where the MAR assumption holds or under random sampling, sample means as estimators of population means will be appreciably biased (especially where N is large).

It might seem wise to include as many potential auxiliaries as possible to reduce the chance of missing a genuine one. For example, Collins et al. (2001) advocated for including all variables exceeding some prescribed correlation with sample inclusion and the variable of interest. This strategy can be a dangerous one, however. Thoemmes & Rose (2014) show that including correlates of sample inclusion and the variable of interest, rather than theoretically justifiable causes, can increase the bias in estimates of population means (also see Thoemmes & Mohan, 2015). Indeed, in a previous version of this manuscript (Boyd, Stewart, et al., 2023), we took a more inclusive approach to the selection of auxiliary variables, and our estimates of *C. vulgaris* occupancy in period two were generally more biased than the naïve estimate from the unadjusted sample.

Identifying appropriate auxiliary variables is likely to be the most challenging part of adjusting samples in biodiversity monitoring. In many situations, causes of the variable of interest and sample inclusion are not known. Taxon and dataset experts might be able to identify potential auxiliary variables, but it is unlikely that they can identify them all (which would be needed to satisfy the MAR assumption). The experts might also erroneously identify auxiliary variables that are not suitable, in which case adjusting for those variables might do more harm than good (Thoemmes & Rose, 2014). Even if experts were able to correctly identify all relevant auxiliaries, those variables might not be reflected in available data. Transparency regarding availability and choice of auxiliary variables should be an important component of reporting for all biodiversity monitoring.

Acknowledging that variables of interest in biodiversity monitoring are likely to be dependent on sample inclusion even after controlling for the available auxiliaries, it might be worth considering adjustment methods that forgo the MAR assumption. For example, Tchetgen Tchetgen & Wirth (2017) showed it is possible to recover a true population regression model (and therefore the population mean) by incorporating "instrumental variables". They define instrumental variables as those that are predictive of sample inclusion, independent of the variable of interest and independent of "selection bias" (the latter defined as the mean of the variable of interest in the sample minus the mean of the variable of interest in non-sampled population units). We screened three additional variables—the proportion of each grid square that is accessible to the public, the density of postcodes in each grid square and its nearest neighbours, and the length of major roads in each grid square and its nearest neighbours—to see if they satisfied these three assumptions, but none did (supplementary material 1). In practical situations, where the variable of interest is not known for non-sampled population units, testing these assumptions would be challenging.

Whilst we are confident that the availability of data on auxiliary variables was the limiting factor in our example, it is possible that improvements to the adjustment methods themselves could have improved matters. Where sampling weights are not of interest, for example, it might be sensible to use a binomial generalised linear model, rather than a general linear regression, for the superpopulation model (Wu & Sitter, 2001). The multilevel modelling component of MRP exploits partial pooling, so we could have used more finely resolved strata on the basis that estimates for sparse strata (with low sample sizes) would be shrunk towards those from strata with more data. The question is whether fine-tuning the adjustment methods is likely to result in large improvements in accuracy. As Mercer et al., (2018), writing in the context of adjusting survey samples, put it, "[t]he right variables make a big difference for accuracy. Complex statistical methods, not so much." The fact that most adjustment methods performed almost similarly in our example is further evidence that the choice of auxiliary variables matters more than the specifics of the adjustment method.

Given that the methods performed similarly in terms of accuracy, it would be sensible to consider those that are quickest to run. As we implemented it, MRP took by far the longest to run of all the methods—about ten hours per time-period on a computer cluster. Bootstrapping to estimate confidence intervals meant that other methods, too, were quite expensive to run. This was particularly true for the quasi-randomisation and doubly robust procedures, both of which involved repeatedly fitting the sample inclusion model—itself a time consuming process. The remainder of the methods—superpopulation modelling, subsampling and poststratification—took a negligible amount of time to run.

Although we have only considered one species and dataset, previous studies (in other disciplines) shed light on the factors that affect the accuracy of inference from nonprobability samples more generally. Omitting genuine auxiliary variables in the adjustment process is more problematic where those variables explain larger proportions of the variance in the variable of interest and sample inclusion (Collins et al., 2001). Equally, inclusion of certain variables that are not appropriate auxiliaries becomes more problematic where they explain larger proportions of the variance in the variable of interest and sample inclusion (Thoemmes & Rose, 2014). In practice, we do not know the strengths of the effects potential auxiliaries on the variable of interest and sample inclusion, or whether they have effects at all, but it is clear that the selection of auxiliary variables will be a critical component of adjusting samples in biodiversity monitoring.

Given the importance of selecting appropriate auxiliary variables, we propose the following general strategy for analysts intending to draw inferences about biodiversity change from geographically unrepresentative nonprobability samples. The first step should be to consult taxon and dataset experts, who might be able to identify relevant auxiliary variables. Where possible, consulting multiple experts to capture their uncertainty about what affects sample inclusion and the variable of interest would be desirable. If data are available on these variables, then their distributions in the sample and population should be compared to assess whether the data are representative with respect to that variable. Several tools are available to perform such comparisons (Boyd et al., 2021; Ruete, 2015). The next step should be to adjust the sample based on the relevant auxiliaries and to draw inferences from the adjusted samples. Like others (e.g. Mercer et al., 2018), we found that it is of little consequence which adjustment method is used, so it is sensible to pick one that is quick to run. Rather than assuming the adjustment worked perfectly, it is important to acknowledge and report the potential for residual bias. As we have shown, traditional uncertainty intervals are not guaranteed (or even likely) to cover the true population parameters of interest unless all relevant auxiliaries are known and reflected in available data (Meng, 2018). Where there is doubt about the relevant auxiliary variables, a safer strategy is to assess the risk of bias qualitatively and to ensure it is reflected in the way that findings are reported (Boyd et al., 2022; Meineke & Daru, 2021; Pescott et al., 2022).

Acknowledgements

Thank you to Richard Valliant and two anonymous reviewers, whose comments improved this paper. All authors were supported by the NERC Exploring the Frontiers award number NE/X010384/1 “Biodiversity indicators from nonprobability samples: Interdisciplinary learning for science and society”. OLP was also supported by the NERC award number NE/R016429/1 as part of the UK Status, Change and Projections of the Environment (UK-SCAPE) programme delivering National Capability.

References

Bailey, M. (2022). Comments on “ Statistical inference with non-probability survey samples .” *Survey Methodology*, *48*(12), 331–338.

Bethlehem, J., Cobben, F., & Schouten, B. (2008). Indicators for the Representativeness of Survey Response. *Statistics Canada’s International Symposium Series: Proceedings*, *11*.

Bowler, D. E., Bhandari, N., Repke, L., Beuthner, C., Callaghan, C. T., Eichenberg, D., Henle, K., Klenke, R., Richter, A., Jansen, F., Bruelheide, H., & Bonn, A. (2022). Decision-making of citizen scientists when recording species observations. *Scientific Reports*, *12*(1), 1–12. https://doi.org/10.1038/s41598-022-15218-2

Bowler, D. E., Klaus-, D. E., Conze, J., Suhling, F., Baumann, K., Benken, T., Bönsel, A., Bittner, T., Drews, A., Günther, A., Isaac, N. J. B., & Petzold, F. (2021). Winners and losers over 35 years of dragonfly and damselfly distributional change in Germany. *Diversity and Distributions*, *27*(August 2020), 1353–1366. https://doi.org/10.1111/ddi.13274

Boyd, R. J., Powney, G., Carvell, C., & Pescott, O. L. (2021). occAssess: An R package for assessing potential biases in species occurrence data. *Ecology and Evolution*, *11*(September), 16177– 16187. https://doi.org/10.1002/ece3.8299

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*(March), 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*, *xx*(xx), 1–11. https://doi.org/10.1016/j.tree.2023.01.001

Boyd, R. J., Stewart, G. B., & Pescott, O. L. (2023). Descriptive inference using large , unrepresentative nonprobability samples : An introduction for ecologists. *Ecoevorxiv*, *April*. https://doi.org/10.32942/X2359P

Chen, Y., Li, P., & Wu, C. (2020). Doubly Robust Inference With Nonprobability Survey Samples. *Journal of the American Statistical Association*, *115*(532), 2011–2021. https://doi.org/10.1080/01621459.2019.1677241

Collins, L. M., Schafer, J., & Kam, C. (2001). A Comparison of Restrictive Strategies in Modern Missing Data Procedures. *Psychological Methods*, *6*(June). https://doi.org/10.1037/1082-989X.6.4.330

Dornelas, M., Gotelli, N. J., McGill, B., Shimadzu, H., Moyes, F., Sievers, C., & Magurran, A. E. (2014). Assemblage time series reveal biodiversity change but not systematic loss.

Science, *344*(6181), 296–299. https://doi.org/10.1126/science.1248484

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

Forister, M. L., Black, S. H., Elphick, C. S., Grames, E. M., Halsch, C. A., Schultz, C. B., & Wagner, D. L. (2023). Missing the bigger picture: Why insect monitoring programs are limited in their ability to document the effects of habitat loss. *Conservation Letters*, *September 2022*, 1–6. https://doi.org/10.1111/conl.12951

Geldmann, J., Heilmann-Clausen, J., Holm, T. E., Levinsky, I., Markussen, B., Olsen, K., Rahbek, C., & Tøttrup, A. P. (2016). What determines spatial bias in citizen science? Exploring four recording schemes with different proficiency requirements. *Diversity and Distributions*, *22*(11), 1139–1149. https://doi.org/10.1111/ddi.12477

Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science*, *22*(2), 153–164. https://doi.org/10.1214/088342306000000691

Gelman, A., & Little, T. (1997). poststratification into many categories using hierarchical regression. *Survey Methodology*, *23*(2), 127–335.

Girardello, M., Chapman, A., Dennis, R., Kaila, L., Borges, P. A. V., & Santangeli, A. (2019). Gaps in butterfly inventory data: A global analysis. *Biological Conservation*, *236*(November 2018), 289–295. https://doi.org/10.1016/j.biocon.2019.05.053

Gonzalez, A., Cardinale, B. J., Allington, G. R. H., Byrnes, J., Endsley, K. A., Brown, D. G., Hooper, D. U., Isbell, F., O'Connor, M. I., & Loreau, M. (2016). Estimating local biodiversity change: A critique of papers claiming no net loss of local diversity. *Ecology*, *97*(8), 1949–1960. https://doi.org/10.1890/15-1759.1

Gregory, R. D., Van Strien, A., Vorisek, P., Meyling, A. W. G., Noble, D. G., Foppen, R. P. B., & Gibbons, D. W. (2005). Developing indicators for European birds. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1454), 269–288. https://doi.org/10.1098/rstb.2004.1602

Hudson, L. N., Newbold, T., Contu, S., Hill, S. L. L., Lysenko, I., Palma, D., Phillips, H. R. P., Senior, R. A., Bennett, D. J., Booth, H., Garon, M., Michelle, L., Correia, D. L. P., Day, J., Echeverr, S., Harrison, K., Ingram, D. J., Jung, M., Kemp, V., … Fernando, A. B. (2014). *The PREDICTS database : a global database of how local terrestrial biodiversity responds to human impacts*. 4701–4735.

Hughes, A., Orr, M., Ma, K., Costello, M., Waller, J., Provoost, P., Zhu, C., & Qiao, H. (2020). Sampling biases shape our view of the natural world. *Ecography*, *44*, 1259–1269. https://doi.org/10.1111/ecog.05926

Intermap. (2009). *NEXTMap British Digital Terrain 50m resolution (DTM10) Model Data by Intermap*. NERC Earth Observation Centre. https://catalogue.ceda.ac.uk/uuid/f5d41db1170f41819497d15dd8052ad2

Johnston, A., Matechou, E., & Dennis, E. B. (2022). Outstanding challenges and future directions for biodiversity monitoring using citizen science data. *Methods in Ecology and Evolution*, *February*. https://doi.org/10.1111/2041-210X.13834

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

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

Lumley, T. (2010). *Complex surveys: A Guide to Analysis Using R* (1st ed.). Wiley.

Makela, S., Si, Y., & Gelman, A. (2014). Statistical Graphics for Survey Weights. *Revista Colombiana de Estadística*, *37*(2Spe), 285–295. https://doi.org/10.15446/rce.v37n2spe.47937

Mandeville, C. P., Nilsen, E. B., & Finstad, A. G. (2022). Spatial distribution of biodiversity citizen science in a natural area depends on area accessibility and differs from other recreational area use. *Ecological Solutions and Evidence*, *3*(4), 1–14. https://doi.org/10.1002/2688-8319.12185

McClure, C. J. W., & Rolek, B. W. (2023). Pitfalls arising from site selection bias in population monitoring defy simple heuristics. *Methods in Ecology and Evolution*, *14*(6), 1489–1499. https://doi.org/10.1111/2041-210X.14120

Meineke, E. K., & Daru, B. H. (2021). Bias assessments to expand research harnessing biological collections. *Trends in Ecology and Evolution*, *36*(12), 1071–1082. https://doi.org/10.1016/j.tree.2021.08.003

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 1 : Miniaturizing data defect correlation : A versatile strategy for handling non-probability samples. *Survey Methodology*, *48*(2), 1–22.

Mercer, A., Lau, A., & Kennedy, C. (2018). For Weighting Online Opt-In Samples, What Matters Most? In *Pew Research Center* (pp. 1–55). https://pewrsr.ch/3heqknn

Morton, R., Marston, C., O'Neil, A., & Rowland, C. (2022). *Land Cover Map 2018 (1km summary rasters, GB and N. Ireland).* NERC EDS Environmental Information Data Centre. https://doi.org/https://doi.org/10.5285/9b68ee52-8a95-41eb-8ef1-8d29e2570b00

Nelson, G., & Ellis, S. (2019). The history and impact of digitization and digital data mobilization on biodiversity research. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *374*(1763), 2–10. https://doi.org/10.1098/rstb.2017.0391

Outhwaite, C., Gregory, R. D., Chandler, R. E., Collen, B., & Isaac, N. J. B. (2020). Complex long-term biodiversity change among invertebrates, bryophytes and lichens. *Nature Ecology & Evolution*. https://doi.org/10.1038/s41559-020-1111-z

Perring, F., & Walters, S. (1962). *Atlas of the British flora*. Thomas Nelson & sons.

Pescott, O. L., Stroh, P. A., Humphrey, T. A., & Walker, K. J. (2022). Simple methods for improving the communication of uncertainty in species ' temporal trends. *Ecological Indicators*, *141*(May). https://doi.org/https://doi.org/10.1016/j.ecolind.2022.109117

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

Pescott, O. L., Walker, K. J., Pocock, M. J. O., Jitlal, M., Outhwaite, C. L., Cheffings, C. M.,

671 Harris, F., & Roy, D. B. (2015). Ecological monitoring with citizen science: The design
672 and implementation of schemes for recording plants in Britain and Ireland. *Biological*
673 *Journal of the Linnean Society*, *115*(3), 505–521. https://doi.org/10.1111/bij.12581

674 Powney, G. D., Carvell, C., Edwards, M., Morris, R. K. A., Roy, H. E., Woodcock, B. A., &
675 Isaac, N. J. B. (2019). Widespread losses of pollinating insects in Britain. *Nature*
676 *Communications*, *10*(2019), 1–6. https://doi.org/10.1038/s41467-019-08974-9

677 Preston, C.D., Pearman, D.A. & Dines, T. D. (2002). *New Atlas of the British and Irish*
678 *Flora.* (eds). Oxford University Press.

679 Rowland, C., Marston, C., Morton, R., & O'Neil, A. (2020). *Land Cover Map 1990 (1km*
680 *dominant target class, GB) v2.* NERC EDS Environmental Information Data Centre.
681 https://doi.org/https://doi.org/10.5285/f5e3bd00-efd0-4dc6-a454-aa597d84764a

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

684 Ruete, A. (2015). Displaying bias in sampling effort of data accessed from biodiversity
685 databases using ignorance maps. *Biodiversity Data Journal*, *3*(1), 1–15.
686 https://doi.org/10.3897/BDJ.3.e5361

687 Stroh, P. A., Walker, K., Humphrey, T. A., Pescott, O. L., & Burkmar, R. (2023). *Plant Atlas*
688 *2020: Mapping Changes in the Distribution of the British and Irish Flora*. Princeton
689 Univ. Press.

690 Tchetgen Tchetgen, E. J., & Wirth, K. E. (2017). A general instrumental variable framework
691 for regression analysis with outcome missing not at random. *Biometrics*, *73*(4), 1123–
692 1131. https://doi.org/10.1111/biom.12670

693 Thoemmes, F., & Mohan, K. (2015). Graphical Representation of Missing Data Problems.
694 *Structural Equation Modeling*, *22*(4), 631–642.
695 https://doi.org/10.1080/10705511.2014.937378

696 Thoemmes, F., & Rose, N. (2014). A Cautious Note on Auxiliary Variables That Can
697 Increase Bias in Missing Data Problems. *Multivariate Behavioral Research*, *49*(5), 443–
698 459. https://doi.org/10.1080/00273171.2014.931799

699 UNEP-WCMC, & IUCN. (2020). *Protected Planet: The World Database on Protected Areas*
700 *(WDPA)/The Global Database on Protected Areas Management Effectiveness*.
701 https://www.protectedplanet.net/en/thematic-areas/wdpa

702 Valliant, R. (2020). Comparing Alternatives for Estimation from Nonprobability Samples.
703 *Journal of Survey Statistics and Methodology*, *8*(2), 231–263.
704 https://doi.org/10.1093/jssam/smz003

705 Valliant, R., Dever, J. A., & Kreuter, F. (2018). *Practical tools for designing and weighting*
706 *survey samples* (2nd ed.). Springer Cham. https://doi.org/https://doi.org/10.1007/978-3-
707 319-93632-1

708 van Strien, A. J., & van Grunsven, R. H. A. (2023). In the past 100 years dragonflies declined
709 and recovered by habitat restoration and climate change. *Biological Conservation*,
710 *277*(December 2022), 109865. https://doi.org/10.1016/j.biocon.2022.109865

711 Van Swaay, C. A. M., Nowicki, P., Settele, J., & Van Strien, A. J. (2008). Butterfly
712 monitoring in Europe: Methods, applications and perspectives. *Biodiversity and*

Conservation, *17*(14), 3455–3469. https://doi.org/10.1007/s10531-008-9491-4

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.

Vellend, M., Baeten, L., Myers-Smith, I. H., Elmendorf, S. C., Beauséjour, R., Brown, C. D., De Frenne, P., Verheyen, K., & Wipf, S. (2013). Global meta-analysis reveals no net change in local-scale plant biodiversity over time. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(48), 19456–19459. https://doi.org/10.1073/pnas.1312779110

Weiser, E. L., Diffendorfer, J. E., Lopez-Hoffman, L., Semmens, D., & Thogmartin, W. E. (2020). Challenges for leveraging citizen science to support statistically robust monitoring programs. *Biological Conservation*, *242*(October 2019). https://doi.org/10.1016/j.biocon.2020.108411

Wu, C. (2022). *Statistical inference with non-probability survey samples*. *12*.

Wu, C., & Sitter, R. R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, *96*(453), 185–193. https://doi.org/10.1198/016214501750333054