A fundamental theory of actual error for species population monitoring

[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 Biological Sciences, University of Reading, UK

Corresponding author email: robboy@ceh.ac.uk

Abstract

Progress towards many national and international targets to halt and reverse declines of species populations (abundances) will be measured against Multispecies Biodiversity Indicators (MSIs). Like any sample-based estimator, MSIs approximate some real-world quantity (the estimand), and the difference between the two is the ‘actual’ or realised statistical error. We propose a general estimator and its corresponding estimand, both of which apply to many high-profile MSIs. Doing so allows us to decompose the error into a within-species component reflecting the impact of missing data for relevant locations and a cross-species component reflecting the impact of non-sampled species. Building on recent developments in sampling theory, we further decompose each of the within- and cross-species errors into three contributing factors: the ‘data defect’ (akin to sampling bias), the ‘data scarcity’ (reflecting the proportion of sites and species sampled) and the ‘problem difficulty’ (variability of abundance across sites and species). Approaches to reducing the error of MSIs can be recast as approaches to minimising one or more of these three quantities: for example, sample weighting reduces the data defect, sampling previously unmonitored species and locations minimises the data scarcity and focusing on functionally similar species may reduce the problem difficulty. Our theoretical framework thus unifies existing approaches to reducing the error of MSIs, reveals alternative approaches that might be considered in future and highlights opportunities for improving the communication of uncertainty.

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 Biodiversity 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 sites and species 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 differ from the population. It follows that the MSI obtained using the data in hand is likely to differ from the one that would have been obtained had all species and locations 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 difference between it and the estimator—the *estimation error*—is small.

In this paper, we develop a theoretical framework in which to consider the estimation error of MSIs. We begin by formalising the concept of the target population and specifying general mathematical expressions for the estimator and estimand. Doing so allows us to decompose the difference between the two, the estimation error, into within- and cross-species components. The within-species component reflects the fact that, for any given species, data may not be available for all sites in the population; the cross-species component reflects the fact that some species in the population might not have been sampled. Building on recent developments in sampling theory, and in particular Meng's (2018) re-expression of the difference between sample and population means, we further decompose the within- and cross-species error components into three fundamental quantities. Existing and prospective approaches to reducing the error of MSIs can be recast in terms of these quantities, and we review these in the final section.

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 'Species-Site Units', or SSUs). We will assume for simplicity that species and sites are classified in the same manner regardless of the time-period. Each SSU 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 SSUs, 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. The sample sets are then defined as

$s_t^J = \{j | \exists k \; such\ that\ R_{jkt} = 1\}$ (species that were sampled at least once at any site) and $s_{tj}^K = \{k | R_{jkt} = 1\}$ (sites at which species j was sampled or 'searched for').

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

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

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,

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

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

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

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 no imputed values of Y for now, 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*.

An alternative estimator based on cumulative per-period 'growth rates' is sometimes used (Collen et al., 2009; Freeman et al., 2021; McRae et al., 2017). If every species is sampled in every time-period, a point we come back to below, the two estimators are equivalent due to the 'telescoping' property of logarithms. Hence, we will focus on the estimator described by equations 1-3, which is simpler to work with.

The population analogue of the per period estimator is

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

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 $\bar{W}_t$ is the estimand.* What value $\bar{W}_t$

takes depends on the precise definition of the population, and we come back to this point below (also see Box 2).

Estimation error

Once the estimand has been defined, it is possible to consider whether the estimator is a good approximation to it. As defined here, MSIs reflect proportional change. Hence, it is natural to consider their relative (rather than absolute) 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

$$\frac{\bar{w}_t}{\bar{W}_t} = \frac{\exp\left(\frac{1}{n^J_{1,t}}\sum_{j\in s^J_{1,t}}\ln\left(w_{jt}\right)\right)}{\exp\left(\frac{1}{N^J_{1,t}}\sum_{j=1}^{N^J_{1,t}}\ln\left(W_{jt}\right)\right)}. \tag{5}$$

Error decomposition

Equation 5 is proportional to the relative error of $\bar{w}_t^J$ as an estimator of $\bar{W}_t^J$ but provides few direct insights into its determinants. By log transforming both sides, the error can be expressed more usefully in terms of cross- and within-species components (appendix A):

$$\ln\left(\frac{\bar{w}_t}{\bar{W}_t}\right) = \ln(\bar{w}_t) - \ln(\bar{W}_t) = \underbrace{\left(\frac{1}{n^J_{1,t}}\sum_{j\in s^J_{1,t}}\ln\left(W_{jt}\right) - \frac{1}{N^J_{1,t}}\sum_{j=1}^{N^J_{1,t}}\ln\left(W_{jt}\right)\right)}_{\substack{cross-species\\ component}} + \underbrace{\frac{1}{n^J_{1,t}}\sum_{j\in s^J_{1,t}}\epsilon_{jt}}_{\substack{within-species\\ component}}, \tag{6}$$

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 can vary arbitrarily among species. The cross-species error component is the difference between the sample and population means of $\ln\left(W_{jt}\right)$ across species and reflects the fact that for any given year some species may not have been sampled. The within-species component is the mean of ϵ_{jt} across sampled species. In the remainder of this section, we further decompose the cross- and within-species errors.

Cross-species error

To decompose the cross-species error component, we can exploit an algebraic identity derived by Meng (2018), which 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 see Fig. 1 and note that each of the quantities has a within-species analogue, which we also explain below). Applying Meng's decomposition to $\ln\left(W_{jt}\right)$, we have

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

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 $\ln(W_{jt})$, 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 $\sqrt{(1-f_{1,t})/f_{1,t}}$ is a measure of data scarcity. To obtain the expected difference between the sample and population means of $\ln(W_{jt})$, one simply substitutes the expected data defect correlation $\mathrm{E}[\rho(R_{1,t}, \ln(W_{jt}))]$ for its realised value $\rho(R_{1,t}, \ln(W_{jt}))$ (Lohr, 2022). $\rho(R_{1,t}, \ln(W_{jt}))$ partly reflects randomness in the way that the sample was collected, whereas $\mathrm{E}[\rho(R_{1,t}, \ln(W_{jt}))]$ is an underlying feature of the sampling design or lack thereof (reflecting the sampling bias).

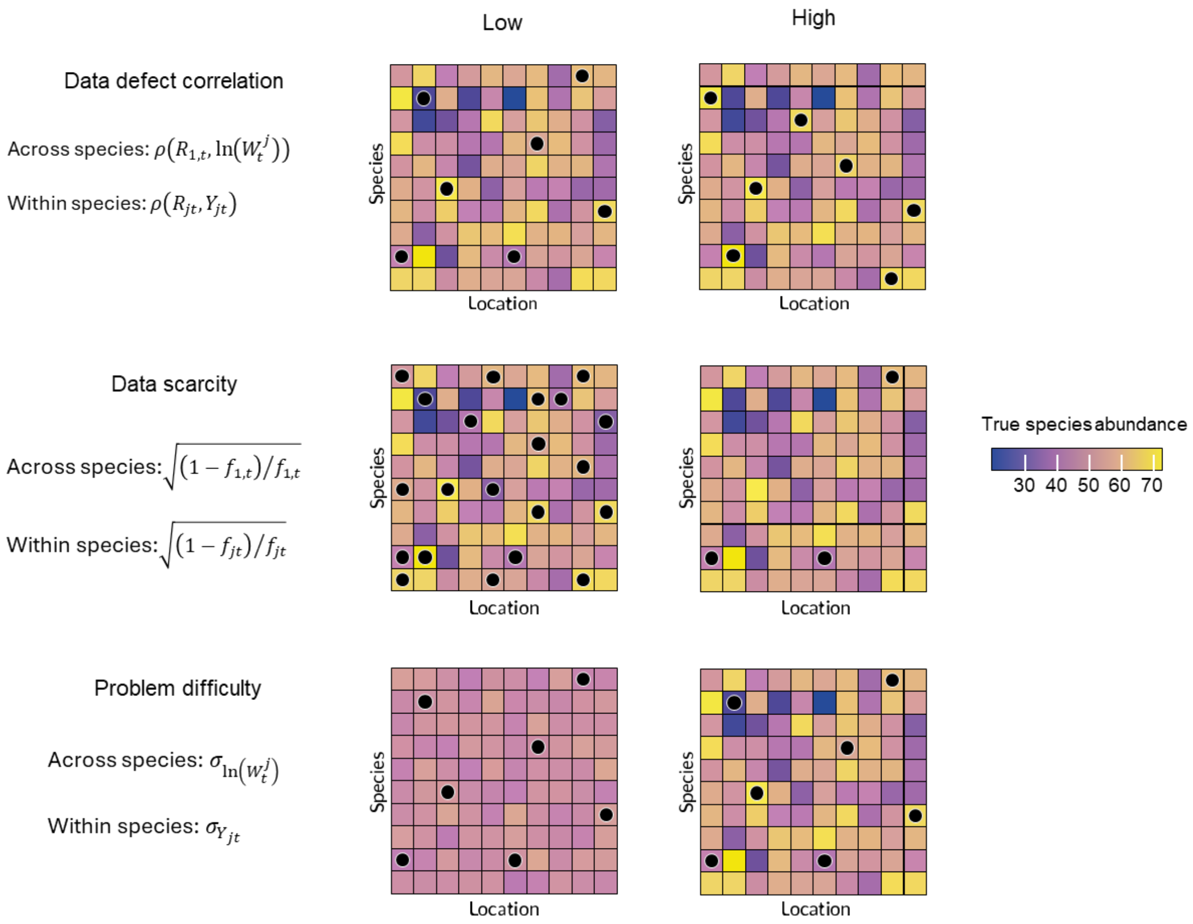

Low
High
Data defect correlation
Across species: $\rho(R_{1,t}, \ln(W_t^j))$
Within species: $\rho(R_{jt}, Y_{jt})$
Data scarcity
Across species: $\sqrt{(1-f_{1,t})/f_{1,t}}$
Within species: $\sqrt{(1-f_{jt})/f_{jt}}$
Problem difficulty
Across species: $\sigma_{\ln(W_t^j)}$
Within species: $\sigma_{Y_{jt}}$
Species
Location
True species abundance
30 40 50 60 70

Figure 1. Six grids depicting 100 species × location combinations, or SSUs. Each grid shows either a high or low value (left to right) of the data defect correlation, the data scarcity or the problem difficulty (top to bottom rows). Each of the three quantities operate both across and within species, and the panels depict situations in which the within- and cross-species variants are simultaneously low or high (e.g. the data defect correlation is low both across species and within species across locations, etc.). Note that in the top right panel, where the data defect is high, it is only SSUs with high abundance that have been sampled. Mathematical notation used elsewhere in the paper for each quantity is also provided.

Within-species error

Meng’s expression can also be applied to the within-species errors of the log relative abundance indices, but to see how we must write them in terms of differences between sample and population means. 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 within-species errors can be expressed as (appendix B)

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

That is, the log within species error for species j is the difference between the log relative errors in time-periods t and 1. The differences between the sample and population mean abundances in each period feature on the right-hand side, and we can substitute Meng's expression for each of them. Equation 8 is an exact identity for any realised sample, but it does not necessarily hold in expectation due to potential dependencies between the sample and population mean abundances. We further examine equation 8 and its implications for how to reduce the within-species errors in the next section.

Applying Meng's decomposition to the differences between the sample and population mean abundances for a given species in time-period t (which could equally be period 1), we have

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

Like 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 quantity. The quantities' meanings are subtly different to their cross-species 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 $\ln\left(W_{jt}\right)$ has been replaced by the abundance of species j in period t Y_{jt}. Hence, the within-species data defect correlation indicates whether the focal species is more abundant on average at sampled than non-sampled locations, and the within-species problem difficulty is the variability of the species' abundance across geographic units.

How to reduce estimation error

Equations 6 through 9 tell us how to reduce the cross-species error, the within-species errors and, consequently, the total estimation error of an MSI. (We consider the related problem of how to *assess* potential estimation error in Box 1.) It is easiest to see how the cross-species error can be reduced, because it is simply the difference between the sample and population means of $\ln\left(W_{jt}\right)$ across species, which is given by the Meng expression. The Meng expression shows that error as 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; reducing any of the quantities whilst the others are held constant will also reduce error.

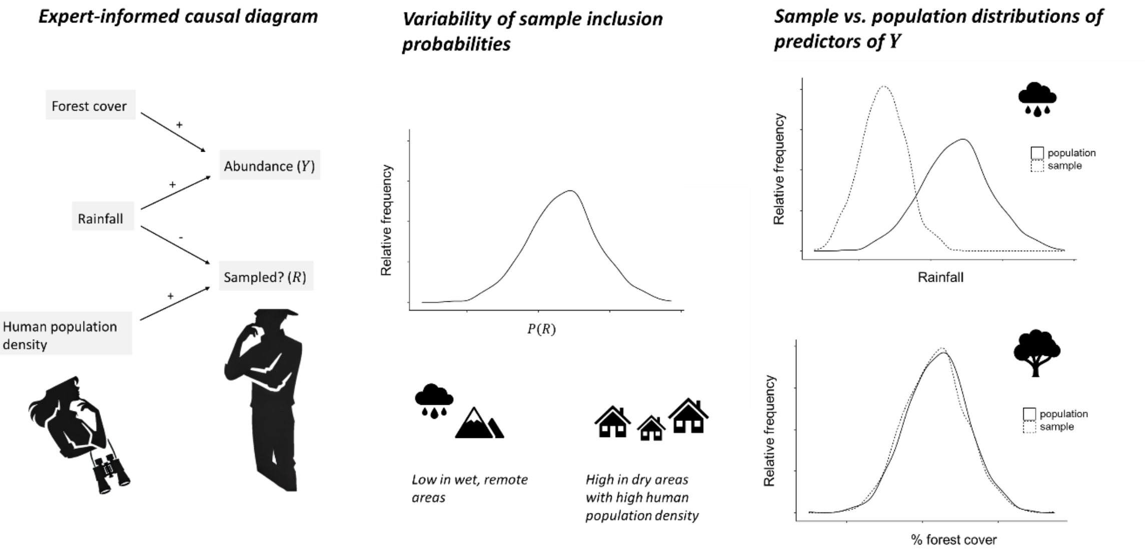
Reducing the within-species error for any given species (equation 8) is best achieved by reducing the per period estimation errors in time-periods 1 and t. It is true that one could get lucky and that the per period errors could have the same signs and similar magnitudes, in which case the within-species 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 (within-species) data defect correlation, data scarcity and problem difficulty will reduce the per period errors and thus the within-species error for a given species.

The total log relative estimation error is the sum of the cross- and within-species components (noting that the within-species component reflects a mean across sampled species). It is theoretically possible to have zero or negligible error if the two components cancel each other out (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 cross-species errors means reducing the cross- and within-species data defect correlations, problem difficulties and data scarcities (the latter being equivalent to maximising the sampling fraction). Starting with the within-species variants, we explain how each of these might be achieved below.

Box 1. How to assess potential estimation error.

To understand the potential error of an MSI, we require information on the within- and cross-species 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 locations in the population that have been sampled, and they are measurable. The data defect correlations and problem difficulties are not directly measurable and must be estimated or qualitatively assessed.

We are aware of three general approaches to assessing the potential for a non-zero data defect correlation. One 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 here 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-zero data defect correlation (Boyd, Botham, et al., 2024; Thoemmes & Mohan, 2015). The second approach is to estimate sample inclusion probabilities $P(R)$ and to calculate their variability in the population (e.g. Schouten et al., 2012). If the variability of $P(R)$ is small, then R and Y can only covary so much, and the data defect correlation is likely to be small (Nishimura et al., 2016)(Nishimura et al., 2016) The third approach 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 indicates a non-zero data defect correlation. Box Fig. 1 summarises our three approaches to estimating data defect correlations in the context of species population monitoring.



Box figure 1. Schematic illustrating how one might diagnose a non-zero within-species 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-zero (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 estimating 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).

Approaches to estimating the problem difficulty (the standard deviation of Y) can also be imagined. One simple option is to use the sample standard deviation of Y as an estimate. Generally, the sample standard deviation is smaller than its population equivalent, so it could serve as a lower bound. A better alternative 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.

Within-species estimation error

Minimising the data defect correlation

The key to reducing the within-species 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 abundanceY_{jt} in the first place. Often, although not always, 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, $\rho(R_{jt}, Y_{jt}) > 0$ (other variables might induce a negative correlation). For a given level of protected area status (e.g. inside or outside), however, the value of $\rho(R_{jt}, Y_{jt})$ should be smaller than its value across all locations, 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-zero 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 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 SSUs (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 (or a better fit for non-sampled sites, which would imply a very poor model!). 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 cross-species data defect correlation in some circumstances, as we explain below).

Increasing the sampling fraction (reducing the data scarcity)

One way to reduce the data scarcity—or, equivalently, to increase the within-species 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., 2024; Schouten & Shlomo, 2017).

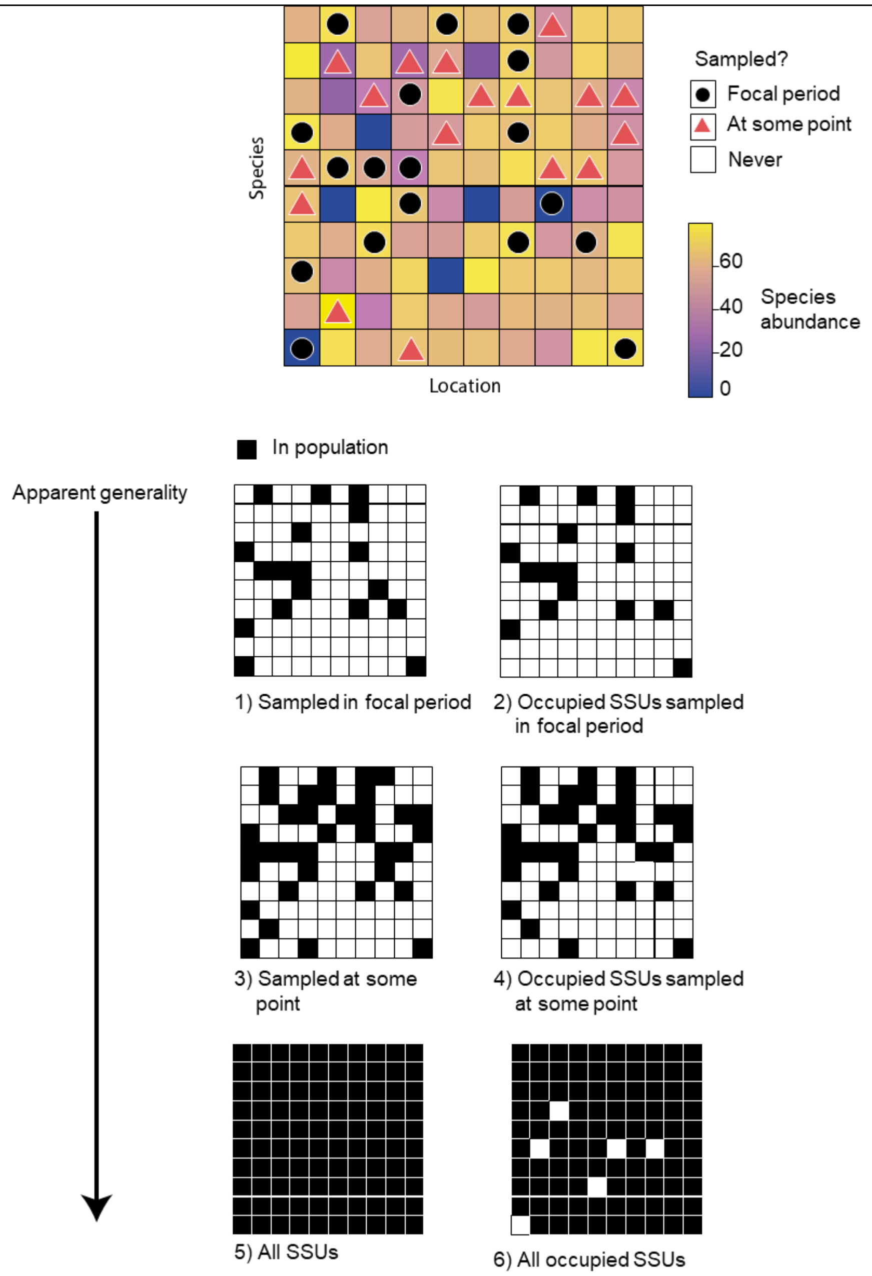
A second and much simpler way to increase f_{jt} is to recognise that the population need not include every site and to constrain it from the outset. Conditioning on (i.e. restricting the population to) the set of sampled geographic units for a given species, for example, means that $f_{jt} = 1$, the data quantity term $\sqrt{(1-f_{jt})/f_{jt}} = 0$ and, consequently, that the within-species 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 the analyst must consider this alongside the desire to minimise error. Conditioning on occupied or sampled sites reduces the number of SSUs 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. 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 Species-Site Unit (SSU). Rather, we might consider a conditional target population given, say, occupancy O_t (i.e. whether $Y_t > 0$) 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 SSUs with zero abundance. 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.

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.



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 SSUs, 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 (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 within-species 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.

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 that Y_{jt} follows a zero-inflated Poisson distribution across sites and let q (which we do not index for simplicity of notation) be the proportion of 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 (appendix C). If occupied sites are omitted, 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).

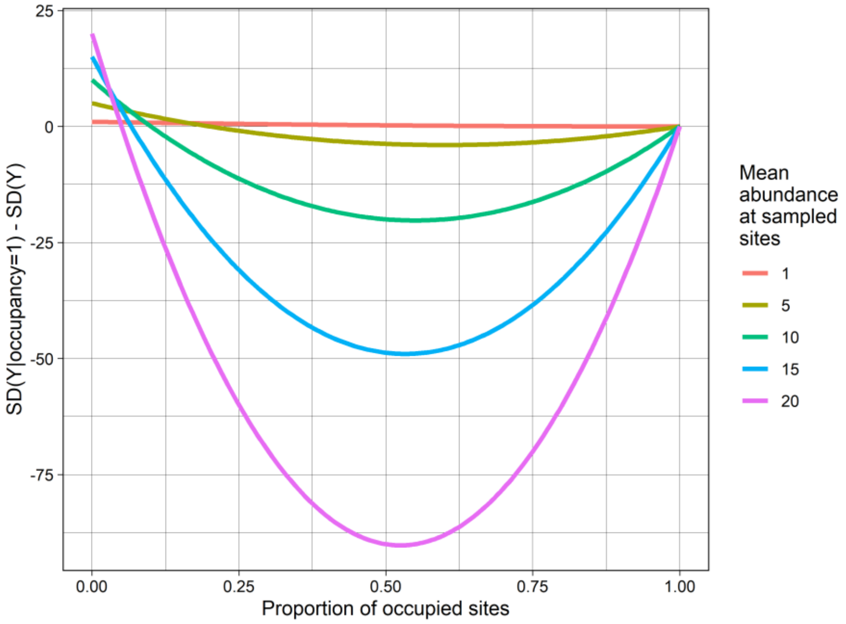


Figure 2. Difference in the problem difficulty (population standard deviation of abundance) when the population is defined as occupied sites only and when it includes all sites. Negative values indicate that omitting unoccupied sites from the population reduces the problem difficulty. Each curve represents one value of mean abundance across occupied sites.

Another way to modify the population, which could also reduce the within-species 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.

Cross-species estimation error

Many of the principles described above apply to minimising the within-species data defect correlation, problem difficulty and sampling fraction, which are conceptually similar to their cross-species counterparts. The only differences are that cross-species 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 cross-species 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 cross-species data defect correlation can be achieved in a similar manner to reducing its within-species 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 cross-species 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 (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 cross-species 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 within-species data defect correlation in some circumstances.

Simpler forms of imputation than the ones described above 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), an assumption we suggest would be more plausible if, say, superpopulation models or weighted estimators were applied.

Increasing the (cross-species) sampling fraction

Increasing the cross-species 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 cross-species error relative to the population MSI.

Reducing the (cross-species) problem difficulty

A reduction in the cross-species 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. Of course, restricting the population in this way will not be appropriate if it means omitting species that are relevant to a species abundance target or if a general MSI reflecting a large fraction of described species is desired.

Estimation error and power to detect change

The actual relative error of an MSI is one way to conceptualise our lack of knowledge about how species are faring; another is in terms of statistical power to detect real change (Leung & Gonzalez, 2024; Valdez et al., 2023). Real change (i.e. a non-zero population MSI) is detectable if the ratio of the sample-based estimate to its standard error exceeds some critical threshold (e.g. 1.96 for the 95% confidence level). Consequently, for a given standard error, if the actual error reduces the magnitude of the estimate, then real change becomes less detectable and vice versa.

Interestingly, the source of the actual error affects its impact on whether a trend can be detected. Although we have not framed it this way so far, actual error may reflect either a systematic bias or sampling variability. A systematic bias occurs when the expected data defect correlations are appreciably non-zero, and sampling variability reflects fluctuations in the data defect correlations across the many possible (and usually hypothetical) samples that could have been obtained. Large sampling variability should be reflected in the standard error of the estimate. Hence, if the actual error primarily reflects variance, then the ratio of the estimate to its standard error can only be so large, and real change can only be so detectable. If the actual error primarily reflects a systematic bias, however, the standard error may be small. In this case, whether real change can be detected depends primarily on whether the true trend and the actual error have the same sign—a bias of the same sign as the trend will make the trend more detectable and vice versa. This insight also highlights a well-known conflict between binary conceptions of "detecting" change (i.e. P -value cut-offs philosophically related to decision-theoretic models of inference; Greenland, 2023) and solely descriptive presentations: if a large contribution of systematic bias to actual error is suspected, then, even if there is evidence that the bias is the same sign as the trend, descriptive MSIs must be wrong. Should the trend be plotted under these circumstances without visual warnings (Pescott et al., 2022)?

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). Consequently, it is not possible to verify a MSI empirically, and the potential for error must be appraised on theoretical grounds. 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. there are no false absences or presences or that the prevalence of these remains constant over time and space). 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 demonstrates 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 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.

Appendix A

Derivation of equation 6

The relative error of the sample-based MSI is

$$\frac{(\bar{w}_t - \bar{W}_t)}{\bar{W}_t} = \frac{\bar{w}_t}{\bar{W}_t} - 1 = \frac{\exp\left[\frac{1}{n^J_{1,t}}\sum_{j\in s^J_{1,t}} \ln(w_{jt})\right]}{\exp\left[\frac{1}{N^J_{1,t}}\sum_{j=1}^{N^J_{1,t}} \ln(W_{jt})\right]} - 1. \tag{A7}$$

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

$$\ln\left(\frac{\bar{w}_t}{\bar{W}_t}\right) = \ln(\bar{w}_t) - \ln(\bar{W}_t) = \frac{1}{n^J_{1,t}}\sum_{j\in s^J_{1,t}} \ln(w_{jt}) - \frac{1}{N^J_{1,t}}\sum_{j=1}^{N^J_{1,t}} \ln(W_{jt}). \tag{A8}$$

Now let ln (w_{jt}) = ln $(W_{jt}) + \epsilon_j$ be the estimated relative abundance index for species j. It follows that the within-species estimation error for species j is ϵ_j = ln (w_{jt}) – ln (W_{jt}), which is an identity and imposes no assumptions about the distribution or behaviour of ϵ. Substituting into equation A8, we have

$$\ln(\bar{w}_t) - \ln(\bar{W}_t) = \frac{1}{n^J_{1,t}}\sum_{j\in s^J_{1,t}} \left(\ln(W_{jt}) + \epsilon_j\right) - \frac{1}{N^J_{1,t}}\sum_{j=1}^{N^J_{1,t}} \ln(W_{jt}), \tag{A9}$$

which expands to

$$\ln(\bar{w}_t) - \ln(\bar{W}_t) = \frac{1}{n^J_{1,t}}\sum_{j\in s^J_{1,t}} \ln(W_{jt}) + \frac{1}{n^J_{1,t}}\sum_{j\in s^J_{1,t}} \epsilon_j - \frac{1}{N^J_{1,t}}\sum_{j=1}^{N^J_{1,t}} \ln(W_{jt}) \tag{A10}$$

or equivalently

$$\ln(\bar{w}_t) - \ln(\bar{W}_t) = \frac{1}{n^J_{1,t}}\sum_{j\in s^J_{1,t}} \ln(W_{jt}) - \frac{1}{N^J_{1,t}}\sum_{j=1}^{N^J_{1,t}} \ln(W_{jt}) + \frac{1}{n^J_{1,t}}\sum_{j\in s^J_{1,t}} \epsilon_j. \tag{A11}$$

Note that while equation 11 is an exact identity for realised relative error given the sample in hand, it does not necessarily hold in expectation due to potential dependencies between terms.

Appendix B

Derivation of equation 8

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

$$\ln(w_{jt}) - \ln(W_{jt}) = \ln\left(\frac{w_{jt}}{W_{jt}}\right) = \ln\left(\frac{\frac{\bar{y}_{jt}}{\bar{y}_{j1}}}{\frac{\bar{Y}_{jt}}{\bar{Y}_{j1}}}\right). \tag{A12}$$

Using the complex fraction and logarithm product rules, equation A12 can be rewritten as

$$\ln\left(\frac{\frac{\bar{y}_{jt}}{\bar{y}_{j1}}}{\frac{\bar{Y}_{jt}}{\bar{Y}_{j1}}}\right) = \ln\left(\frac{\bar{y}_{jt}}{\bar{y}_{j1}} \times \frac{\bar{Y}_{j1}}{\bar{Y}_{jt}}\right) = \ln\left(\frac{\bar{y}_{jt}}{\bar{y}_{j1}}\right) + \ln\left(\frac{\bar{Y}_{j1}}{\bar{Y}_{jt}}\right). \tag{A13}$$

We can then apply the logarithm quotient rule to expand each term on the right-hand side:

$$\ln\left(\frac{\bar{y}_{jt}}{\bar{y}_{j1}}\right) + \ln\left(\frac{\bar{Y}_{j1}}{\bar{Y}_{jt}}\right) = \left(\ln(\bar{y}_{jt}) - \ln(\bar{y}_{j1})\right) + \left(\ln(\bar{Y}_{j1}) - \ln(\bar{Y}_{jt})\right). \tag{A14}$$

Rearranging the terms on the right-hand side yields

$$\ln\left(\frac{\bar{y}_{jt}}{\bar{y}_{j1}}\right) + \ln\left(\frac{\bar{Y}_{j1}}{\bar{Y}_{jt}}\right) = \left(\ln(\bar{y}_{jt}) - \ln(\bar{Y}_{jt})\right) - \left(\ln(\bar{y}_{j1}) - \ln(\bar{Y}_{j1})\right). \tag{A15}$$

It is also evident from the logarithm quotient rule that

$$\ln(\bar{y}_{jt}) - \ln(\bar{Y}_{jt}) = \ln\left(\frac{\bar{y}_{jt}}{\bar{Y}_{jt}}\right) \tag{A16}$$

and that

$$\ln(\bar{y}_{j1}) - \ln(\bar{Y}_{j1}) = \ln\left(\frac{\bar{y}_{j1}}{\bar{Y}_{j1}}\right). \tag{A17}$$

We can rewrite the fractions on the right-hand sides of equations A16 and A17 as

$$\frac{\bar{y}_{jt}}{\bar{Y}_{jt}} = \frac{\bar{Y}_{jt} + (\bar{y}_{jt} - \bar{Y}_{jt})}{\bar{Y}_{jt}} = 1 + \frac{\bar{y}_{jt} - \bar{Y}_{jt}}{\bar{Y}_{jt}} \tag{A18}$$

and

$$\frac{\bar{y}_{j1}}{\bar{Y}_{j1}} = \frac{\bar{Y}_{j1} + (\bar{y}_{j1} - \bar{Y}_{j1})}{\bar{Y}_{j1}} = 1 + \frac{\bar{y}_{j1} - \bar{Y}_{j1}}{\bar{Y}_{j1}}. \tag{A19}$$

Substituting the right-hand sides of equations A18 and A19, we have

$$\ln(w_{jt}) - \ln(W_{jt}) = \ln\left(\frac{\bar{y}_{jt}}{\bar{Y}_{jt}}\right) - \ln\left(\frac{\bar{y}_{j1}}{\bar{Y}_{j1}}\right) = \ln\left(1 + \frac{\bar{y}_{jt} - \bar{Y}_{jt}}{\bar{Y}_{jt}}\right) - \ln\left(1 + \frac{\bar{y}_{j1} - \bar{Y}_{j1}}{\bar{Y}_{j1}}\right). \tag{A16}$$

Like equation A11, equation A16 is an exact identity given the sample in hand but does not necessarily hold in expectation.

Appendix C

Variance of the ZIP model

The ZIP (zero-inflated Poisson) model assumes that abundance Y is generated from two processes. The first process determines occupancy O and follows a Bernoulli distribution:

$$O \sim Bernoulli(q), \tag{A17}$$

where $q = 1 - p$ is the proportion of occupied sites and p is the proportion of unoccupied sites. The second process follows a Poisson distribution:

$$X \sim Poisson(\mu), \tag{A18}$$

where μ is the mean of X across occupied sites. Assuming O and X are independent, abundance is given by $Y = OX$. That is, if $O = 1$, then $Y = X$, and if $O = 0$, then $Y = 0$. The independence of O and X also implies that

$$E[Y] = E[OX] = E[O]E[X] = q\mu. \tag{A19}$$

From the law of total variance,

$$V[Y] = V[E(Y|O)] + E[V(Y|O)], \tag{A21}$$

where

$$E[V(Y|O)] = P(O = 1)V(Y|O = 1) + P(O = 0)V(Y|O = 0). \tag{A22}$$

Since $E(Y|O) = O\mu$, the first term on the right-hand side of equation A21 is $V[O\mu]$. Now, recognising that $V[aX] = a^2V[X]$ (for constant a),

$$V[O\mu] = \mu^2V[O]. \tag{A23}$$

As O is Bernoulli distributed,

$$V[O\mu] = \mu^2q(1 - q). \tag{A24}$$

The second term on the right-hand side of equation A21 is $E[V(Y|O)]$. If $O = 1$, then, since X is Poisson distributed, $V[Y|O = 1] = V[X] = \mu$. If $O = 0$, $V[Y|O = 0] = 0$. Hence,

$$E[V(Y|O)] = E[O\mu]. \tag{A25}$$

Due to the linearity of expectations,

$$E[O\mu] = \mu E[O]. \tag{A26}$$

And since $E[O] = q$,

$$E[V(Y|O)] = E[O\mu] = \mu q. \tag{A27}$$

Summing the terms give the total variance:

$$V[Y] = \mu^2q(1 - q) + \mu q. \tag{A28}$$

The expression in A28 tells us that the variance of the ZIP has two components: $\mu^2 q(1-q)$, which represents the variance of occupancy O, and μq, which represents the variance of Y at occupied sites. Equation A28 can be derived more simply using standard results for the variance of a product of random variables: $Y = OX$, $V[Y] = V[OX] = E[O^2]V[X] + V[O]E[X]^2 = q\mu + q(1-q)\mu^2$. Nevertheless, we include the more complete derivation for pedagogical purposes.

References

Backstrom, L. J., Callaghan, C. T., Worthington, H., Fuller, R. A., & Johnston, A. (2024). 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. (2024). Treating gaps and biases in biodiversity data as a missing data problem. *Biological Reviews*. 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*. https://doi.org/10.1111/2041-210X.14492

Boyd, R. J., Botham, M., Dennis, E., Fox, R., Harrower, C., Middlebrook, I., Roy, D., & Pescott, O. (2024). Using causal diagrams and superpopulation models to correct geographic biases in biodiversity monitoring data. *EcoEvoRxiv*.

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. (2023a). Descriptive inference using large, unrepresentative nonprobability samples: An introduction for ecologists. *Ecology*. https://doi.org/10.1002/ecy.4214

Boyd, R. J., Stewart, G. B., & Pescott, O. L. (2023b). Descriptive Inference using large, unrepresentative nonprobability samples: An introduction for ecologists. *Ecology*, *forthcomin*.

Bradley, V. C., Kuriwaki, S., Isakov, M., Sejdinovic, D., Meng, X. L., & Flaxman, S. (2021). Unrepresentative big surveys significantly overestimated US vaccine uptake. *Nature*, *600*(7890), 695–700. https://doi.org/10.1038/s41586-021-04198-4

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. In *Appl. Statist* (Issue 2). http://www.blackwellpublishing.com/rss

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

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*. 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

Greenland, S. (2023). 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., & van Strien, A. (2010). Wild bird indicators: using composite population trends of birds as measures of environmental health. *Ornithological Science*.

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

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*. https://doi.org/10.1038/s41586-024-07236-z

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

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. (2024). *Adaptive sampling for ecological monitoring using biased data: A stratum-based approach*.

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

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., 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*. 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

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

Van Swaay, C. A. M., 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