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

In the age of big data, it is essential to remember that the size of a dataset is not all that matters. This is particularly true where the goal is to draw inferences about some wider population, in which case it is far more important that the data are *representative* of that population. It is possible to adjust unrepresentative samples so that they more closely resemble the population in terms of "auxiliary variables". If the auxiliaries predict sample inclusion and/or the variable of interest well, then the adjusted sample estimates will be closer to the truth. Several survey sampling techniques exist to perform such adjustments, but most are not familiar to ecologists. We applied five types of adjustment—subsampling, quasi-randomisation, poststratification, superpopulation modelling, 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). *Calluna vulgaris* is an attractive case study because we have good estimates of its true distribution in both time-periods. We estimated the means and trend using large, but (originally) unrepresentative, samples. Compared to the unadjusted estimates, the means and trends estimated using most adjustment methods were more accurate, although their uncertainty intervals generally did not cover the true values. Quasi-randomisation performed especially poorly, and we explain why. Most adjustments were far more successful at bringing the distributions of the auxiliary variables in the samples closer to those in the population than they were at improving the estimates of population means and trends. This implies that the major challenge for adjusting unrepresentative samples in biodiversity monitoring is assembling a suitable set of auxiliary variables (i.e. predictors of sample inclusion and the variable of interest). This challenge will be particularly acute for poorly studied taxa and those whose habitat requirements or sampling biases are not reflected in available data.

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, in which case the sample would be a subset of those units. 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 patches of land 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.

More important than the size of a sample for descriptive inference is whether it is representative of the population (X. L. 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" (ddc; Meng, 2018). Intuitively, statistics derived from a representative sample, such as means and proportions, will be similar to their population equivalents. The challenge is that the variable of interest is unknown in non-sampled population units, so it is typically impossible to calculate a sample's representativeness exactly.

Rather than measuring sample representativeness in terms of the variable of interest, which is not known for all population units, it is possible to approximate it using "auxiliary variables". Auxiliary variables are those that are thought to predict the variable of interest or the probability that each population unit was sampled. Such variables might be available for every population unit. For example, climate variables might explain a species' occupancy, and data on these variables are available for every 1 km^2 grid square across the globe (Fick & Hijmans, 2017). If the distributions of auxiliary variables in the sample are different to those in the population, then the sample is likely to be unrepresentative, at least with respect to those variables (Bethlehem et al., 2008).

It is possible to adjust an unrepresentative sample by placing more weight on population units that were less likely to be sampled. Weighting is simple where the probability that each population unit was sampled is known (i.e. in a probability sample). For example, rather than using a sample mean to estimate a population mean, the researcher would use a weighted mean with the weights being the inverses of the inclusion probabilities (Lohr, 2022). Weighting of this type is known as "design-based inference", because the inclusion probabilities are a feature of the sampling design. Unfortunately, design-based inference is not applicable for many "big" biodiversity datasets, whose sample inclusion probabilities are not known (i.e. they are nonprobability samples), so alternatives are required.

Most approaches to descriptive inference from nonprobability samples make use of auxiliary variables. The details differ, but the general strategy is to weight the sample in such a way that the distributions of the auxiliary variables in the sample more closely resemble those in the population (Valliant et al., 2018). If the auxiliaries predict the variable of interest and sample inclusion well, then this is essentially the same as bringing the distribution of the variable of interest in the sample closer to its distribution in the population (i.e. making the sample more representative; see X. Meng, [2022], who demonstrates this mathematically).

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; C.A.M. Van Swaay et al., 2002; Chris A.M. Van Swaay et al., 2008; Weiser et al., 2020). But it is also common to see sample representativeness ignored, an issue that has led to some high-profile controversies in the biodiversity monitoring literature (Boyd, Powney, et al., 2023). 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 five 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 trend). *Calluna vulgaris* is an attractive case study because we have good estimates of its true geographic distribution in both periods from several sources. The approaches to inference that we demonstrate are subsampling, quasi-randomisation (Elliott and Valliant, 2017), poststratification (Little, 1993), superpopulation modelling (Valliant, 2009) 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

Estimating the 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). 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 used cover the main periods of recording for two national distribution atlases (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.

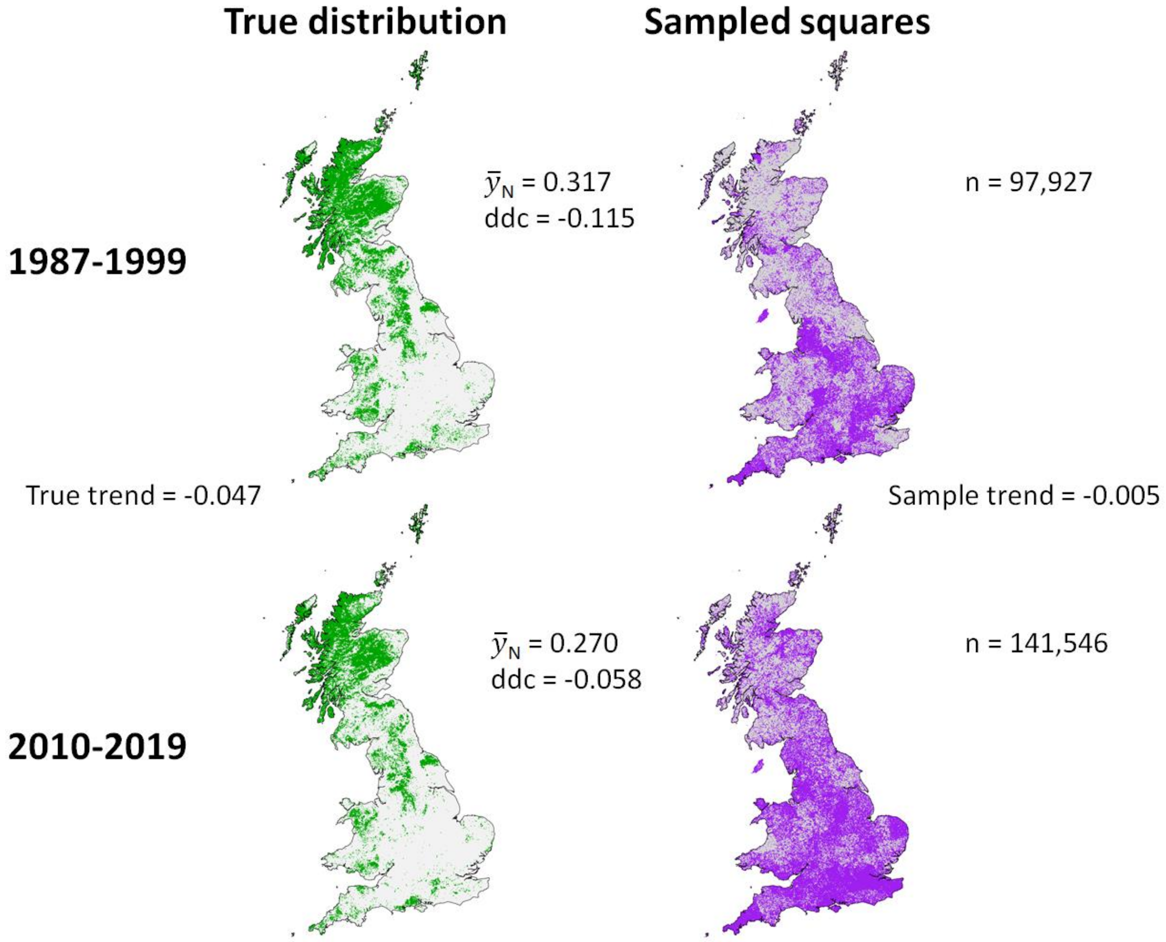

Figure 1. Left column: the distribution of *Calluna vulgaris* in both time-periods. Green squares are occupied and grey squares are not. $\bar{y}_N$ is mean occupancy or, equivalently, the proportion of squares occupied. The ddc's are the correlations between sample inclusion (1 if the square is in the 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 assigned to a single 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).

Auxiliary data

We used five auxiliary variables for which data are available for all 1 km grid squares in Great Britain (Table 1). Most of the auxiliary variables indicate the accessibility or attractiveness of grid squares, which tend to be associated with site selection in citizen science datasets (Geldmann et al., 2016). Elevation is a potential predictor of *C. vulgaris* occupancy.

Originally, we included three additional predictors of *C. vulgaris* occupancy—the first and third principal components of climate space in Britain and soil pH—but later omitted them. We had previously found the climate variables to be important predictors of 1 km habitat suitability for *C. vulgaris* using species distribution models (Boyd, Harvey, et al., 2023). Including these predictors did not improve the estimates of mean occupancy, a point that we expand on in the discussion. Reducing the set of auxiliary variables simplifies matters for some of the adjustment methods that we present below.

For simplicity, we assume that the auxiliary variables are constant between time-periods. This assumption is obviously violated for some variables (e.g. road length and postcode density). However, this should not matter if, in reality, the variables in period one are correlated with those in period two, because any given grid square will generally have a higher or lower value than the others regardless of the period. We think that this situation is plausible: for example, there is a higher density of postcodes in London in period two than in period one, but in either period, it has a higher density than elsewhere. Another reason to use one set of auxiliary data for both time-periods is to make our findings more applicable to situations in which temporally resolved data are not available (e.g. in data poor countries or periods in the distant past).

Table 1. Auxiliary variables used for sample adjustment.

Variable	Reason for inclusion	Details	Reference(s)
Postcode density	Indicates population density in vicinity	Total number of postcodes in the focal grid square and its 299 nearest neighbours	ONS (2021)
Road length	Indicates accessibility	The total length of all "Roads" and "Link roads" ("Highways" class of the OpenStreetMap ontology) in the focal grid square and its 299 nearest neighbours	Data from https://www.openstreetmap.org/ under an open database license
Proportion in protected area	Indicates potential attractiveness to surveyor	Proportion of the focal grid square with some level of "protection". Includes everything from SSSIs to e.g. local nature reserves	UNEP-WCMC & IUCN (2020)
Proportion open access land	Indicates accessibility	Proportion of land legally designated as open access within 1 km grid square	All open access land datasets in GB are available via an Open Government License. For England, we used the CRoW act 2000 layer. For Wales, we combined the registered common land, other statutory access land, open country and public forest datasets. All of the Scottish countryside is open access.
Average elevation	Predictor of *C. vulgaris* occupancy	Average elevation of 1 km grid square calculated from 50 m digital terrain model	Intermap (2009)

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

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

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. 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 1)

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

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 five 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 2). The other two, subsampling and MRP, 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).

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, 2) is 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 auxiliary data 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 the auxiliary data in Table 1 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 most auxiliary variables into three categories using their terciles (i.e. cut points at the 33rd and 67th percentiles). This did not make sense for the variables denoting the proportion of each grid square that is open access land and protected area, because most squares took the value one or zero. We split these variables into two categories, 0 and >0, i.e. whether or not there is some open or protected land in the grid square. Discretization initially gave $3 \times 3 \times 3 \times 2 \times 2 = 108$ poststrata, from which we subtracted one poststratum that contained no population units, leaving 107.

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.

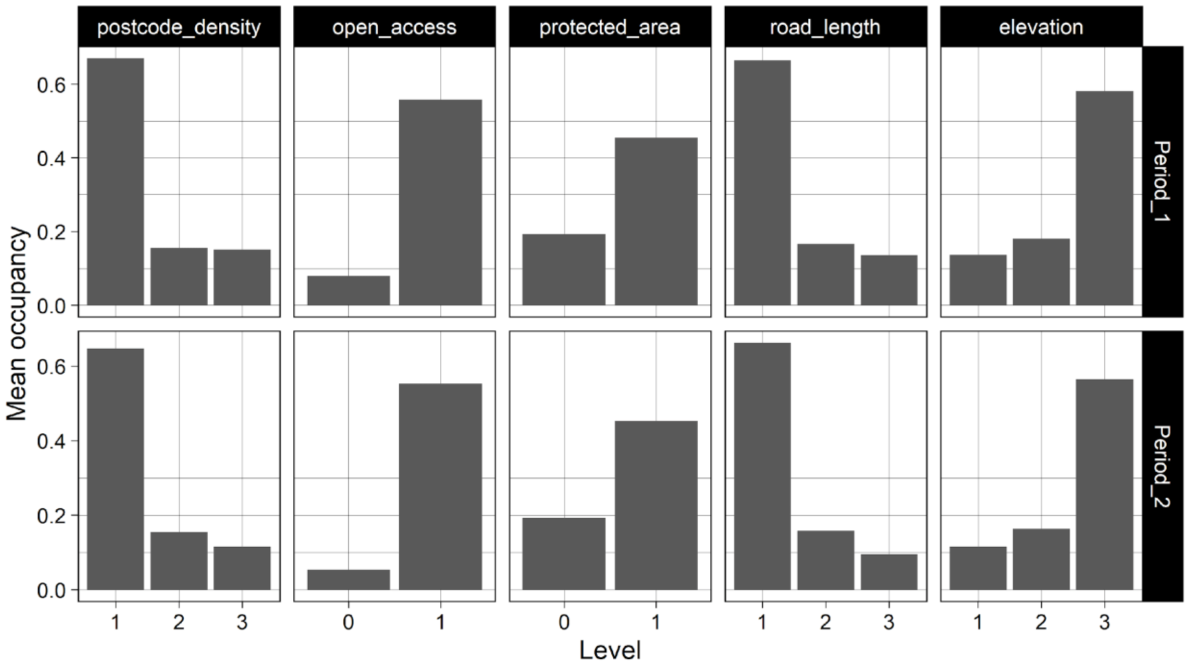

Figure 2. Mean occupancy of *Calluna vulgaris* for each level of the auxiliary variables (Table 1) in each time-period. The auxiliary variables were originally on a continuous scale, but we discretized them to enable poststratification. See the main text for details.

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

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

where the subscript M indicates that the expectation (mean) is with respect to the model, $\boldsymbol{x}_i$ is a vector of predictors 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. There is some matrix notation in 3) and what follows, but the logic should be apparent to those who do not understand the precise detail. A prediction of y for unit i is

$$\hat{y}_i = x_i^T\ \hat{\beta}. \tag{4)}$$

The accent on β indicates that it is an estimate. Given a sample s, one estimator of β is $\hat{\beta} = A_s^{-1} X_s^T y_s$, where $A_s = X_s^T X_s$, X_s is an $n\ x\ p$ matrix of covariates and $\boldsymbol{y}_s$ is an n vector of y's (Valliant, 2020). If $\bar{s}$ is the set of non-sampled population units, the superpopulation model prediction of the population mean is

$$\bar{y}_{sp} = \frac{\sum_{i \in s} y + \sum_{i \in \bar{s}} \hat{y}}{N}. \tag{5)}$$

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 2), with $w_i = 1 + \boldsymbol{t}_{\bar{s}}^x A_s^{-1} x_i$ and $\boldsymbol{t}_{\bar{s}}^x$ being the vector of population totals of the auxiliary variables in non-sampled population units (Elliott and 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 of the form 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). 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 in Table 1 as predictors.

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 weighted random samples of size 500 with replacement from the original samples [note that these weights are different to sampling weights in 2)]. The decision to set $n = 500$ 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). We assigned each grid square i in poststratum j (using the same strata as described above under Poststratification) a weight of n_j/N_j. The result was subsamples whose members were more likely to be from strata comprising a larger fraction of the population. The subsample mean is the estimator of the population mean.

Rather than using a single subsample, we repeated the process 1000 times and used the mean of the estimated means (i.e. bootstrapping). This was necessary, because the estimated means were sensitive to the random component of the subsampling.

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 simple one. It is a binomial GLM with a logit link function, a fixed intercept and a random intercept for each auxiliary variable (see https://mc-stan.org/rstanarm/articles/mrp.html for a similar formulation). A more complex model might include interactions among the auxiliaries (e.g. Ghitza and Gelman, 2013), but we found these take several times longer to run. Long run times may be undesirable for production-type statistical workflows in biodiversity monitoring, where models might need to be fitted for thousands of species in tens of time-periods. Even without interactions, and on a computer cluster, the models took around ten hours to run per time-period. 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 was implemented in a Bayesian framework). The *survey* package (Lumley, 2010), which we used to calculate the sample means, the superpopulation model estimates and the poststratified estimates, calculates the confidence intervals automatically. It accounts for the sampling weights where relevant. We used percentile confidence intervals from the bootstrapped subsamples.

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). The standard errors of the trends are $\sqrt{var\left(\underline{y}_2\right) + var\left(\underline{y}_1\right)}$ (Gelman, 2007), where the variances are sampling *not sample* variances (i.e. the square of the standard error rather than a measure of variability in the samples). We used the standard errors returned by the *survey* package, which accounts for the sampling weights. We present 95% confidence intervals for the trends from most estimators ($\pm 1.96 \times$ the standard errors). MRP is one exception, because the 95% credible interval can be calculated directly from the posterior distribution of $\bar{y}_2 - \bar{y}_1$. Similarly, we extracted percentile 95% confidence intervals for the subsampling estimator from the bootstrapped distribution of trends.

Evaluating the effects of the adjustments

We used relative frequency plots (c.f. 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 construct adjusted relative frequency distributions from MRP so omit it from this portion of the analysis. The problem is that, whilst it has been shown how to derive unit-level sampling weights where the multilevel model is linear (Gelman, 2007), no formula has yet been derived for the case of the binomial GLM (Valliant et al., 2018). 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.

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 -0.058. 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).

The adjustment methods did not always result in improved point estimates of mean occupancy relative to the naive sample means (Fig. 3). In period one, the adjusted estimates were generally better in terms of absolute errors, with the exception of the subsample estimate, which was worse. In period two, on the other hand, the estimate from the subsample was the only one to get closer than the naive sample mean (again, in terms of absolute error). The absolute errors are provided in supplementary material 3.

In terms of confidence/credible interval coverage, the estimators were generally very poor. With the exception of the subsample means, none covered the population mean in either period. The fact that the confidence intervals from the subsamples did cover the population means is not surprising: the subsamples are small ($n = 500$), so the confidence intervals are wide. Of course, increasing the size of the subsamples reduces the width of the confidence intervals, as we show in supplementary material 1.

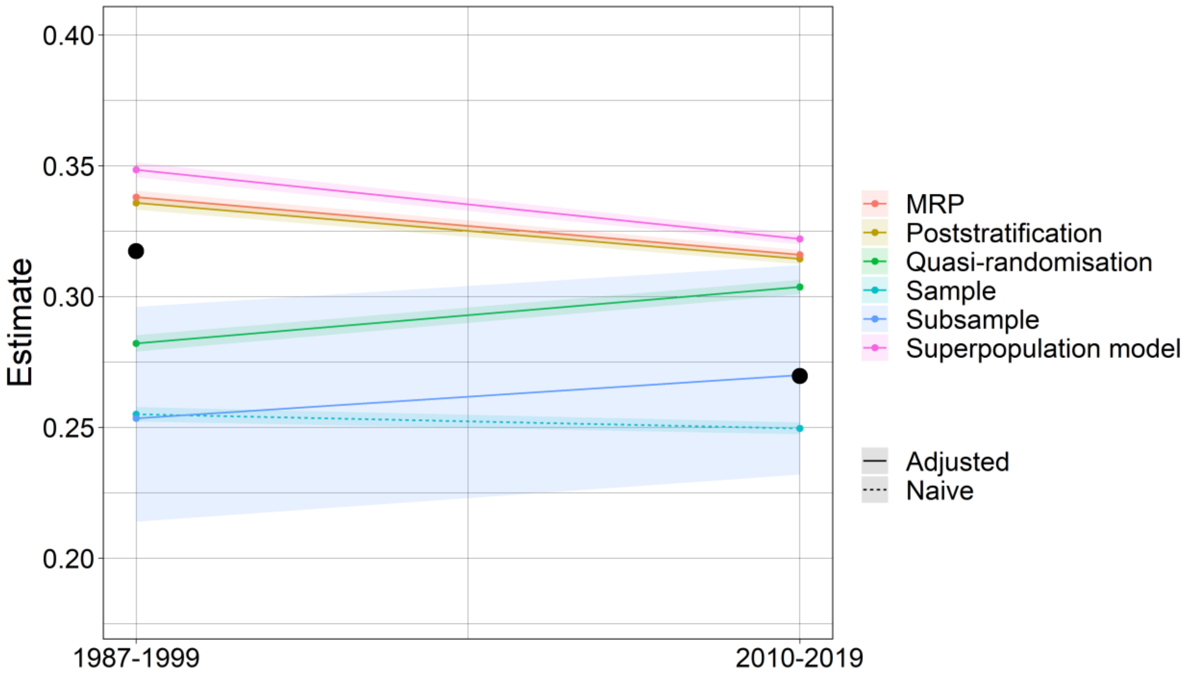

Figure 3. Naive (i.e. unadjusted) and adjusted sample-based estimates of mean occupancy in each time-period. The shaded regions are 95% confidence/credible intervals (see the main text for information on how these have been constructed for each method). The large black circles are the true population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

Estimated trend in mean occupancy

Three of the five adjusted point estimates of the trend in mean occupancy are closer than the difference in naive sample means to the true population trends. The other two, the trends from quasi-randomisation and subsampling, are poor. Their point estimates even have the wrong sign. No estimator's credible/confidence interval covers the true trend. The fact that the naïve sample trend underestimates the true trend is a consequence of the time varying representativeness (Bowler et al., 2022; Oliver L. Pescott et al., 2019).

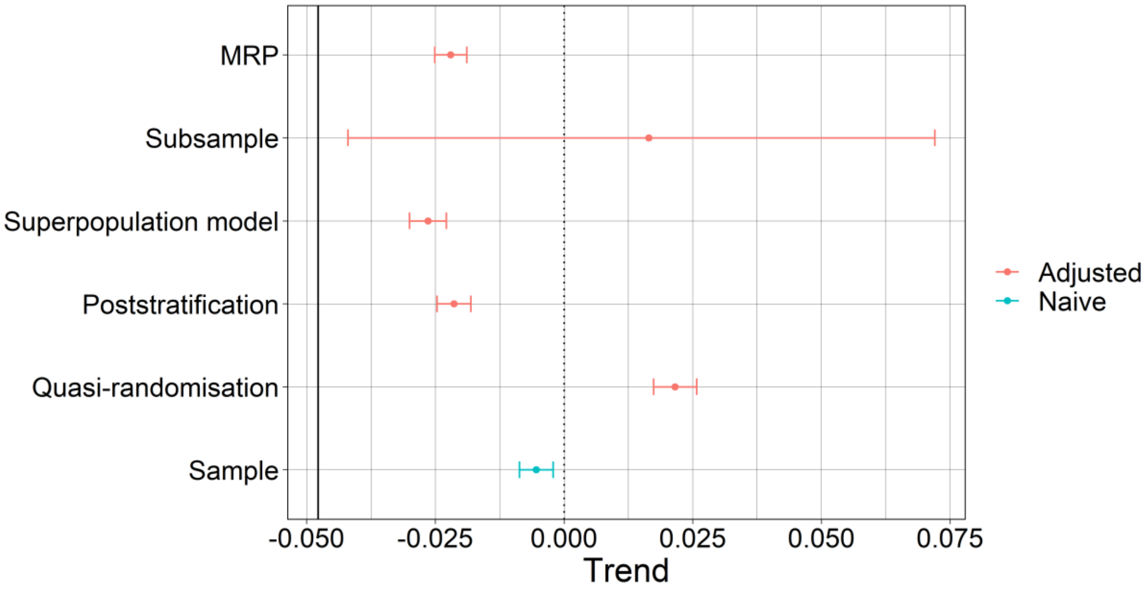

Figure 4. Trends in mean occupancy between periods one and two produced by the estimator from each adjustment method, in addition to the naive sample estimate. Error bars delimit 95% confidence/credible intervals. The solid vertical black line denotes the true population trend (-0.047).

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. Figs 5 and 6 show the sample and population distributions of two auxiliary variables, road length and elevation, but the MAEs for these and the others can be found in supplementary material 3. Superpopulation modelling and poststratification performed particularly well. Quasi-randomisation offered only a minor improvement in period one. Subsampling was the only approach that did not bring the distributions of the auxiliaries in the sample closer to those in the population.

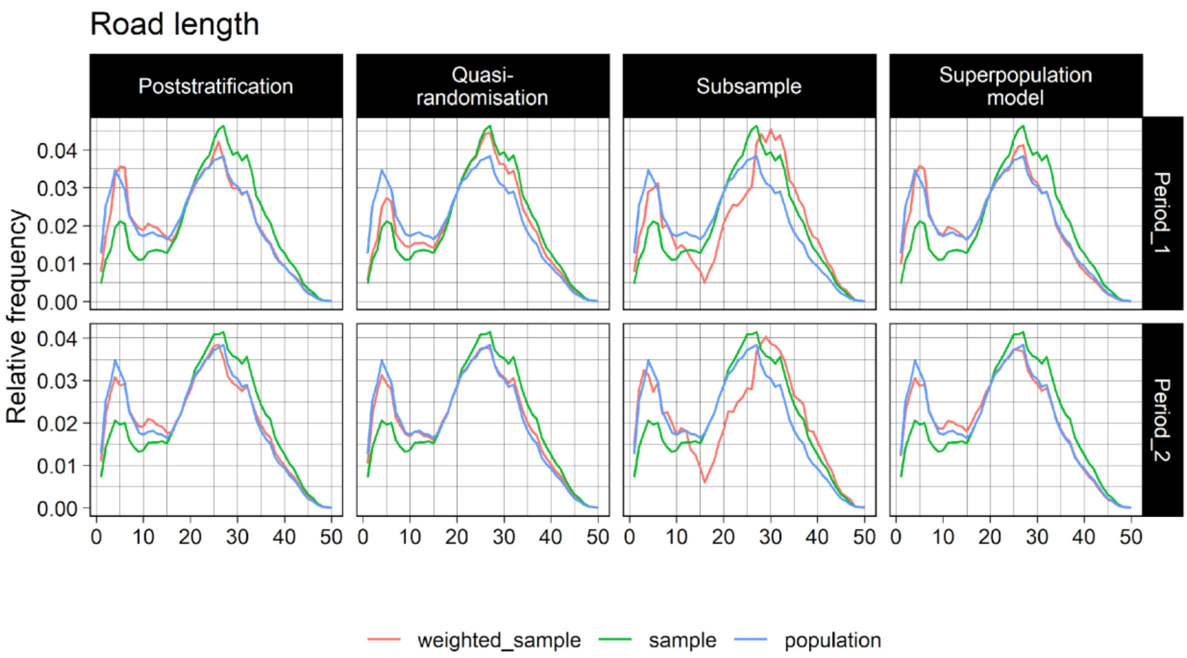

Figure 5. Sample, population and weighted sample distributions of the auxiliary variable road length (Table 1) in periods one and two.

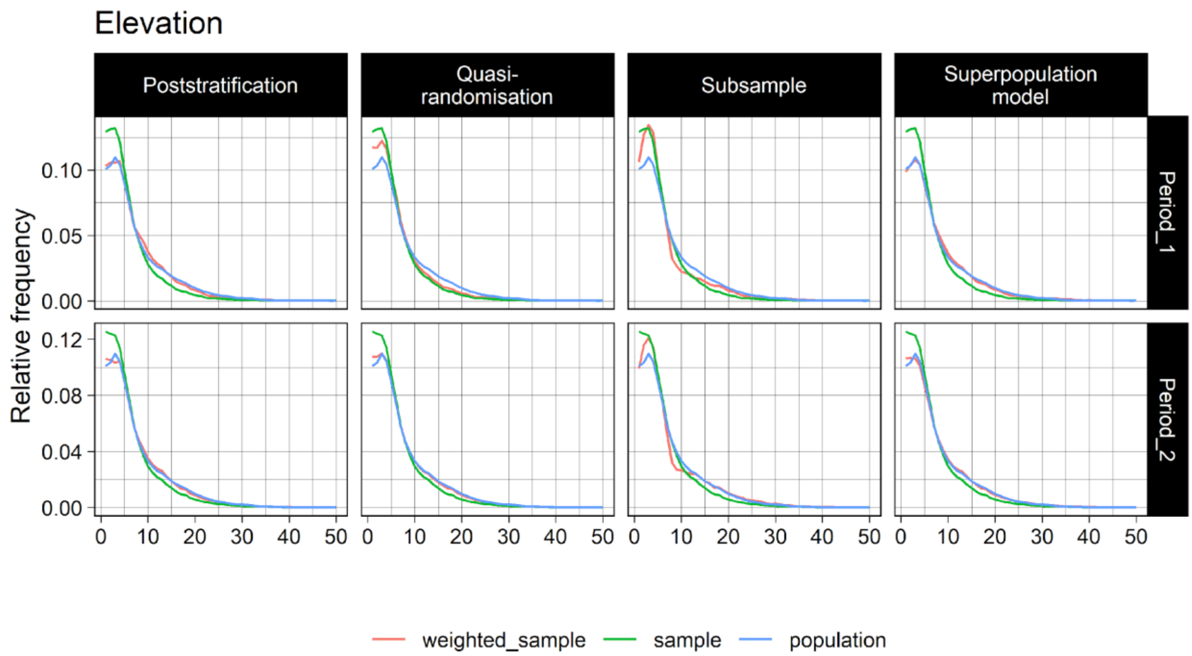

Figure 6. Sample, population and weighted sample distributions of the auxiliary variable elevation (Table 1) in periods one and two.

Discussion

Our experience is that analysts using large, nonprobability samples to monitor biodiversity tend not to account for issues of representativeness. Even where such issues are dealt with, there has been little acknowledgement of the broader panoply of relevant survey sampling methods available to the analyst, no exploration of how these are conceptually (or mathematically) linked and no comparison of their performance in realistic (i.e. relatively data poor) biodiversity monitoring situations. Evidence that a method can work in some discipline, or in simulation studies, is not proof that it will work in all situations. We have demonstrated how such adjustments might be applied using a realistic example of distribution change in a vascular plant over a period of 32 years. This example is realistic in that we do not have access to perfect predictors of occupancy or of sample inclusion. However, it is still likely to be closer to a best-case scenario than otherwise, due to the intense survey effort expended on vascular plants over the British landscape in the recent past (Stroh et al., 2023) and the fact that auxiliary data are relatively accessible in this area.

Our key finding is that the ability to bring the distributions of auxiliary variables in the sample closer to those in the population does not automatically mean that an adjustment will produce a more accurate estimate of a population quantity. For example, poststratification and superpopulation modelling were highly successful at redistributing the auxiliary variables in the samples (Figs 5 and 6). However, this did not translate into large improvements in the estimates of mean occupancy in each time-period or the trend (Figs 3 and 4). It must be the case that the auxiliary variables were not sufficient to describe the key differences between sample and population.

So, what makes a good auxiliary variable? Caughey et al. (2020) listed three criteria: 1) it should predict the response; 2) it should predict sample inclusion; and 3) its distribution in the population should be known. Four of the five auxiliary variables in our example were chosen on the basis that they predict sample inclusion, whereas only one was thought to be a reasonable predictor of the response (occupancy). Whilst it might seem like we prioritised the second criterion over the first, note that we originally included additional predictors of the response. These included soil pH, a known predictor of *C. vulgaris* occupancy (Stroh et al., 2023), and the first and second principal components of climate space in Britain, which we previously found to be important predictors of *C. vulgaris* habitat suitability (Boyd, Harvey, et al., 2023). Including these variables did not improve the estimates of mean occupancy or the trend, as we show in supplementary material 2. We suspect that these variables were redundant, because they are highly correlated with those in Table 1, so it is of little consequence which of these auxiliaries we included.

Identifying auxiliary variables that satisfy Caughey and colleagues' (2020) criteria is likely to be the most challenging part of adjusting samples in biodiversity monitoring. In many situations, predictors of the variable of interest and sample inclusion are unknown. Where they are known, data might not be available at the required scale (i.e. their distribution in the population is not known). To illustrate this point, consider the hoverfly *Criorhina asilica*, whose larvae require decaying timber from particular tree species (Stubbs & Falk, 2002). Without data on the locations of those decaying trees, it would likely be impossible to adjust for what is presumably a major determinant of its distribution. For taxa whose habitat requirements are well understood and reflected in available data (e.g. birds), selecting auxiliary variables should be simpler. Nevertheless, in practice, the analyst does not know the truth, so there will always be some guesswork (if this were not the case, statistical modelling would not be required). Transparency regarding availability and choice of auxiliary variables should be an important component of reporting for all biodiversity monitoring.

Whilst we are confident that the appropriateness of the auxiliary variables was the limiting factor in our example, it is possible that improvements to the estimators themselves could have improved matters. Potential improvements to MRP are most obvious. For example, interactions between the auxiliary variables could be included in the multilevel model (Ghitza & Gelman, 2013), and multiple time-periods could be modelled at once (Gelman et al., 2018). The question is whether fine-tuning models, which might make them more expensive to run (including interactions in MRP certainly does), is worth marginal gains 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.”

Some have questioned whether it is worth weighting nonprobability samples at all. In opinion polling, for example, there are many examples where weighting or other adjustments did little or nothing to improve the accuracy of inference from nonprobability samples, or even made matters worse (Bailey, 2023). In terms of the accuracy of the sample-based estimates, our results suggest that the situation in biodiversity monitoring is similar. Importantly, however, we have also showed that most adjustment methods do what they are supposed to: they turn an unrepresentative sample into a representative one, albeit strictly with respect to the chosen auxiliaries. First principles dictate that, if the auxiliaries are appropriate, this would translate into a more representative sample in terms of the variable of interest and improve the accuracy of inference. We see taxon experts as having a crucial role in identifying appropriate auxiliaries (e.g. Boyd, Harvey, et al., 2023; Smart et al., 2019).

It is worth commenting on how we measured the accuracy of the estimated trends. We compared the magnitudes of the estimated and true trends and assessed whether the estimates’ confidence/credible intervals covered the true value. Others defined accuracy as the power to “detect” a trend, whereby a method is considered successful if it gets the sign of the trend correct and its uncertainty interval does not span zero (e.g. Valdez et al., 2023). In this power setting, four of the estimators that we considered were able to detect the true trend, including the difference between the naive sample means.

We prefer to use the magnitude of the trend as a measure of accuracy for biodiversity monitoring, because many applications in this area are descriptive-inferential, not decision-theoretic (Greenland, 2022; Hurlbert et al., 2019; Oliver L. Pescott et al., 2019). That is to say, the final objective of exercises in biodiversity trend estimation is frequently a descriptive indicator, not a binary accept/reject conclusion of change or no change (e.g. Dennis et al., 2019; Powney et al., 2019). The link between Neyman-Pearsonian power and such exercises is often unclear (Amrhein et al., 2019):

they are essentially descriptive exercises and as such should be evaluated in terms of the closeness of the sample-based estimate to the truth, not merely in terms of rejecting (typically unrealistic) null hypotheses. The ability to report and consider uncertainty in the trend estimation is essential in making judgements about the risk of bias in biodiversity data (Boyd et al., 2022).

Also worth remembering is that we have only applied the adjustments to one species and using a relatively “good” dataset. It is plausible that the adjustments would improve estimates from a smaller or less representative dataset to a greater extent. They will certainly work better for species whose auxiliary variables are easier to identify and reflected in available data.

Repeating our analysis with other taxa and datasets would provide a better understanding of in what circumstances we can expect adjustments to perform well. The difficulty will be finding species whose true occupancy (or other variable of interest) is known. One option is to use simulations, but it is extremely important that they are not designed in such a way that the auxiliary variables explain sample inclusion and the variable of interest completely. In this case, the methods will all work very well, but that is not a true reflection of reality.

Our concluding message is that statistical adjustments might improve descriptive statistical inference in ecology, but only when combined with expert knowledge and appropriate data. Where there is doubt about the suitability of available auxiliary variables, a safer strategy is to assess the risk of bias qualitatively (Boyd et al., 2022; Meineke & Daru, 2021). If there is deemed to be a risk, it should be reflected in the way that findings are reported (Boyd, Powney, et al., 2023; O L Pescott et al., 2022). This might include using more conservative language and acknowledging that traditional uncertainty intervals are not guaranteed (or even likely) to cover the truth (X.-L. Meng, 2022).

Acknowledgement

Thank you to Richard Valliant, 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

Amrhein, V., Trafimow, D., & Greenland, S. (2019). Inferential Statistics as Descriptive Statistics: There Is No Replication Crisis if We Don’t Expect Replication. *American Statistician*, *73*(sup1), 262–270. https://doi.org/10.1080/00031305.2018.1543137

Bailey, M. (2023). *Polling at a crossroads: Rethinking modern survey research*. Cambridge University Press (forthcoming). https://doi.org/10.2753/RSL1061-1975230163

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., 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

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., 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, 1–11. https://doi.org/10.1111/ddi.13698

Boyd, R., 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., 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

Caughey, D., Berinsky, A., Chatfield, S., Hartman, E., Schickler, E., & Sekhon, J. (2020). *Target Estimation and Calibration Weighting for Unrepresentative Survey Samples.* Cambridge University Press.

Dennis, E. B., Brereton, T. M., Morgan, B. J. T., Fox, R., Shortall, C. R., Prescott, T., & Foster, S. (2019). Trends and indicators for quantifying moth abundance and occupancy in Scotland. *Journal of Insect Conservation*, *23*(2), 369–380. https://doi.org/10.1007/s10841-019-00135-z

Fick, S. E., & Hijmans, R. J. (2017). WorldClim 2 : new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*. https://doi.org/10.1002/joc.5086

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.

Gelman, A., Phillips, J., Lax, J., Gabry, J., & Trangucci, R. (2018). *Using Multilevel Regression and Poststratification to Estimate Dynamic Public Opinion*. http://www.stat.columbia.edu/~gelman/research/unpublished/MRT(1).pdf

Ghitza, Y., & Gelman, A. (2013). Deep interactions with MRP: Election turnout and voting patterns among small electoral subgroups. *American Journal of Political Science*, *57*(3), 762–776. https://doi.org/10.1111/ajps.12004

Greenland, S. (2022). Divergence versus decision P-values: A distinction worth making in theory and keeping in practice: Or, how divergence P-values measure evidence even when decision P-values do not. *Scandinavian Journal of Statistics*, *50*(1), 54–88. https://doi.org/10.1111/sjos.12625

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.

Hurlbert, S. H., Levine, R. A., & Utts, J. (2019). Coup de Grâce for a Tough Old Bull: “Statistically Significant” Expires. *American Statistician*, *73*(sup1), 352–357. https://doi.org/10.1080/00031305.2018.1543616

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

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. (2022). Double Your Variance, Dirtify Your Bayes, Devour Your Pufferfish, and Draw your Kidstrogram. *The New England Journal of Statistics in Data Science*, *0*, 1–20. https://doi.org/10.51387/22-nejsds6

Meng, X. (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.

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

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

ONS. (2021). *UK Office for National Statistics Postcode Directory*. https://geoportal.statistics.gov.uk/datasets/ons-postcode-directory-november-2022/about

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, Oliver L., Humphrey, T. A., Stroh, P. A., & Walker, K. J. (2019). Temporal changes in distributions and the species atlas: How can British and Irish plant data shoulder the inferential burden? *British & Irish Botany*, *1*(4), 250–282. https://doi.org/10.33928/bib.2019.01.250

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

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

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

Smart, S. M., Jarvis, S. G., Mizunuma, T., Herrero-Jáuregui, C., Fang, Z., Butler, A., Alison, J., Wilson, M., & Marrs, R. H. (2019). Assessment of a large number of empirical plant species niche models by elicitation of knowledge from two national experts. *Ecology and Evolution*, *9*(22), 12858–12868. https://doi.org/10.1002/ece3.5766

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

Stubbs, A., & Falk, S. (2002). *British Hoverflies*. British Entomological and Natural History Society.

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

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), 1–14. https://doi.org/10.1111/ecog.06604

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

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

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

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.

Van Swaay, Chris A.M., Nowicki, P., Settele, J., & Van Strien, A. J. (2008). Butterfly monitoring in Europe: Methods, applications and perspectives. *Biodiversity and Conservation*, *17*(14), 3455–3469. https://doi.org/10.1007/s10531-008-9491-4

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., & 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