How to reduce sampling error in species population monitoring: from theory to methods

^{1}Boyd, R. J., ^{2}Jarvis, S., ^{3}Meng, X-L., ^{1}Powney, G. D., ^{4}Spake, R., ^{1}Pescott, O.

^{1}UK Centre for Ecology and Hydrology, Maclean Building, Benson Ln, Crowmarsh Gifford, OX10 8BB

^{2}UK Centre for Ecology and Hydrology, Lancaster, Lancaster Environment Centre, Bailrigg, LA1 4AP

^{3}Department of Statistics, Harvard University, 1 Oxford St., Cambridge, MA 02138, USA

^{4}School of Geography and Environmental Sciences, University of Southampton, UK

Corresponding author email: robboy@ceh.ac.uk

Abstract

Progress towards national and international targets to halt and reverse declines in species' abundances will be assessed using *Multispecies Indicators* (MSIs). A distinction must be drawn between two MSIs. One is the *ideal*, but unobserved, MSI that would have been estimated had all species and sites within the scope of the target been sampled. The other is the empirical MSI estimated from the sample in hand. The discrepancy between the two, the *sampling error*, determines whether the empirical MSI faithfully reflects progress towards abundance targets.

We decompose the sampling error of common sample-based MSIs algebraically into a *geographic component* reflecting the effect of non-sampled sites and a *taxonomic component* reflecting the effect of non-sampled species. Building on established results from sampling theory, we further decompose each component into three contributing factors: the *data defect* (capturing the bias of the sampling process), the *data scarcity* (capturing the odds that species and sites are not sampled) and the *problem difficulty* (caused by variation in abundance across sites and species).

Having shown that both error components are determined by the same three factors, we review approaches to mitigating them. The approaches can be categorised broadly as obtaining more data, estimating the sample-based MSI in a different way (e.g. using a model), or redefining the 'target population'. The target population is effectively the spatial and taxonomic scope of the MSI, so redefining it *changes what is being estimated* and modifies the problem at hand. Hence, it can be justified only when an accurate answer to a different question (e.g. pertaining to a subset of species from the original population) is preferable to an inaccurate answer to the original one.

Key words: Biodiversity indicator; Data defect correlation; Essential Biodiversity Variable; Missing data; Species abundance; Sampling theory

Introduction

From a legislative perspective, world leaders have never been more committed to halting and reversing declines in species' abundances. In December 2022, parties to the Convention on Biological Diversity agreed on the latest Global Biodiversity Framework (GBF), which states that "the abundances of native wild species [should be] increased to healthy and resilient levels" (Convention on Biological Diversity, n.d.). Not long after, the UK and the European Union (EU) set a precedent by enshrining specific targets that echo this sentiment in law (DEFRA, 2024; European Commission, 2024). That species abundance targets are becoming enforceable is clearly a positive development for nature conservation, but it does mean that the evidence used to monitor progress towards those targets must stand up to scrutiny.

A common benchmark for monitoring progress towards species abundance targets is the Multispecies Indicator (MSI). MSIs have been defined in various ways (Freeman et al., 2021; Gregory & van Strien, 2010), but to us the term is best described as *an estimate of the 'average' rate of change in abundance, relative to some reference time, across a predefined set of species and geographic area*. A prominent example, which was recently reinstated as a 'component' indicator for monitoring progress towards the GBF, is the Living Planet Index (LPI; Collen et al., 2009; Loh et al., 2005). According to its website, the LPI measures the "the average rate of change in … population sizes of native [vertebrate] species" globally (ZSL & WWF, 2024). Other examples include the EU's grassland butterfly index and England's 'all species' index, which will be used to measure progress towards the respective governments' legal commitments (DEFRA, 2024; European Parliament, 2024).

MSIs have nominal spatial and taxonomic extents that should, in theory, align with the relevant species abundance target. Spatial extents might be defined in terms of, say, a country or administrative unit (or even globally in the case of the LPI), and they can be divided conceptually into areal units or 'sites' (e.g. grid squares on a map). Taxonomic extents are usually defined in terms of a set of species. In statistical parlance, the complete set of species × site combinations to which an MSI nominally pertains is known as the *target population* or simply the *population* (not to be confused with the ecological concept of a population).

Given the limited spatial and taxonomic coverage of biodiversity data (Gonzalez et al., 2016; Hughes et al., 2020; Meyer et al., 2016), it is likely that the set of sites and species for which abundance data are available will represent a small subset of the population. It follows that the MSI obtained using the data in hand is almost certain to differ from the one that would have been obtained had all species and sites in the population been sampled. To use more statistical language, the sample-based MSI is known as the *estimator*, and the population MSI is the target parameter or *estimand*. Since it is the estimand that is of interest, the hope is that the discrepancy between it and the estimator is small.

There are several ways to define the discrepancy between an estimator and its corresponding estimand. Here, we focus solely on the *sampling error* induced by sampling a subset of species and sites in the population. We do not consider *measurement error*, which occurs when the observed number of individuals of a given species across sampled sites is not directly proportional to the true numbers. Nor do we invoke hypothetical replicate samples to characterise the statistical properties of the estimator, such as its bias or variance. Throughout, we reason *conditionally on the sample in hand*—that is, given the data actually collected.

The purpose of this paper is to reveal the determinants of sampling error and what can be done to mitigate them. Since sampling error is caused by missing data, its determinants cannot be evaluated empirically. Rather, they must be identified on theoretical grounds or explored using in-silico experiments. We opt for a theoretical treatment because it does not require any assumptions about the underlying ecological processes and provides more general insights. That said, we acknowledge that simulations are useful for understanding sampling error (Guzman et al., 2022; Wilkes et al., 2025) and that they could always inspire—or be informed by—theory (Albert et al., 2010; Boyd et al., 2024).

The remainder of the paper is organised as follows. We begin by formalising the concept of the target population and specifying general mathematical expressions for the estimator and the estimand. This framework allows us to decompose the sampling error of the estimator algebraically into a geographic component reflecting the impact of non-sampled sites and a taxonomic component reflecting the impact of non-sampled species. Building on established results from sampling theory, we further decompose the geographic and taxonomic error components into three fundamental sources: the data defect, data scarcity and the problem difficulty. The final section reviews methods for reducing the magnitude each of these three quantities and hence for reducing the total sampling error of MSIs. A simulated example that demonstrates some of the points made throughout can be found at https://nerc-ceh.github.io/data-science-toolbox/methods/ds-toolbox-notebook-multispecies-biodiversity-indicators/msbi-error.html.

Theory

Life on Earth as a finite population

For a given time-period t, life on Earth—or any subset thereof—can be considered a statistical population comprising $j = 1, \dots, J$ species, $k = 1, \dots, K$ sites and $N = J \times K$ combinations thereof (hereafter 'Spatio-Taxonomic Units', or STUs). We assume for simplicity that species and sites are classified in the same manner regardless of the time-period. Each STU is characterised by its abundance Y_{jkt} (or e.g. biomass) and its occupancy (i.e. whether $Y_{jkt} > 0$). We do not impose a mathematical model for abundance and hence do not need to treat it as a random variable.

The sample

In any one time-period, data on abundance Y_{jkt} are available for a sample of the N STUs, K sites and J species in the population. We denote sample inclusion using a binary indicator R, where $R_{jkt} = 1$ if species j is sampled at site k in time-period t and 0 otherwise.

Two samples are defined. One is the set of species that were counted at least once at any site; we denote this set $s_t^J = \{j | \exists k \text{ such that } R_{jkt} = 1\}$. The other is the set of sites at which species j was counted, which we denote $s_{tj}^K = \{k | R_{jkt} = 1\}$.

The estimand and the estimator

The details differ, but the general approach to constructing a MSI is to average Y_{jkt} in two stages for each time-period: first across sampled sites for each species and then across species (Freeman et al., 2021). Assuming for now that the arithmetic mean is used at the first stage, the average abundance of species j across sampled sites in time-period t is

$$\bar{y}_{jt} = \frac{1}{n_{jt}^K} \sum_{k \in s_{jt}^K} Y_{jkt}, \tag{1}$$

where n_{jt}^K is the number of sites at which species j was sampled. It is common practice to convert $\bar{y}_{jt}$ to a relative index w_{jt} by dividing by its value in the first time-period (Buckland et al., 2011): that is,

$$w_{jt} = \frac{\bar{y}_{jt}}{\bar{y}_{j1}}. \tag{2}$$

There is no requirement that the same set of sites were sampled in time-periods 1 and t, but mean abundance may not be zero in either period (this “zero problem” has been covered elsewhere; (Toszogyova et al., 2024).

The geometric mean is typically used to average the relative abundance indices across species (Gregory & van Strien, 2010; McRae et al., 2017):

$$\bar{w}_t^J = \exp\left(\frac{1}{n_{1,t}^J} \sum_{j \in s_{1,t}^J} \ln(w_{jt}) \right), \tag{3}$$

where $s_{1,t}^J = s_1^J \cap s_t^J$ is the set of species sampled in both time-periods 1 and t and $n_{1,t}^J$ is the number of elements therein. (Assuming for now that there are no imputed values of Y, a point we come back to below, it is only those species sampled in periods 1 and t whose relative abundance indices are defined.) We will refer to $\bar{w}_t^J$ as *the per time-period estimator or simply the estimator*.

The reader should be aware that the sample averages in equations (1) and (3) are special cases of a broader class of estimators for population averages. Both can be expressed as weighted sums of the observed data, a form that encompasses a wide range of estimators used in practice (Boyd et al., 2023; McRae et al., 2017). This observation will be useful below, where we show that the error decomposition applies more generally than to the specific estimator defined by equations (1–3).

The LPI estimator is slightly different to one given by equations (1-3). Rather than representing the growth rate of annual average abundance, w_{jt} represents the average per-site growth rate across sites. Since it corresponds more closely to existing national biodiversity indicators, we focus on the error of the estimator given by equation 3. However, the error of the sample-based LPI decomposes in a similar manner (appendix C), so the general insights described in the remainder of the paper apply regardless of which of these estimators is used.

The population analogue of the per period estimator is

$$\bar{W}_t^J = \exp\left(\frac{1}{N_{1,t}^J}\sum_{j=1}^{N_{1,t}^J}\ln\left(W_{jt}\right)\right), \tag{4}$$

where $N_{1,t}^J$ is the total number of species in the population in both time-periods 1 and t, $W_{jt} = \bar{Y}_{jt}/\bar{Y}_{j1}$ is the population relative abundance index for species j, $\bar{Y}_{jt} = \sum_{i=1}^{N_{jt}^K} Y_{ijt} / N_{jt}^K$ is the population mean of Y for species j in time-period t, and N_{jt}^K is the total number of sites at which species j was sampled in period t. It is standard practice in statistics, and indeed in many areas of applied science, to define one's estimand before considering an estimator (Lundberg et al., 2021). Although this convention does not appear to be standard in biodiversity monitoring, we argue that *the use of a biodiversity indicator with a similar form to equation 3 strongly implies that $\bar{W}_t^J$ is the estimand*. What value $\bar{W}_t^J$ takes depends on the precise definition of the population, and we come back to this point below (also see Box 2).

Sampling error

Now let us consider the discrepancy between the estimand and the estimator. As defined here, MSIs reflect proportional change. Hence, it is natural to consider their relative (rather than absolute) sampling error, which is given by $(\bar{w}_t - \bar{W}_t)/\bar{W}_t = (\bar{w}_t/\bar{W}_t) - 1$. Focusing on $\bar{w}_t/\bar{W}_t$, since -1 is a constant and provides no insight into the determinants of error, we have from equations (3) and (4) that

$$\frac{\bar{w}_t^J}{\bar{W}_t^J} = \frac{\exp\left(\frac{1}{n_{1,t}^J}\sum_{j \in s_{1,t}^J}\ln\left(w_{jt}\right)\right)}{\exp\left(\frac{1}{N_{1,t}^J}\sum_{j=1}^{N_{1,t}^J}\ln\left(W_{jt}\right)\right)}. \tag{5}$$

Error decomposition

Equation 5 gives the relative sampling error of $\bar{w}_t^J$ as an estimator of $\bar{W}_t^J$ (up to an additive constant -1) but provides few direct insights into its determinants. In this section, we decompose the relative error algebraically into its component sources.

For ease of exposition, we begin with the estimator defined in equation (3). However, as we explained in the previous section, this estimator is a special case of a more general class that can be expressed as weighted sums of the observed data (i.e. linear estimators). By redefining certain terms, the same decomposition extends to this wider class of estimators (Meng, 2022), making it quite general. We return to these alternative estimators below, where we frame them as strategies to reduce the sampling error relative to the baseline estimator in equation (3).

The first step in the decomposition is to apply a log transformation. Doing so does not alter the determinants of the error; it simply re-expresses them on a scale that makes their components additive. The resulting expression separates the sampling error into geographic and taxonomic components (appendix A):

$$\ln\left(\frac{\bar{w}_t^J}{\bar{W}_t^J}\right) = \ln\left(\bar{w}_t^J\right) - \ln\left(\bar{W}_t^J\right) = \underbrace{\left(\frac{1}{n_{1,t}^J}\sum_{j\in s_{1,t}^J}\ln\left(W_{jt}\right) - \frac{1}{N_{1,t}^J}\sum_{j=1}^{N_{1,t}^J}\ln\left(W_{jt}\right)\right)}_{\substack{taxonomic\\ component}} + \underbrace{\frac{1}{n_{1,t}^J}\sum_{j\in s_{1,t}^J}\epsilon_{jt},}_{\substack{geographic\\ component}} \tag{6}$$

where $\epsilon_{jt} = \ln\left(w_{jt}\right) - \ln\left(W_{jt}\right)$ is the error of the log relative abundance index for species j and reflects the change in the distribution of the sample relative to the species' abundance over time.

The taxonomic error component is the difference between the sample and population means of $\ln\left(W_{jt}\right)$ across species and reflects the fact that some species may not have been sampled. The geographic component is the mean of ϵ_{jt} across sampled species. In the remainder of this section, we further decompose the geographic and taxonomic errors.

Taxonomic sampling error

To decompose the taxonomic error component, we can exploit an algebraic identity derived by Meng (2018). Assuming no measurement error, the identity shows that the difference between the sample and population means of an arbitrary variable in a finite population is the product of three fundamental quantities (defined below; also note that each of the quantities has a geographic analogue, which we also explain below). Applying Meng's decomposition to $\ln\left(W_{jt}\right)$, we have

$$\frac{1}{n_{1,t}^J}\sum_{j\in s_{1,t}^J}\ln\left(W_{jt}\right) - \frac{1}{N_{1,t}^J}\sum_{j=1}^{N_{1,t}^J}\ln\left(W_{jt}\right) = \underbrace{\rho\left(R_{1,t},\ln\left(W_{jt}\right)\right)}_{\substack{data\\ defect\\ correlation}} \underbrace{\sigma_{\ln\left(W_{jt}\right)}}_{\substack{problem\\ difficulty}} \underbrace{\sqrt{\frac{1-f_{1,t}}{f_{1,t}}}}_{\substack{data\\ scarcity}}. \tag{7}$$

The first quantity on the right-hand side, the data defect correlation $\rho\left(R_{1,t},\ln\left(W_{jt}\right)\right)$, is the correlation between $\ln\left(W_{jt}\right)$ and a binary variable $R_{1,t}$ taking the value 1 for species sampled in both periods 1 and t and 0 otherwise. A positive data defect correlation implies that $\ln\left(W_{jt}\right)$ is larger on average for sampled than non-sampled species and vice versa. The second quantity $\sigma_{\ln\left(W_{jt}\right)}$ is the population standard deviation of $\ln\left(W_{jt}\right)$ across species. It takes the value 0 when $\ln\left(W_{jt}\right)$ is a constant, in which case the sample mean is equivalent to the population mean regardless of which species were sampled. Hence, it can be considered a measure of "problem difficulty" (Meng, 2018), because the higher the variability of $\ln\left(W_{jt}\right)$, the harder it is to accurately estimate its population average. $f_{1,t}$ is the proportion of species in the population that were sampled in periods 1 and t, and $\sqrt{\left(1-f_{1,t}\right)/f_{1,t}}$ is a measure of data scarcity.

Before going further, it is worth pointing out that general structure of equation (7) has been known to (survey) statisticians for some time. Others have tended to express it in terms of sample inclusion probabilities rather than the binary sample inclusion indicator (Aubry et al., 2024; Schouten, 2007), in which case it gives the expected error of the sample mean (i.e. its bias). Nevertheless, we refer to the identity as the "Meng expression", since (Meng, 2018, 2022) formalised its general form and clarified its implications for a wide class of estimators and data collection mechanisms.

Geographic sampling error

We now turn to the geographic sampling error component. Recalling that $\bar{y}_{jt}$ is the mean abundance of species j across sampled sites in time-period t and that $\bar{Y}_{jt}$ is its population equivalent, the geographic sampling error for species j can be expressed as (appendix B)

$$\epsilon_{jt} = \ln\left(1 + \frac{\bar{y}_{jt} - \bar{Y}_{jt}}{\bar{Y}_{jt}}\right) - \ln\left(1 + \frac{\bar{y}_{j1} - \bar{Y}_{j1}}{\bar{Y}_{j1}}\right). \tag{8}$$

Equation (8) yields two insights. One is that the differences between the sample and population mean abundances for species j in time-periods t and 1 feature in the numerators on the right-hand side. Consequently, the Meng expression in equation (7), which gives the difference between sample and population means, can be applied to the geographic sampling error component (see below). The second insight is that the geographic sampling error for a given species reflects not only the discrepancy between the sample and population mean abundances in time-period t but also how this discrepancy differs from time-period 1. Hence, while we use the term "geographic error" for convenience, it could just as easily be described as the "spatio-temporal error component". We further discuss equation (8) and its implications for how to reduce the geographic sampling error component in the next section.

Applying Meng's decomposition to the differences between the sample and population mean abundances for species j in time-period t, we have

$$\bar{y}_{jt} - \bar{Y}_{jt} = \rho(R_{jt}, Y_{jt})\, \sigma_{Y_{jt}} \sqrt{\frac{1 - f_{jt}}{f_{jt}}}. \tag{9}$$

Like in equation (7), the three quantities on the right-hand side of equation 9 are, respectively, the data defect correlation, the problem difficulty and a measure of data scarcity (see Fig. 1 for a graphical representation of each component). The quantities' meanings are subtly different to their taxonomic counterparts, because R_{jt} indicates whether a site—rather than a species—was sampled for species j in time-period t, f_{jt} is the proportion of sites at which species j was sampled in time-period t and $\ln(W_{jt})$ has been replaced by the abundance of species j in period t, Y_{jt}. Hence, the geographic data defect correlation indicates whether the focal species is more abundant on average at sampled than non-sampled sites, and the geographic problem difficulty is the variability of the species' abundance across geographic units within a given time-period.

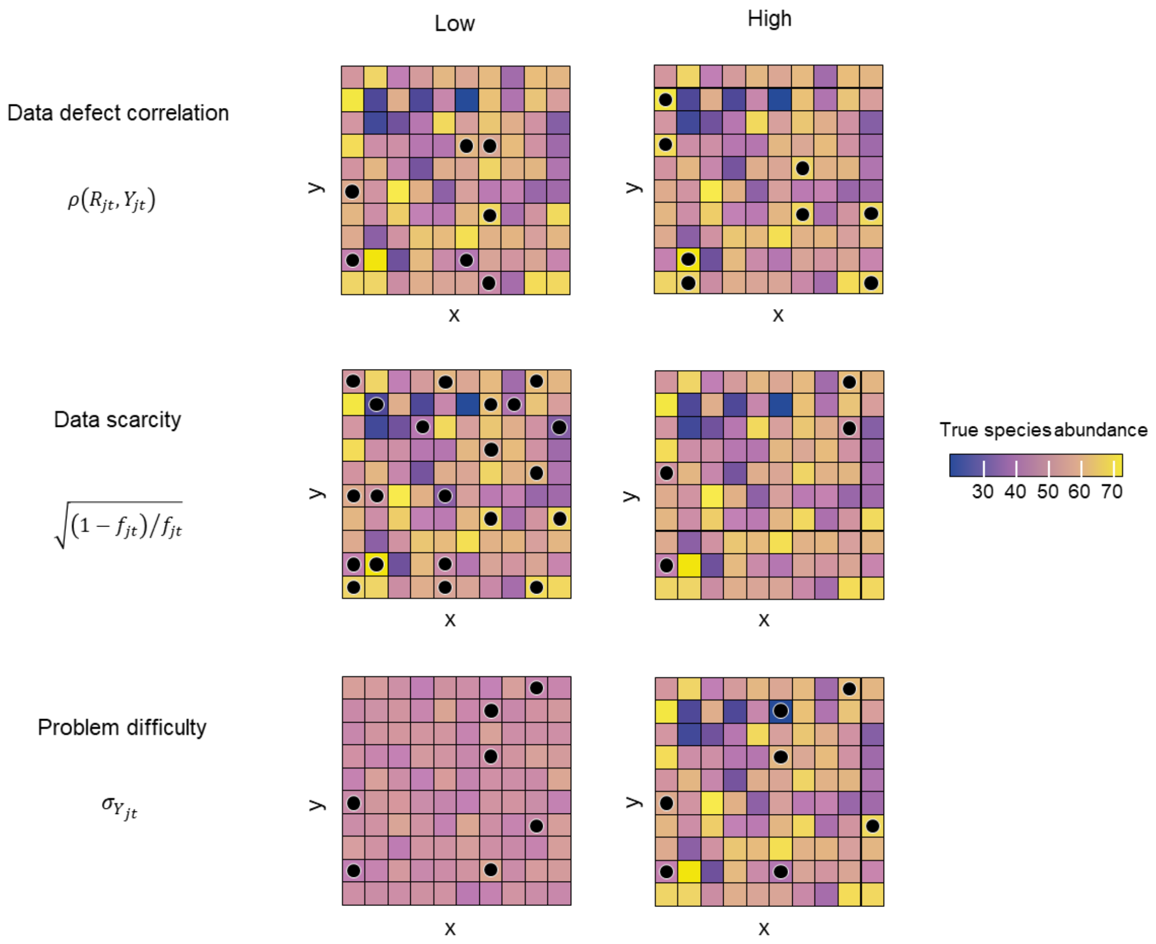


Figure 1. Six grids depicting 100 sites. Each grid shows either a high or low value (left to right) of the geographic data defect correlation, the data scarcity or the problem difficulty (top to bottom rows). Note that in the top right panel, where the data defect is high, it is only sites with high abundance that have been sampled. Mathematical notation used elsewhere in the paper for each quantity is also provided.

How to reduce sampling error

Insights from the decomposition

Equations 6 through 9 tell us how to reduce the taxonomic error, the geographic errors and, consequently, the total sampling error of an MSI. (We consider the related problem of how to *assess* potential estimation error in Box 1.)

Box 1. How to assess the potential sampling error of a Multispecies Biodiversity Indicator (MSI).

The potential sampling error of an MSI determines whether mitigating action is needed. To understand the potential for error, we require information on the geographic and taxonomic data defect correlations, data scarcities and problem difficulties (see equations 7 and 9 and refer to Fig. 1). The data scarcities reflect the proportions of species and sites in the population that have not been sampled, and they are measurable (assuming the total number of species is known). The data defect correlations and problem difficulties are not directly measurable and must be estimated or qualitatively assessed.

Assessing the data defect

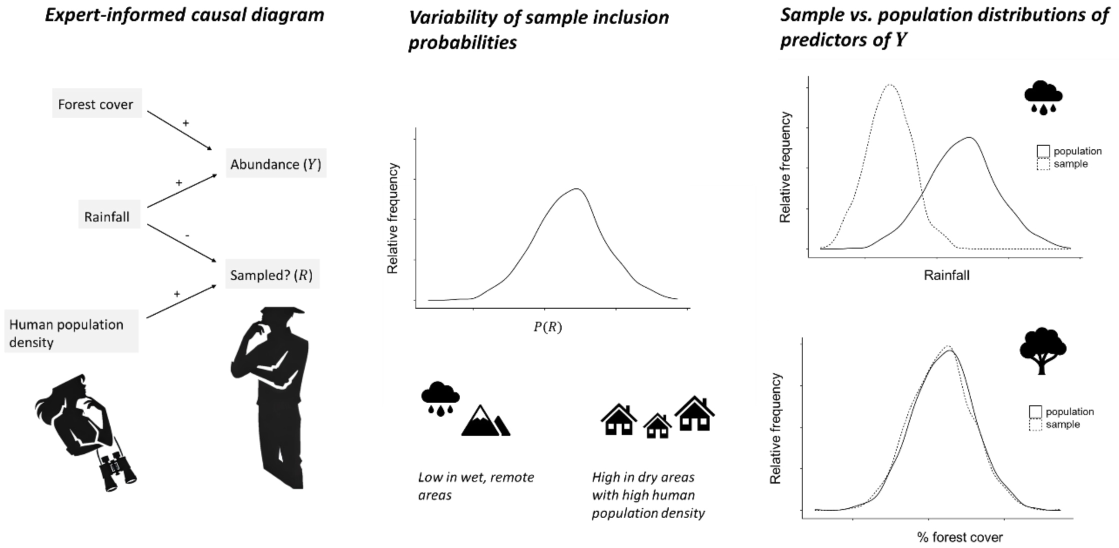
We are aware of three general approaches to assessing the potential for a non-negligible data defect correlation.

The first approach leverages the existing machinery of causal diagrams and the 'd-separation' algorithm, which are widely used in causal inference (Pearl et al., 2016). For notational simplicity, we will not index the time-period, will let R be sample inclusion (which could be species or site inclusion) and will let Y be the variable of interest (which could be abundance or a relative abundance index). The idea is to construct a causal diagram depicting causes and effects of R and Y; given the structure of the diagram, the d-separation algorithm determines whether two are dependent and thus whether we might expect a non-negligible data defect correlation (Boyd et al., 2025; Thoemmes & Mohan, 2015).

The second approach is to estimate sample inclusion probabilities $P(R = 1)$ and to calculate their variability in the population (e.g. Schouten et al., 2012). If the variability of $P(R = 1)$ is small, then R and Y can only covary so much, and the data defect correlation is likely to be small (Aubry et al., 2024; Nishimura et al., 2016).

*The third approach is to assess covariate balance***.** The idea is to identify variables that are predictive of Y and whose distributions in the population are known and to compare their sample and population distributions (Backstrom et al., 2024; Boyd et al., 2023a; cf. Makela et al., 2014). A mismatch signals that sampling was more or less likely at different levels of the predictor, which suggests a non-negligible data defect correlation.

Box Fig. 1 summarises the three approaches to assessing data defect correlations in the context of species population monitoring.



Expert-informed causal diagram

Variability of sample inclusion probabilities

Sample vs. population distributions of predictors of Y

Box figure 1. Schematic illustrating how one might diagnose a non-negligible geographic data defect correlation for a given species (the sample principles apply across species). It depicts a simple hypothetical situation in which rainfall is a common cause of sample inclusion (negative effect) and abundance and induces a non-negligible (data defect) correlation between the two. Forest cover and human population density solely affect abundance and sample inclusion, respectively, and do not contribute to the data defect correlation.

Each of the three approaches to assessing the data defect correlations could presented as part of a "risk-of-bias" assessment (Pescott et al., 2023). Risk-of-bias assessment comprise a series of questions about the potential for sampling bias, which is very closely related to the data defect correlation (sampling bias being proportional to its expected value). One risk-of-bias tool, ROBITT, was designed specifically for the purpose of biodiversity monitoring (Boyd, Powney, et al., 2022).

Assessing the problem difficulty
Approaches to estimating the problem difficulty (the standard deviation of Y) can also be imagined. One approach might be to identify predictors of Y whose population distributions are known and to calculate their variability. For example, Y might be a species' abundance, and the predictor might be habitat type. If the population is variable in terms of habitat, and habitat is predictive of abundance, then we would expect abundance to be variable too.

It is easiest to see how the taxonomic error can be reduced, because it is simply the difference between the sample and population means of $\ln(W_{jt})$ across species, which is given by the Meng expression. The Meng expression shows that the error is the product of the data defect correlation, the data scarcity and the problem difficulty. Consequently, it reduces to zero when any of those quantities is zero; all else being equal, reducing any of the quantities will also reduce error (although note that the quantities cannot vary independently in practice).

Reducing the geographic error for any given species (equation 8) is best achieved by reducing the per period estimation errors given by equation (9) in time-periods 1 and t. It is true that one could get lucky and that the per period errors could have the same signs and similar magnitudes, in which case the geographic error would be small. However, given that the error in any one period generally cannot be known, a better strategy is to aim for zero error in both periods. Since the per period errors can be expressed using Meng's decomposition, reducing the (geographic) data defect correlation, data scarcity and problem difficulty will reduce the per period errors and thus the geographic error for a given species.

The total log relative sampling error is the sum of the taxonomic and geographic components (noting that the geographic component reflects a mean across sampled species). It is theoretically possible to have zero or negligible error if the two components cancel (i.e. if one is positive and the other is negative). How the analyst would know they are in this situation is unclear, however, so a more sensible approach is to try to minimise both error components. As we have seen, minimising the within- and taxonomic errors means reducing the taxonomic and geographic data defect correlations, problem difficulties and data scarcities (the latter being equivalent to maximising the sampling fraction).

Problem preserving versus problem-modifying approaches

We have now seen that to reduce the sampling error of an MSI is to reduce one or more of the three quantities in the Meng expression, whether their geographic or taxonomic variants. Approaches to reducing these quantities fall in one of three broad categories: obtaining new data, replacing the sample-based MSI with an alternative estimator or redefining the target population. Each type can help to address more than one quantity in the Meng expression, as illustrated in Table 1.

Redefining the target population means modifying the estimand and hence the problem at hand. As such it can be justified only on the basis that obtaining an accurate answer to a different question is preferable to obtaining an inaccurate answer to the original one. Neither obtaining more data nor opting for an alternative estimator change the problem in this sense, since they generally do not affect the estimand.

Modifying the estimator nevertheless warrants more discussion. The decomposition in the previous section assumes the particular estimator defined in equation (3). As we explained in that section, however, the estimator in equation (3) is part of a wider class that can be expressed as weighted sums of the observed data (i.e. estimators that are linear in the observed data). The decomposition applies in structure, albeit after redefining some quantities, to any estimator in this class (Meng, 2018, 2022).

(There are inferential frameworks in which modifying the estimator necessarily modifies the estimand. One example is the estimating-equation framework. See appendix D for more on the relationship between the estimator and the estimand in this framework and its implications for our decomposition.)

Starting with the geographic variants, we review approaches to reducing the data defect, the problem difficulty and the data scarcity in the remainder of this section. See Table 1 for an overview, which indicates whether each approach modifies the original problem.

Table 1. A non-exhaustive list of approaches to reducing the sampling error of a Multispecies Indicator (MSI). The high-level approach is listed in column one: obtaining more data, modifying the estimator or redefining the target population. Column two lists the more specific approach within each higher-level class. The error component(s) targeted by each approach are listed in column three. Column four indicates whether the approach modifies the estimand and therefore the problem at hand. The mechanism(s) by which the relevant error components are reduced are described in column five. Column six lists the assumptions that must hold for a reduction in the error component to be achieved or, for those approaches that redefine the target population, for the new question to remain valid.

High-level approach	Specific approach	Error component(s) targeted	Problem modified?	Mechanism(s)	Condition required
Obtain more data.	Collect new data.	Sampling fraction, data defect.	No, if the new data are collected from the same target population	Increase coverage of target population. Adaptive sampling of underrepresented strata might reduce data defect.	Data defect given new data is smaller than before.
	Mobilise historic data.	Sampling fraction, data defect.	No, if the target population has not changed over time	As above but for historic time-periods.	Data defect given newly mobilised data is smaller than before.
Switch to an alternative estimator	Quasi-randomisation (i.e. propensity score weighting)	Data defect.	No, if the altered estimator does not lead to an altered estimand	Diminishes variability of sample inclusion propensities via weighting. Balances covariates between sample and population.	Conditional data defect given covariates is smaller than the unconditional one.
	Superpopulation model	Data defect, problem difficulty.	No.	Including confounders of sample inclusion and the response reduces the data defect; including predictors of the response reduces	Conditional data defect given covariates is smaller than the unconditional one. Likewise

				the problem difficulty.	for the problem difficulty.
Redefine the target population	Coarsen the spatial resolution.	Sampling fraction, problem difficulty	Yes.	Generally lowers variability in abundance and growth rates.	Aggregation preserves the usefulness of the estimand.
	Condition target population on occupied sites.	Problem difficulty, sampling fraction.	Yes.	Removing zeros lowers population variability. Might increase sampling fraction if occupied sites were preferentially sampled.	New target population is relevant to inferential goal.
	Condition target population on sampled sites.	Sampling fraction.	Yes.	Geographic sampling fraction becomes 1.	New target population is relevant to inferential goal.
	Condition target population on a subset of species.	Data defect, problem difficulty sampling fraction.	Yes.	Reduces data defect if sample inclusion becomes less correlated with growth rates and problem difficulty if growth rates become less variable.	New target population is relevant to inferential goal.

Geographic sampling error

Minimising the data defect correlation

The key to reducing the geographic data defect correlation for species j in time-period t, $\rho(R_{jt}, Y_{jt})$, is to recognise that its conditional value once some variable or set of variables is held constant (i.e. stratified on or "adjusted for"; we come back to how this is achieved in practice below) might be smaller than its unconditional value when they are not. More formally, there usually exists a set of variables $\boldsymbol{X}$ (or some other observed information) that satisfies $|\rho(R_{jt}, Y_{jt}|\boldsymbol{X})| < |\rho(R_{jt}, Y_{jt})|$. The first step towards reducing $\rho(R_{jt}, Y_{jt})$ is to identify these variables.

The variables that satisfy $|\rho(R_{jt}, Y_{jt}|\boldsymbol{X})| < |\rho(R_{jt}, Y_{jt})|$ *when included in* $\boldsymbol{X}$ *are generally the ones that induced the (data defect) correlation between whether sites were sampled* R_{jt} *and abundance* Y_{jt} *in the first place.* Often, these variables will be direct common causes of the two. For example, abundance Y_{jt} might be larger within protected areas, as they tend to be relatively well managed for species (Cooke et al., 2023). Likewise, data collectors might preferentially visit protected areas in the hope of seeing wildlife. In this case, when both R_{jt} and Y_{jt} are greater within protected areas, $\rho(R_{jt}, Y_{jt}) > 0$. For a given level of protected area status (e.g. inside or outside), however, the (conditional) value of $\rho(R_{jt}, Y_{jt})$ should be smaller than its value across all sites, which is to say $\rho(R_{jt}, Y_{jt}|\boldsymbol{X}) < \rho(R_{jt}, Y_{jt})$.

Variables that are not direct common causes of R_{jt} and Y_{jt} can also induce a non-zero data defect correlation, so the "common cause principle" (Mathur et al., 2023) will not always suffice. A more formal and comprehensive (but laborious) approach to identifying the variables that should be included in $\boldsymbol{X}$ is to construct causal diagrams (see Pearl et al., 2016) depicting causes and effects of R_{jt} and Y_{jt} (Boyd et al., 2025; Thoemmes & Mohan, 2015; Box 1). We will not go into the theory behind causal diagrams; the important point is that it is possible to deduce from their structures the sets of variables that induce a dependence between R_{jt} and Y_{jt} and potentially a (data defect) correlation. As we saw earlier, it is the variables that induce a non-negligible data defect correlation that should be included in $\boldsymbol{X}$, so causal diagrams are a good way to identify them. Critically, however, the use of a causal diagram supposes that it is a true reflection of reality, which is difficult to verify in practice (Grace & Irvine, 2020), and it generally provides no information on the form of the relationships between $\boldsymbol{X}$, Y_{jt} and R_{jt}.

Once the variables in $\boldsymbol{X}$ have been identified, the next step is to account for or 'condition on' them in the hope that it reduces $\rho(R_{jt}, Y_{jt})$. One option is to replace the arithmetic mean used to estimate $\bar{Y}_{jt}$ in equation 1 with a *weighted* sample mean, where the weights are selected in such a way that they balance the variables in $\boldsymbol{X}$ between sample and population (i.e. propensity score weighting a.k.a. quasi-randomisation; Boyd et al., 2023; Fink et al., 2023; McRae et al., 2017). Another is to impute values for Y_{jt} given $\boldsymbol{X}$ and to estimate $\bar{Y}_{jt}$ from the complete dataset obtained by combining the observed and imputed values (i.e. "superpopulation modelling"; Dorfman & Valliant, 2005). More complex approaches are available (e.g. Ghitza & Gelman, 2013), but we will not consider them here.

Equation 9, which gives the error of the sample mean of Y_{jt} as an estimator of its population mean, can be modified to give the error of both the weighted mean and the superpopulation model estimate. For the weighted mean, $\rho(R_{jt}, Y_{jt})$ is replaced by $\rho(\tilde{R}_{jt}, Y_{jt})$, where $\tilde{R}_{jtk} = R_{jtk}\, W_{jtk}$, and W_{jtk} is the weight applied to site k (Meng, 2018). The data scarcity term also needs to be adjusted to account for the fact that weights reduce the 'effective' sample size, but this too is a simple modification (Meng, 2022). To obtain the error of the superpopulation model estimate, the key is to substitute the model's residuals $Z_{jt} = Y_{jt} - m(\boldsymbol{X})$ for Y_{jt}, including those hypothetical residuals for non-sampled STUs (Meng, 2022). Switching the focus from Y_{jt} to the model's residuals means that $\rho(R_{jt}, Y_{jt})$ is replaced by $\rho(R_{jt}, Z_{jt})$, which indicates whether the model is better fit for sampled than non-sampled sites. Given a judicious choice of $\boldsymbol{X}$, weighting and imputation should ensure that $|\rho(\tilde{R}_{jt}, Y_{jt})| < |\rho(R_{jt}, Y_{jt})|$ and $|\rho(R_{jt}, Z_{jt})| < |\rho(R_{jt}, Y_{jt})|$, respectively.

In practice, the analyst will not possess knowledge of and data on all variables that should be included in $\boldsymbol{X}$, so alternative types of information might be conditioned on (e.g. used to construct weights or included in a superpopulation model). One practical option is to exploit shared autocorrelation between R_{jt} and Y_{jt} induced by autocorrelation in $\boldsymbol{X}$. Adjusting for shared autocorrelation between R_{jt} and Y_{jt} (e.g. by including autocorrelation terms in a superpopulation model) moves one closer to rendering the two uncorrelated and potentially even independent (Diggle et al., 2010). Most examples of this approach in ecology have focused on spatial autocorrelation (Mostert & O'Hara, 2023; Seaton et al., 2024; Simmonds et al., 2020), but Johnson et al. (2024) recently extended the idea to account for spatial, temporal and phylogenetic autocorrelation simultaneously (this approach could also help to deal with the taxonomic data defect correlation in some circumstances, as we explain below).

Increasing the sampling fraction (reducing the data scarcity)

One obvious way to reduce the data scarcity—or, equivalently, to increase the geographic sampling fraction f_{jt}—is to obtain data on sites for which no data was previously available. Since biodiversity indicators measure historic change in species' populations, the effects of collecting new data will not be seen for some years. Mobilising previously inaccessible historic data, however, could have an immediate impact (e.g. Ellwood et al., 2015).

When obtaining data for previously unsampled sites, there is a risk of inadvertently increasing the data defect correlation $\rho(R_{jt}, Y_{jt})$. Indeed, Boyd et al. (2022) showed that adding newly digitised data on bee distributions in Chile to Global Biodiversity Information Facility increased some measures of sampling bias [and hence the expected value of $\rho(R_{jt}, Y_{jt})$]. Following an adaptive sampling plan that explicitly targets a reduction in $\rho(R_{jt}, Y_{jt})$, for example by prioritising underrepresented strata, may be one way to guard against this issue (Pescott et al., 2025; Schouten & Shlomo, 2017)

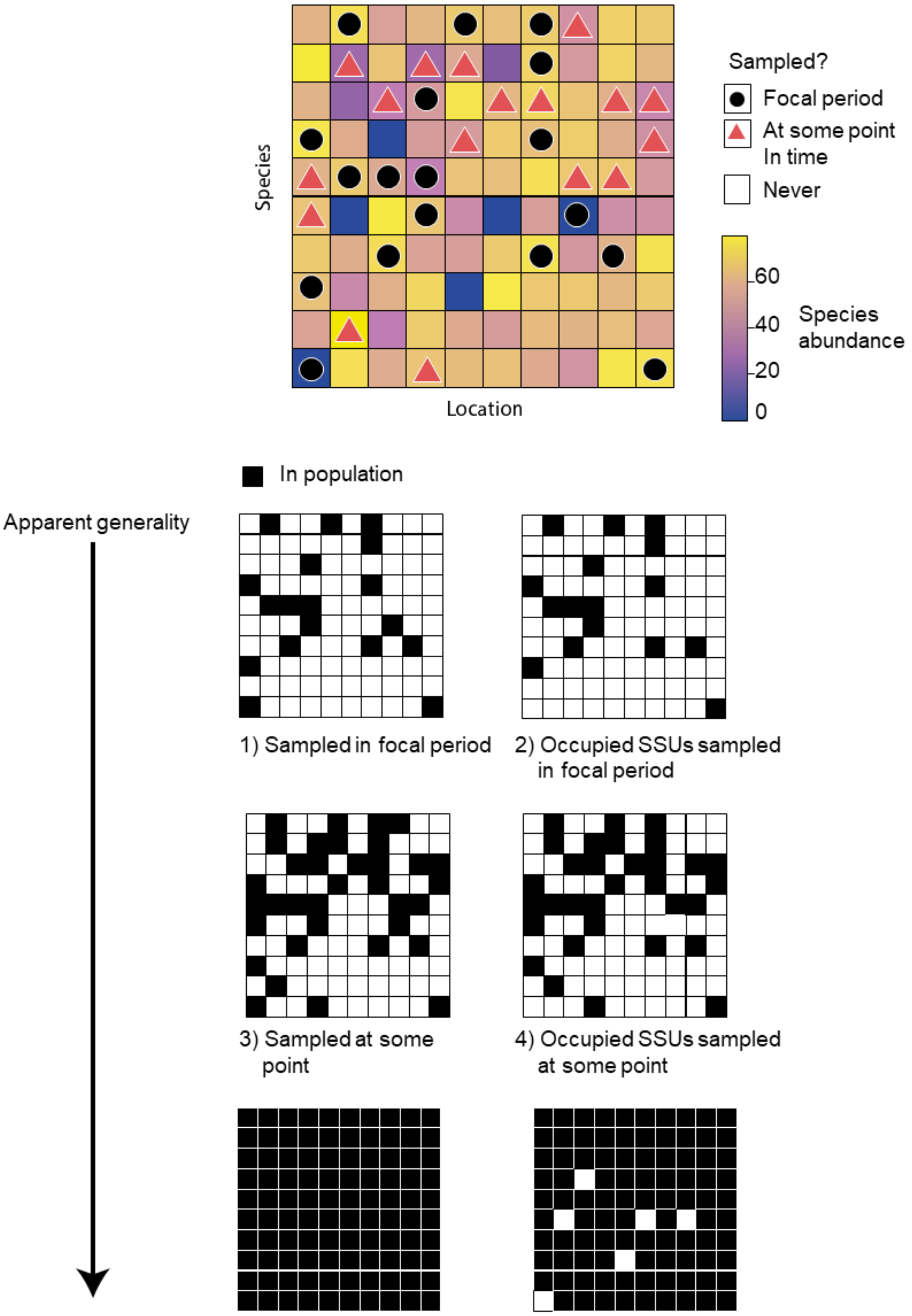
A second and much simpler way to increase f_{jt} is to recognise that the population need not include every site and to constrain it from the outset. Conditioning on (i.e. restricting the population to) the set of sampled geographic units for a given species, for example, means that $f_{jt} = 1$, the data scarcity term $\sqrt{(1 - f_{jt})/f_{jt}} = 0$ and, consequently, that the geographic estimation error $\rho(R_{jt}, Y_{jt})\, \sigma_{Y_{jt}} \sqrt{(1 - f_{jt})/f_{jt}} = 0$. Conditioning on occupied sites (either occupied in the focal time-period or in some time-period since monitoring began), too, could increase f_{jt}. Data collectors are usually interested in seeing wildlife as opposed to recording absences, so it is reasonable to suppose that, on average across species, occupied geographic units are more likely to have been sampled than unoccupied ones.

Of course, modifying the target population means modifying the estimand and changing the problem at hand. Conditioning on occupied or sampled sites reduces the number of STUs in the population and therefore the generality of the MSI. Doing so could be problematic if, say, it means omitting a species or geographic area that is relevant to a species abundance target. A reviewer pointed out a special case of this problem that deserves mention: species' range expansions would not affect the MSI if the target population were conditioned on sites that were occupied before that expansion took place. See Box 2 for more on the implications of conditioning the target population.

Box 2. Six ways to define the target population in each time-period. The list is not exhaustive, and other definitions could be imagined.

For a given set of species, geographic area and time-period, the population need not include every possible Spatio-Taxonomic Unit (STU). Rather, we might consider a conditional target population given, say, occupancy O_t or sample inclusion R_t (or indeed other variables such as habitat). Conditioning on $R_t = 1$ means focusing on sampled species and sites, and conditioning on $O_t = 1$ means ignoring STUs with zero abundance. Of course, which sites are occupied by a given species is generally not known and would have to be estimated based on, say, the presence of relevant habitat. We explain in the main text why conditioning on R and O might reduce error, but the analyst must also recognise that modifying the target population means modifying the estimand and therefore the problem at hand (Table 1).

Constraining the population can be done on a per period or cross-period basis: that is, we can condition on $O_t = 1$ and $R_t = 1$ or on $O_{1,t} = 1$ and $O_{1,t} = 1$, respectively. Since MSIs reflect change in abundance between two time-periods, it is perhaps most natural to condition the population on a cross time-period basis, in which case it does not change over time. If we condition the population on O or R on a cross time-period basis, it can change over time. From a mathematical perspective, one may not condition on $R_t = 1$ or O_t=1 on a per time-period basis if it means that there is a different set of species in time-period 1 to time-period t. Doing so would invalidate the relative abundance indices, since they require a defined abundance for any given species in both time-periods. From a conceptual perspective, defining the population in such a way that it can vary over time means that the error is not defined with respect to a clear reference population and partly reflects shifts in which sites are included in the population (noting again that the set of species must remain constant between periods). Box Fig. 2 depicts six possible definitions of the population depending on whether it is unconditional, conditioned on O across time-periods, conditioned on R across time-periods or conditioned on R for each time-period.



1) Sampled in focal period

2) Occupied SSUs sampled in focal period

3) Sampled at some point

4) Occupied SSUs sampled at some point

5) All SSUs

6) All occupied SSUs

Box figure 2. Six definitions of the target population for a given time-period. Each grid represents the total set of site × species combinations, or STUs, that might be considered. Black cells in the smaller grids represent the set that are considered under each definition of the population. In the top grid, cells with black circles were sampled in the focal time-period, and cells with red triangles were not sampled in the focal period but were sampled at some point (i.e. another period).

Reducing the problem difficulty

One approach to reducing the problem difficulty is covariate adjustment. The idea is to construct a model of abundance Y_{jt} given some covariates $\boldsymbol{X}$. In this setting, the problem difficulty is no longer the standard deviation of Y_{jt}, $\sigma_{Y_{jt}}$, but the standard deviation of the model's residuals $\sigma_{Z_{jt}}$ (Meng, 2022). If $\boldsymbol{X}$ explains a portion of Y_{jt}, then $\sigma_Z < \sigma_Y$, which is to say the problem difficulty has been reduced. $\boldsymbol{X}$ might include, say, land cover or environmental variables, for which high-resolution data are available globally (and therefore for any conceivable target population; Fick & Hijmans, 2017). Other estimators that condition on or "account for" $\boldsymbol{X}$ (e.g. poststratification) can reduce the problem difficulty for similar reasons (Lohr, 2022).

Another potential way to reduce the geographic problem difficulty is to modify the spatial resolution at which the analysis is conducted. For example, Boyd, Bowler, et al., (2024) showed that coarsening the resolution at which species occupancy is estimated can reduce the problem difficulty and reasoned on theoretical grounds that the same is likely to be true of abundance. Of course, for a given problem difficulty, estimates of species occupancy or abundance may be less practically useful at coarser resolutions, so there is a trade-off between potential error and the perceived usefulness of any given estimate across scales. This is known as the relevance–robustness trade-off for multi-resolution inference (Liu & Meng, 2016), a manifestation of the well-known bias-variance trade-off.

A third approach to reducing the problem difficulty is to condition the population on (i.e. restrict it to) the set of occupied sites for which $Y_{jt} > 0$. Assume as an example that Y_{jt} follows a zero-inflated Poisson (ZIP) distribution, which separates sites into 'structural' zeros governed by a Bernoulli distribution and counts governed by Poisson distribution. Now let q be the proportion of sites that are not structural zeros (i.e. occupied sites). When we do not condition on occupied sites, the problem difficulty is $\sqrt{\mu^2 q(1-q) + \mu q}$, where μ is the mean abundance across occupied sites. When we do condition on occupied sites, then the problem difficulty is $\sqrt{\mu}$. The difference between the two is $D = \sqrt{\mu} - \sqrt{\mu^2 q(1-q) + \mu q}$. For most levels of q and μ (when $q > 1/\mu$ to be precise), $D < 0$, which is to say that conditioning on occupied sites reduces the problem difficulty (Fig. 2).

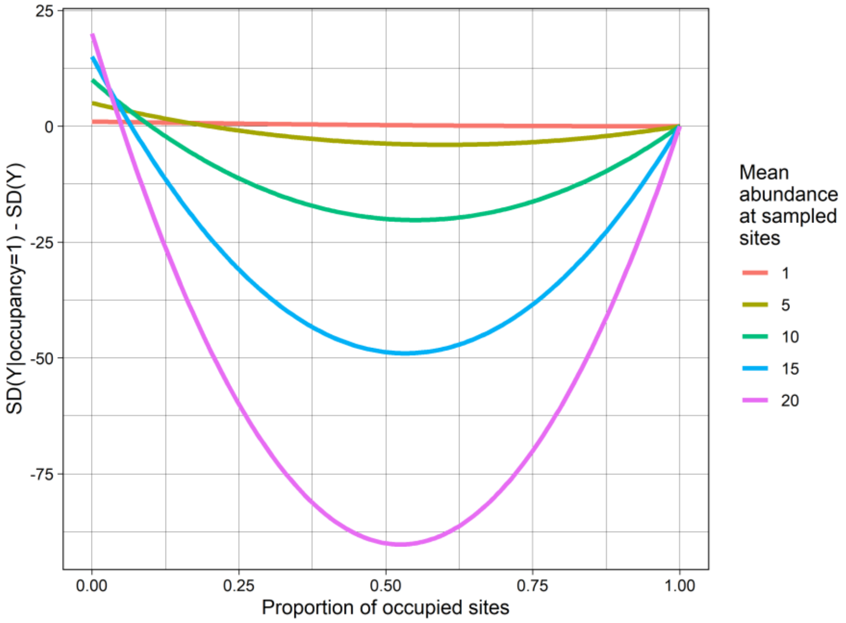


Figure 2. Difference in the problem difficulty (population standard deviation of abundance) when the population is defined as occupied sites only and when it includes all sites. Negative values indicate that omitting unoccupied sites from the population reduces the problem difficulty. Each curve represents one value of mean abundance across occupied sites. The results in this figure assume a zero-inflated Poisson model for abundance.

Another way to modify the population, which could also reduce the geographic problem difficulty, is to condition on sites with certain environmental conditions. Species' abundances tend to vary between environments and habitats. Conditioning on sites that fall within certain environmental strata may therefore reduce its variability in the population.

Taxonomic sampling error

Many of the principles described above apply to minimising the geographic data defect correlation, problem difficulty and sampling fraction, which are conceptually similar to their taxonomic counterparts. The only differences are that taxonomic variants are calculated across species rather than geographic units and pertain to $\ln(W_{jt})$, i.e. the log transformed relative abundance indices for some time-period after monitoring has begun, rather than abundance. Hence, the taxonomic problem difficulty is the variability of $\ln(W_{jt})$ across species, the data defect correlation is the correlation between whether a species was sampled (in time-periods 1 and t) and its value of $\ln(W_{jt})$, and the sampling fraction is the proportion of species that were sampled in both time-periods 1 and t.

Minimising the data defect correlation

In principle, reducing the taxonomic data defect correlation can be achieved in a similar manner to reducing its geographic counterpart. A set of variables could be sought that, once accounted for, reduce its conditional value relative to its unconditional value. Recall that the variables that satisfy this condition are generally the ones that induced the data defect correlation in the first place. Often, although not exclusively, these variables are common causes sample inclusion (here whether a species was sampled) and the variable of interest (here the relative abundance indices). Traits might be good candidates, since they could affect whether a species was sampled and its relative abundance index (e.g. a habitat specialist might be more likely to have been sampled because it is rare and more likely to be responding poorly to habitat loss). Once the data defect-inducing variables have been identified, sample weighting, superpopulation modelling and/or related approaches can then be used to correct for their effects.

If the variables that induced the taxonomic data defect correlation prove hard to identify or measure, a more practical option might be to exploit the fact that closely related species *could* be faring similarly (but see e.g. Losos, 2008). For example, Johnson et al. (2024) proposed a “correlated effects” model for relative abundance, which includes species level random effects whose covariance matrix encodes phylogenetic relatedness. If phylogeny explains an appreciable portion of the taxonomic data defect correlation, then the conditional data defect correlation given these random effects should be smaller than its unconditional value. This approach is closely related to (and can be combined with) the use of spatial random effects and autocorrelation terms, which might help to reduce the geographic data defect correlation in some circumstances.

Simpler forms of imputation than the ones described in the previous paragraph are generally used to deal with missing species in MSIs. One approach is to interpolate between years for which data are available on a per species basis (Collen et al., 2009). Others have proposed imputing values for missing species based on values for species that were sampled in the focal time-period (Freeman et al., 2021; Soldaat et al., 2017). Both of these approaches operate on the very strong assumption that non-sampled species are “Missing At Random” given the observed data (Rubin, 1976). We suggest that this assumption would be more plausible if estimators that condition on available data (e.g. superpopulation modelling or quasi-randomisation) were applied.

Increasing the (taxonomic) sampling fraction

Increasing the taxonomic sampling fraction can be achieved by obtaining data for underrepresented species or by modifying the definition of the population (Box 2). Obtaining data on underrepresented species means either collecting new data or mobilising previously inaccessible data. Modifying the population might mean restricting it to only those species sampled in every year, in which case the sampling fraction $f_{1,t} = 1$ and there is no taxonomic error relative to the population MSI.

Reducing the (taxonomic) problem difficulty

A reduction in the taxonomic problem difficulty, i.e. the standard deviation of the log relative abundance indices across species, could be achieved by restricting the population to a set of species that are thought to be faring similarly. In practice, this would probably mean focusing on species in a particular taxonomic or functional group on the assumption that they are responding similarly to environmental change. Species are included in existing MSIs, including the European farmland bird (Gregory et al., 2005) and grassland butterfly indicators (Van Swaay et al., 2008), based on their functional traits, so there is a precedent.

For some MSIs, conditioning the target population on a subset of species is not an option. One example is England’s ‘all species’ index (DEFRA, 2024), whose taxonomic scope is written into law. When the species set is fixed, the problem difficulty could be reduced by fitting a model for the growth rates. In this case, the (effective) problem difficulty becomes the unexplained rather than total variation in the growth rates across species. The more of the variation that the model explains, the greater the reduction in the problem difficulty.

Concluding remarks

Monitoring species’ populations using MSIs is generally a missing data problem in the sense that data on abundance are available for some species and sites in the target population but not others (Bowler et al., 2024; Dumelle et al., 2025). Consequently, it is not possible to verify a MSI empirically, and the potential for error must be appraised on theoretical grounds (and/or using in-silico experiments). Our theoretical framework is helpful in this respect, and, since it is merely an algebraic re-expression of the difference between the sample-based and population MSIs, it invokes very few assumptions. One notable exception is the assumption that abundance is measured without error (i.e. detection is perfect or at least consistently imperfect over space and time). This assumption is unlikely to hold in practice and should be relaxed in future work (e.g. Dempsey, 2023).

On a practical level, our framework can act as a guide to developers of MSIs. It reminds us that the first and most critical step is to clearly define the estimand, which should include a specification of the target parameter (e.g. mean growth rate) and the target population (the set of sites and species of interest). Once the estimand has been defined, the next step is to systematically assess the potential for error by considering the issues of data quantity, data quality, and problem difficulty, as reflected in the following questions:

- What fraction of sites in the target population were sampled, and has this changed over time?
- What fraction of species in the target population were sampled in all time-periods of interest?
- Are species similarly abundant at sampled and non-sampled sites, and has this changed over time?
- Are sampled species faring differently to the rest in terms of relative abundance?
- How variable is abundance across sites for any one species?
- How variable are the growth rates or relative abundance indices across species?

While most of these questions cannot be answered with certainty, carefully considering them is likely to reveal much about the potential for error and to guide more principled MSI development. Without such principles, the interpretation of biodiversity indicators and linked legislative targets is likely to be subject to so much model-based and epistemological uncertainty that scientific and political agreement on what they mean will remain out of reach.

Acknowledgements

Thank you to Kate Randall, whose modifications vastly improved Box figures 1 and 2. RJB and SGJ were supported by the UK CEH National Capability for UK Challenges programme NE/Y006208/1. RB was also supported by the MAMBO (Modern Approaches to Monitoring Biodiversity) project, which receives funding from the European Union’s Horizon Europe research and innovation programme under grant agreement No.101060639.

Appendix A

Derivation of equation 6

The relative error of the sample-based MSI is

$$\frac{(\bar{w}_t^J - \bar{W}_t^J)}{\bar{W}_t^J} = \frac{\bar{w}_t^J}{\bar{W}_t^J} - 1 = \frac{\exp\left[\frac{1}{n_{1,t}^J}\sum_{j\in s_{1,t}^J}\ln(w_{jt})\right]}{\exp\left[\frac{1}{N_{1,t}^J}\sum_{j=1}^{N_{1,t}^J}\ln(W_{jt})\right]} - 1. \tag{A1}$$

Focusing on $\bar{w}_t/\bar{W}_t$ (since -1 is a constant and provides no insight into the determinants of the error) and applying a log transformation yields

$$\ln\left(\frac{\bar{w}_t^J}{\bar{W}_t^J}\right) = \ln(\bar{w}_t^J) - \ln(\bar{W}_t^J) = \frac{1}{n_{1,t}^J}\sum_{j\in s_{1,t}^J}\ln(w_{jt}) - \frac{1}{N_{1,t}^J}\sum_{j=1}^{N_{1,t}^J}\ln(W_{jt}). \tag{A2}$$

Now let $\ln(w_{jt}) = \ln(W_{jt}) + \epsilon_{jt}$ be the estimated relative abundance index for species j in time-period t. It follows that the geographic estimation error for species j is $\epsilon_{jt} = \ln(w_{jt}) - \ln(W_{jt})$, which is an identity and imposes no assumptions about the distribution or behaviour of ϵ. Substituting into equation A2, we have

$$\ln\left(\overline{w}_t^J\right) - \ln\left(\overline{W}_t^J\right) = \frac{1}{n_{1,t}^J} \sum_{j \in s_{1,t}^J} \left(\ln\left(W_{jt}\right) + \epsilon_{jt}\right) - \frac{1}{N_{1,t}^J} \sum_{j=1}^{N_{1,t}^J} \ln\left(W_{jt}\right), \tag{A3}$$

which expands to

$$\ln\left(\overline{w}_t^J\right) - \ln\left(\overline{W}_t^J\right) = \frac{1}{n_{1,t}^J} \sum_{j \in s_{1,t}^J} \ln\left(W_{jt}\right) + \frac{1}{n_{1,t}^J} \sum_{j \in s_{1,t}^J} \epsilon_j - \frac{1}{N_{1,t}^J} \sum_{j=1}^{N_{1,t}^J} \ln\left(W_{jt}\right) \tag{A4}$$

or equivalently

$$\ln\left(\overline{w}_t^J\right) - \ln\left(\overline{W}_t^J\right) = \frac{1}{n_{1,t}^J} \sum_{j \in s_{1,t}^J} \ln\left(W_{jt}\right) - \frac{1}{N_{1,t}^J} \sum_{j=1}^{N_{1,t}^J} \ln\left(W_{jt}\right) + \frac{1}{n_{1,t}^J} \sum_{j \in s_{1,t}^J} \epsilon_{jt}. \tag{A5}$$

Appendix B

Derivation of equation 8

For any species j sampled in both time-periods 1 and t, the (log) geographic error component is

$$\ln\left(w_{jt}\right) - \ln\left(W_{jt}\right) = \ln\left(\frac{w_{jt}}{W_{jt}}\right) = \ln\left(\frac{\frac{\bar{y}_{jt}}{\bar{y}_{j1}}}{\frac{\bar{Y}_{jt}}{\bar{Y}_{j1}}}\right). \tag{B1}$$

Using the complex fraction and logarithm product rules, equation B1 can be rewritten as

$$\begin{aligned} \ln\left(\frac{\frac{\bar{y}_{jt}}{\bar{y}_{j1}}}{\frac{\bar{Y}_{jt}}{\bar{Y}_{j1}}}\right) &= \ln\left(\frac{\bar{y}_{jt}}{\bar{y}_{j1}} \times \frac{\bar{Y}_{j1}}{\bar{Y}_{jt}}\right) = \ln\left(\frac{\bar{y}_{jt}}{\bar{Y}_{jt}} \times \frac{\bar{Y}_{j1}}{\bar{y}_{j1}}\right) = \ln\left(\frac{\bar{y}_{jt}}{\bar{Y}_{jt}}\right) + \ln\left(\frac{\bar{Y}_{j1}}{\bar{y}_{j1}}\right) \\ &= \ln\left(\frac{\bar{y}_{jt}}{\bar{Y}_{jt}}\right) - \ln\left(\frac{\bar{y}_{j1}}{\bar{Y}_{j1}}\right). \end{aligned} \tag{B2}$$

We can rewrite the fractions on the right-hand sides of equations B2 as

$$\frac{\bar{y}_{jt}}{\bar{Y}_{jt}} = \frac{\bar{Y}_{jt} + \left(\bar{y}_{jt} - \bar{Y}_{jt}\right)}{\bar{Y}_{jt}} = 1 + \frac{\bar{y}_{jt} - \bar{Y}_{jt}}{\bar{Y}_{jt}} \tag{B3}$$

and

$$\frac{\bar{y}_{j1}}{\bar{Y}_{j1}} = \frac{\bar{Y}_{j1} + \left(\bar{y}_{j1} - \bar{Y}_{j1}\right)}{\bar{Y}_{j1}} = 1 + \frac{\bar{y}_{j1} - \bar{Y}_{j1}}{\bar{Y}_{j1}}. \tag{B4}$$

Substituting the right-hand sides of equations B3 and B4, we have

$$\ln\left(w_{jt}\right) - \ln\left(W_{jt}\right) = \ln\left(\frac{\bar{y}_{jt}}{\bar{Y}_{jt}}\right) - \ln\left(\frac{\bar{y}_{j1}}{\bar{Y}_{j1}}\right) = \ln\left(1 + \frac{\bar{y}_{jt} - \bar{Y}_{jt}}{\bar{Y}_{jt}}\right) - \ln\left(1 + \frac{\bar{y}_{j1} - \bar{Y}_{j1}}{\bar{Y}_{j1}}\right). \tag{B5}$$

Appendix C

Error decomposition of the sample LPI estimator

Here we show that the sample LPI estimator can be decomposed into within- and taxonomic components. Both components reflect the difference between sample and population means, so the Meng expression applies, and the decomposition is almost identical to the one presented in the main text.

Let $g_{jk} = \ln\left(y_{jkt}/y_{jk1}\right)$ be the log relative abundance index for species j at site k. Its mean across sampled sites (in both time periods 1 and t) is $\bar{g}_j$, and its mean across all sites in the population is $\bar{G}_j$. Now let the set of species present in both time periods 1 and t be ζ, the set species sampled in both periods be s_ζ, and the number of species in the population and sample be N_ζ and n_ζ, respectively. The basic LPI estimator is (Collen et al., 2009)

$$w = \exp\left(\frac{1}{n_\zeta}\sum_{j\in s_\zeta} \bar{g}_j\right), \tag{C1}$$

and the estimand is

$$W = \exp\left(\frac{1}{N_\zeta}\sum_{j=1}^{N_\zeta} \bar{G}_j\right). \tag{C2}$$

In practice, a weighted estimator is now used in place of equation C1 (McRae et al., 2017), the effects of which we explain in the main text.

The relative error of C1 as an estimator of C2 is $(w - W)/W = (w/W) - 1$. Focusing on w/W, since -1 is a constant and provides no insight into the determinants of the error, we have

$$\frac{w}{W} = \frac{\exp\left(\frac{1}{n_\zeta}\sum_{j\in s_\zeta} \bar{g}_j,\right)}{\exp\left(\frac{1}{N_\zeta}\sum_{j=1}^{N_\zeta} \bar{G}_j\right)} = \exp\left(\frac{1}{n_\zeta}\sum_{j\in s_\zeta} \bar{g}_j - \frac{1}{N_\zeta}\sum_{j=1}^{N_\zeta} \bar{G}_j\right). \tag{D3}$$

We can now use the identity $\bar{g}_j = \bar{G}_j + \varepsilon_j$, where $\varepsilon_j = \bar{g}_j - \bar{G}_j$ is the geographic error component, to write the relative error as

$$\frac{w}{W} = \exp\left(\underbrace{\frac{1}{n_\zeta}\sum_{j\in s_\zeta} \bar{G}_j - \frac{1}{N_\zeta}\sum_{j=1}^{N_\zeta} \bar{G}_j}_{\substack{cross-species\\ error}} + \underbrace{\frac{1}{n_\zeta}\sum_{j\in s_\zeta} \varepsilon_j}_{\substack{within-species\\ error}}\right). \tag{D4}$$

Both terms inside the exponential represent differences between sample and population means, so Meng's identity can be applied to each component, just as in the decomposition in the main text.

Appendix D

In the main text, we make the argument that changing one's estimator does not have a bearing on one's estimand. This statement is not strictly true in the estimating equation inferential framework, which is common in economics. In this setting, the estimand is defined implicitly as the solution to a population estimating equation: that is, as the population analogue of the sample-based estimating equation used to obtain the estimator.

Zhou & Meng (2026) have derived a general decomposition for estimators based on estimating equations. It is similar to the Meng expression in (equation 7) but includes a fourth quantity that measures the efficiency of the set of estimating equations. When we use a linear estimator for a linear estimand (e.g. using a sample average to estimate a population average), this quantity takes the value one, which is why it does not appear in equation (7). However, it is vital to recognize that changing a set of estimation equations can affect both estimand and estimator, even if initially one's desire is to alter the estimator only. We exercise the same caution here, since a seemingly impressive answer to a wrong question can be more harmful than being merely useless or wasting research resources.

References

Albert, C. H., Yoccoz, N. G., Edwards, T. C., Graham, C. H., Zimmermann, N. E., & Thuiller, W. (2010). *Sampling in ecology and evolution — bridging the gap between theory and practice*. *33*(6), 1028–1037. https://doi.org/10.1111/j.l600-0587.2010.0642

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

Backstrom, L. J., Callaghan, C. T., Worthington, H., Fuller, R. A., & Johnston, A. (2025). Estimating sampling biases in citizen science datasets. *Ibis*. https://doi.org/10.1111/ibi.13343

Bowler, D. E., Boyd, R. J., Callaghan, C. T., Robinson, R. A., Isaac, N. J. B., & Pocock, M. J. O. (2025). Treating gaps and biases in biodiversity data as a missing data problem. *Biological Reviews*, *100*(1), 50–67. https://doi.org/10.1111/brv.13127

Boyd, R. J., Aizen, M. A., Prado, L. F.-, Fontúrbel, F. E., Francoy, T. M., Martinez, L., Morales, C. L., Ollerton, J., Pescott, O. L., Powney, G. D., Mauro, A., Reto, S., Eduardo, S., & Carvell, C. (2022). Inferring trends in pollinator distributions across the Neotropics from publicly available data remains challenging despite mobilization efforts. *Diversity and Distributions*, *28*(May), 1404– 1415. https://doi.org/10.1111/ddi.13551

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

Boyd, R. J., Bowler, D. E., Isaac, N. J. B., & Pescott, O. L. (2024). On the trade-off between accuracy and spatial resolution when estimating species occupancy from geographically biased samples. *Ecological Modelling*, *493*. https://doi.org/10.1016/j.ecolmodel.2024.110739

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., Stewart, G. B., & Pescott, O. L. (2023). Descriptive inference using large, unrepresentative nonprobability samples: An introduction for ecologists. *Ecology*. https://doi.org/10.1002/ecy.4214

Buckland, S. T., Studeny, A. C., Magurran, A. E., Illian, J. B., & Newson, S. E. (2011). The geometric mean of relative abundance indices: a biodiversity measure with a difference. *Ecosphere*, *2*(9), art100. https://doi.org/10.1890/es11-00186.1

Collen, B., Loh, J., Whitmee, S., Mcrae, L., Amin, R., & Baillie, J. E. M. (2009). Monitoring Change in Vertebrate Abundance: The Living Planet Index. *Biology*, *23*(2), 317–327. https://doi.org/10.1111/j

Cooke, R., Mancini, F., Boyd, R., Evans, K. L., Shaw, A., Webb, T. J., & Isaac, N. J. B. (2023). Protected areas support more species than unprotected areas in Great Britain , but lose them equally rapidly. *Biological Conservation*, *278*(December 2022), 109884. https://doi.org/10.1016/j.biocon.2022.109884

DEFRA. (2024). *Indicators of species abundance in England*. https://www.gov.uk/government/statistics/indicators-of-species-abundance-in-england/indicators-of-species-abundance-in-england-frequently-asked-questions

Dempsey, W. (2023). Addressing selection bias and measurement error in covid-19 case count data using auxiliary information. *Annals of Applied Statistics*, *17*(4), 2903–2923. https://doi.org/10.1214/23-AOAS1744

Diggle, P. J., Menezes, R., & Su, T.-L. (2010). Geostatistical inference under preferential sampling. *Appl. Statist*, *59*(2), 191–232. http://www.blackwellpublishing.com/rss

Dorfman, A. H., & Valliant, R. (2005). Superpopulation Models in Survey Sampling. In P. Armitage & T. Colton (Eds.), *Encyclopedia of Biostatistics* (Issue July). https://doi.org/10.1002/0470011815.b2a16076

Dumelle, M., Trangucci, R., Nahlik, A. M., Olsen, A. R., Irvine, K. M., Blocksom, K., Ver Hoef, J. M., & Fuentes, C. (2025). Missing data in ecology: Syntheses, clarifications, and considerations. *Ecological Monographs*, *95*(4). https://doi.org/10.1002/ecm.70037

Ellwood, E. R., Dunckel, B. A., Flemons, P., Guralnick, R., Nelson, G., Newman, G., Newman, S., Paul, D., Riccardi, G., Rios, N., Seltmann, K. C., & Mast, A. R. (2015). Accelerating the digitization of biodiversity research specimens through online public participation. *BioScience*, *65*(4), 383–396. https://doi.org/10.1093/biosci/biv005

European Commission. (2024). *Nature Restoration Law*. https://environment.ec.europa.eu/topics/nature-and-biodiversity/nature-restoration-law_en

European Parliament. (2024). *Nature restoration: Parliament adopts law to restore 20% of EU's land and sea*. https://www.europarl.europa.eu/news/en/press-room/20240223IPR18078/nature-restoration-parliament-adopts-law-to-restore-20-of-eu-s-land-and-sea

Fick, S. E., & Hijmans, R. J. (2017). WorldClim 2 : new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, *37*(2), 4302–4315. https://doi.org/10.1002/joc.5086

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

Freeman, S. N., Isaac, N. J. B., Besbeas, P., Dennis, E. B., & Morgan, B. J. T. (2021). A Generic Method for Estimating and Smoothing Multispecies Biodiversity Indicators Using Intermittent Data. *Journal of Agricultural, Biological, and Environmental Statistics*, *26*(1), 71–89. https://doi.org/10.1007/s13253-020-00410-6

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

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

Grace, J. B., & Irvine, K. M. (2020). Scientist's guide to developing explanatory statistical models using causal analysis principles. *Ecology*, *101*(4), 1–14. https://doi.org/10.1002/ecy.2962

Gregory, R., & van Strien, A. (2010). Wild bird indicators: using composite population trends of birds as measures of environmental health. *Ornithological Science*, *9*(1), 3–22.

Gregory, R., 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

Guzman, L. M., Thompson, P. L., Viana, D. S., Vanschoenwinkel, B., Horváth, Z., Ptacnik, R., Jeliazkov, A., Gascón, S., Lemmens, P., Anton-Pardo, M., Langenheder, S., De Meester, L., & Chase, J. M. (2022). Accounting for temporal change in multiple biodiversity patterns improves the inference of metacommunity processes. *Ecology*, *103*(6). https://doi.org/10.1002/ecy.3683

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

Johnson, T. F., Beckerman, A. P., Childs, D. Z., Webb, T. J., Evans, K. L., Griffiths, C. A., Capdevila, P., Clements, C. F., Besson, M., Gregory, R. D., Thomas, G. H., Delmas, E., & Freckleton, R. P. (2024). Revealing uncertainty in the status of biodiversity change. *Nature*, *628*, 788–794. https://doi.org/10.1038/s41586-024-07236-z

Liu, K., & Meng, X. L. (2016). There Is Individualized Treatment. Why Not Individualized Inference? *Annual Review of Statistics and Its Application*, *3*, 79–111. https://doi.org/10.1146/annurev-statistics-010814-020310

Loh, J., Green, R. E., Ricketts, T., Lamoreux, J., Jenkins, M., Kapos, V., & Randers, J. (2005). The Living Planet Index: Using species population time series to track trends in biodiversity. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1454), 289–295. https://doi.org/10.1098/rstb.2004.1584

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

Losos, J. B. (2008). Phylogenetic niche conservatism, phylogenetic signal and the relationship between phylogenetic relatedness and ecological similarity among species. In *Ecology Letters* (Vol. 11, Issue 10, pp. 995–1003). https://doi.org/10.1111/j.1461-0248.2008.01229.x

Lundberg, I., Johnson, R., & Stewart, B. M. (2021). What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory. *American Sociological Review*, *86*(3), 532–565. https://doi.org/10.1177/00031224211004187

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

Mathur, M., Shpitser, I., & VanderWeele, T. (2023). A common-cause principle for eliminating selection bias in causal estimands through covariate adjustment. *OSF Preprints*. https://osf.io/ths4e/

McRae, L., Deinet, S., & Freeman, R. (2017). The diversity-weighted living planet index: Controlling for taxonomic bias in a global biodiversity indicator. *PLoS ONE*, *12*(1), 1–20. https://doi.org/10.1371/journal.pone.0169156

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.

Meyer, C., Weigelt, P., & Kreft, H. (2016). Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecology Letters*, *19*(8), 992–1006. https://doi.org/10.1111/ele.12624

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

Nishimura, R., Wagner, J., & Elliott, M. (2016). Alternative Indicators for the Risk of Non-response Bias: A Simulation Study. *International Statistical Review*, *84*(1), 43–62. https://doi.org/10.1111/insr.12100

Pearl, J., Glymour, M., & Jewell, N. (2016). *Causal inference in statistics: A primer*. Wiley.

Pescott, O. L., Boyd, R. J., Powney, G. D., & Stewart, G. B. (2023). Towards a unified approach to formal risk of bias assessments for causal and descriptive inference. *Arxiv*. https://doi.org/https://doi.org/10.48550/arXiv.2308.11458

Pescott, O. L., Powney, G. D., & Boyd, R. J. (2025). Adaptive sampling for ecological monitoring using biased data: a stratum-based approach. *Oikos*, *10*. https://doi.org/10.1002/oik.11115

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

Schouten, B. (2007). A Selection Strategy for Weighting Variables Under a Not-Missing-at-Random Assumption. *Journal of Official Statistics*, *23*(1), 51–68.

Schouten, B., Bethlehem, J., Beullens, K., Kleven, Ø., Loosveldt, G., Luiten, A., Rutar, K., Shlomo, N., & Skinner, C. (2012). Evaluating, Comparing, Monitoring, and Improving Representativeness of Survey Response Through R-Indicators and Partial R-Indicators. *International Statistical Review*, *80*(3), 382–399. https://doi.org/10.1111/j.1751-5823.2012.00189.x

Schouten, B., & Shlomo, N. (2017). Selecting Adaptive Survey Design Strata with Partial R-indicators. *International Statistical Review*, *85*(1), 143–163. https://doi.org/10.1111/insr.12159

Seaton, F. M., Jarvis, S. G., & Henrys, P. A. (2024). Spatio-temporal data integration for species distribution modelling in R-INLA. *Methods in Ecology and Evolution*, *15*(7), 1221–1232. https://doi.org/10.1111/2041-210X.14356

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

Soldaat, L. L., Pannekoek, J., Verweij, R. J. T., van Turnhout, C. A. M., & van Strien, A. J. (2017). A Monte Carlo method to account for sampling error in multi-species indicators. *Ecological Indicators*, *81*(May), 340–347. https://doi.org/10.1016/j.ecolind.2017.05.033

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

Van Swaay, C. 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

Wilkes, M. A., Mckenzie, M., Johnson, A., Hassall, C., Kelly, M., Willby, N., & Brown, L. E. (2025). Revealing hidden sources of uncertainty in biodiversity trend assessments. *Ecography*. https://doi.org/10.1111/ecog.07441

Zhou, J., & Meng, X.-L. (2026). Error - Free Error Attribution For Estimating Equations with Data of Arbitrary Quality. *In Preparation*.