Treating gaps and biases in biodiversity data as a missing data problem

Diana E. Bowler[1]*, Robin J. Boyd[1], Corey T. Callaghan[2], Robert A. Robinson[3], Nick J. B. Isaac[1], Michael J. O. Pocock[1]

Affiliations

1 UK Centre for Ecology &amp; Hydrology, Wallingford, OX10 8BB, UK

2 Department of Wildlife Ecology and Conservation, Fort Lauderdale Research and Education Center, University of Florida, Davie, FL 33314-7719

3 British Trust for Ornithology, The Nunnery, Thetford, IP24 2PU, UK

* Corresponding author: diana.e.bowler@gmail.com

Abstract

Big biodiversity datasets have great potential for monitoring and research because of their large taxonomic, geographic and temporal scope. Such datasets have become especially important for assessing the temporal change of species' populations and distributions. Gaps in the available data, however, often hinder drawing large-scale inferences about species' trends. Here, we conceptualise biodiversity data gaps as a missing data problem, which provides a unifying framework for the challenges and potential solutions across different types of biodiversity datasets. We characterise the typical types of data gaps in biodiversity data as different classes of missing data and then use missing data theory to explore the implications for different research questions. By using this framework, we show that bias due to data gaps can arise when the factors affecting sampling and/or data availability overlap with those affecting biodiversity. But the outcome also depends on the ecological questions, which determines choices around the analytical approach. We argue that typical approaches to long-term species trend modelling are especially susceptible to data gaps since such models do not tend to account for the factors that drive missingness. To identify general solutions, we review empirical studies and use simulation studies to compare some of the most frequently employed approaches to deal with data gaps, including subsampling, weighting and imputation. All these methods have the potential to reduce bias but may come at the cost of increased uncertainty of parameter estimates. Weighting approaches are arguably the least used so far in ecology and have the potential to reduce both the bias and variance of parameter estimates. Regardless of the method, the ability to reduce bias critically depends on knowledge of, and the availability of data on, the factors creating data gaps. We use our review to outline the necessary considerations when dealing with data gaps at different stages of the data collection and analysis workflow.

Keywords: Biodiversity change; Citizen Science; Ecological Modelling; Macroecology; Spatial bias

Contents

I. Introduction: uneven sampling of biodiversity

Ecologists have ever-growing access to data on species' occurrence and abundances. Potential sources of data include long-term citizen-science monitoring schemes (such as the North American Breeding Bird Survey) (Bled *et al.*, 2013), data aggregators (such as the Global Biodiversity Information Facility) (Garcia-Rosello *et al.*, 2015), remote-sensing platforms (Fretwell, Scofield & Phillips, 2017) and synthesis databases (such as BioTIME or the Living Planet Database) (Dornelas *et al.*, 2014). Since these data cover broad spatial and temporal scales, they are especially useful for large-scale questions, for instance, about species' distributions, population and community-level trends, and ecological niches (Chandler *et al.*, 2017; Sullivan *et al.*, 2017; Fink *et al.*, 2020). These data also underpin many biodiversity trend indicators that are central for national and international conservation policy (Gregory *et al.*, 2005; van Swaay *et al.*, 2008; Fraisl *et al.*, 2020).

Despite the impressive volume of data, biodiversity data, regardless of the source, tend to be filled with gaps and redundancies (Boakes *et al.*, 2010). Data gaps are not necessarily problematic; indeed, most studies rely on statistical inference to make inferences about a broader region of interest from a sample. Data gaps, however, can be problematic when they lead to biases (Boakes *et al.*, 2010; Bled *et al.*, 2013; Amano, Lamming & Sutherland, 2016). Already many ecologists have raised concerns about the impacts of bias on estimated spatial or temporal biodiversity patterns (Bayraktarov *et al.*, 2019; Valdez *et al.*, 2023). Developing methods to deal with data gaps and associated biases within large-scale biodiversity data is an increasingly important task to make full use of the growing big data sources.

Patterns in the availability of biodiversity data can be affected by the original motivations for, and constraints on, data collection activities. While some data are collected as part of scientific studies, much of the available data on species' occurrence and abundance are collected through citizen science initiatives (Chandler *et al.*, 2017). Spatial patterns in data availability from citizen science have been especially well-studied. Citizen science programs have varying degrees of protocol and sampling designs (Isaac & Pocock, 2015; Pocock *et al.*, 2017) but more data are typically collected in accessible areas such as near roads and urban areas (Geldmann *et al.*, 2016). Such biases are not unique to citizen science data, as even data collected during formal scientific studies have potential sampling biases towards regions undergoing less habitat change, which may lead to underestimates of biodiversity change (Gonzalez *et al.*, 2016; Forister *et al.*, 2023; Cardinale *et al.*, 2018). Various solutions have been proposed to deal with these biases (Hefley *et al.*, 2013; Cretois *et al.*, 2021; Johnston *et al.*, 2020; Ver Hoef *et al.*, 2021), but there is still a lack of a general framework for ecologists to guide decisions on when and how to deal with data gaps.

Here, we show how using missing data theory (Rubin, 1976) can unify problems associated with data gaps across different types of biodiversity datasets. Missing data are a widespread problem crossing disciplines, with a large body of literature on the implications and possible solutions (Little & Rubin, 2019; Carpenter & Kenward, 2012). We expect that aligning the generalized problem of missing data, conceptualized within missing data theory, to the problem of biodiversity data biases discussed above will yield opportunities so far overlooked. We mostly focus our review on modelling trends in species occupancy or abundance using monitoring data collected by volunteer citizen scientists, but the ideas transfer to other types of biodiversity data or questions. We show that bias is not a property of a dataset but rather a property of the use of a dataset for a specific question and target population that are imposed by the data analyst. We review some commonly used solutions to missing data to highlight potential approaches that could be considered in biodiversity analyses.

II. Classifying data gaps using missing data theory

(1) Biodiversity data gaps

Species occurrence or abundance data can have gaps in different dimensions. We distinguish between spatial, annual and within-year gaps (Fig. 1). We define spatial gaps as those formed by sites with no data, and annual gaps as those formed by a lack of data in some years at sites that have been otherwise sampled. Together, spatial and annual gaps determine the spatial and temporal coverage of a dataset. Within-year gaps arise when data are lacking in specific seasons or months, which can be important because most organisms are seasonal and multiple visits are usually necessary to robustly estimate detection probabilities. Considering why these gaps arise can help understand their likely impact, for instance, on long-term species trend estimation. Data gaps are found in different types of monitoring data including highly structured monitoring schemes with a standardised protocol, such as many national bird survey schemes, as well as opportunistic monitoring data that are typically an aggregation of heterogeneous observations. Biodiversity datasets can also have taxonomic gaps (Troudet *et al.*, 2017), but this is outside the scope of this paper.

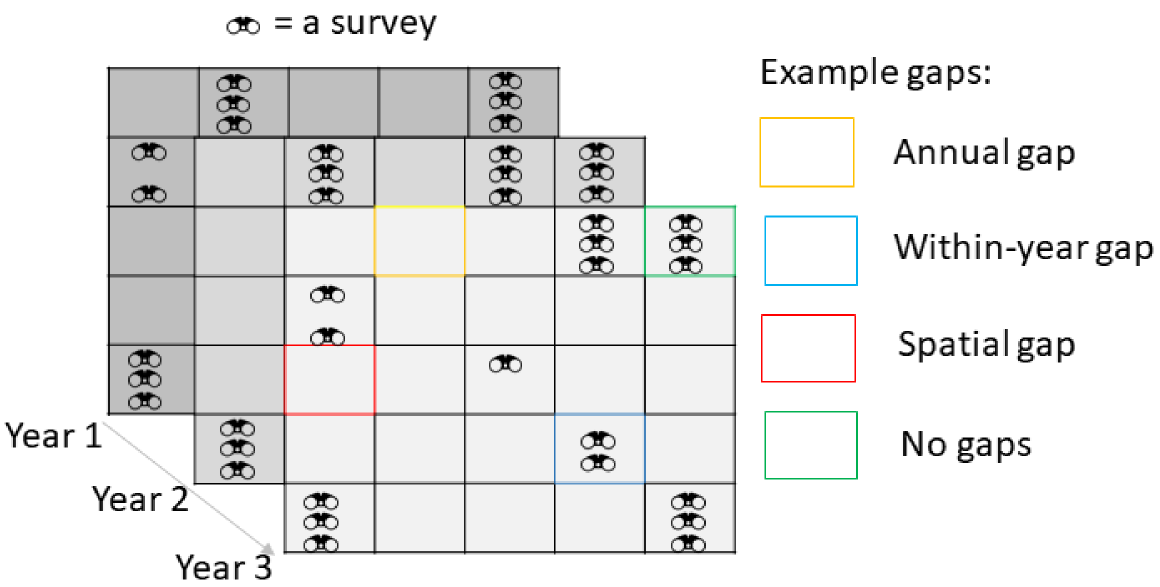

Fig. 1 Different types of data gaps within biodiversity data.

We imagine a scenario where there are multiple survey visits across sites and years. Visits can be in response to a protocol (‘structured’ data) or opportunistic (‘unstructured’), and repeat visits can be by the same or multiple recorders. Data gaps, or more generally uneven data availability, can arise due to (a) within-year gaps (e.g., blue square, i.e., ordinarily there are three visits, but some sites are only visited once or twice in a year); (b) annual gaps (e.g., yellow square, i.e., some sites that are usually sampled are entirely unvisited in some years) or (c) spatial gaps (e.g., red square, i.e., some sites within the region of interest are never visited across all years). Some sites are well-sampled within and across years and hence have no missing data (e.g., green square).

While both structured and opportunistic monitoring data can be affected by similar data gaps (Binley & Bennett, 2023), there are some key differences between these types of monitoring data. Moreover, structured schemes themselves vary in the degree of structure and standardisation. In structured schemes with a formal spatial sampling design, data gaps include both planned and unplanned gaps. Planned gaps arise because only a sample of sites was ever intended to be sampled. Unplanned gaps occur because of failure to recruit and retain surveyors at sites that were intended to be sampled (Zhang *et al.*, 2021; Marsh & Cosentino, 2019). In most other types of data, gaps are neither planned or unplanned. Some monitoring schemes have sampling protocols but participants are free to choose their own sampling sites. In fully opportunistic monitoring schemes, participants make individual decisions about where to sample and gaps emerge from unevenness in the cumulative sampling effort of all participants. Due to the high number of participants, and lack of coordination of their effects, sampling effort is generally more strongly skewed across space and time in opportunistic schemes than in structured schemes, leading to more pervasive data gaps (Geldmann *et al.*, 2016). Synthesis databases such as BioTIME and the Living Planet Database, and data aggregators such as GBIF, are similar in these respects to schemes without a formal spatial sampling design since they contain data that were independently collected as part of separate studies, without coordinated efforts.

Despite these differences, correlates of data gaps tend to be similar across monitoring schemes, especially those involving citizen scientists. Spatial gaps often occur in remote areas because there is a smaller pool of potential participants nearby (Geldmann *et al.*, 2016; Mandeville, Nilsen & Finstad, 2022). Spatial gaps can also be more common where species have lower abundance or land cover is perceived to be less attractive for biodiversity and for surveying e.g., agricultural land (Tulloch *et al.*, 2013; Dambly *et al.*, 2021; Marsh & Cosentino, 2019). Annual gaps can arise due to project turnover or because of external factors (e.g. the 2020 season for most countries was highly compromised by the Covid-19 pandemic). Annual gaps have also been linked with local land use changes that negatively affected species abundance and the attractiveness of a site for sampling (Zhang *et al.*, 2021; Marsh & Cosentino, 2019). Within-year data gaps can be caused by periods of inclement weather (Zimney & Smart, 2022; Diekert *et al.*, 2023) or vary seasonally e.g., missing surveys for butterflies are more common at start and end of the main flight period (Dennis *et al.*, 2016), while bird sampling can be higher during their migration periods (La Sorte & Somveille, 2020).

(2) Classes of missing data

Within the classic missing data theory, there are three classes of missing data (Missing Completely at Random, Missing at Random, Missing Not at Random), defined below, each with different consequences for bias (Table 1) (Rubin, 1976; Nakagawa & Freckleton, 2008; Little & Rubin, 2019). These classes vary in their missing data mechanism, which describes the relationship between the probability of missing data (or sampling effort in the monitoring context) and the values of other variables. Hefley et al. (2013) already proposed viewing spatial biases in presence-only data as a form of missing data. Here, we extend it more broadly across different types of biodiversity data.

Within the context of biodiversity data, missingness can be regarded as Missing Completely at Random (MCAR) if the factors affecting biodiversity sampling, and causing missingness, are independent of those affecting biodiversity (Table 1). Under MCAR, the observed data are effectively a random sample of the whole population, and the values of the variable of interest are similar in sampled and non-sampled sites or times. For instance, if site selection is driven by human accessibility, but species distribution is primarily driven by climate, and if accessibility and climate are not correlated, then spatial data gaps would be MCAR. Within-year gaps associated with weekdays (Evans & Day, 2002; Courter *et al.*, 2013), or annual gaps associated with project turnover, are also examples likely to cause MCAR data gaps since such gaps are probably not associated with biodiversity patterns (Table 1). In this case, missing data could reduce the precision of parameters estimates through reduced sample size, but not increase the bias.

When the factors affecting sampling are the same as, or correlated with, those affecting biodiversity, the missing data mechanism can either be Missing at Random (MAR) or Missing Not at Random (MNAR). For instance, if road density affects both sampling probability and species abundance, then spatial gaps are not MCAR. Road density might affect sampling probability directly (e.g., if people are more often looking for wildlife along roads) or indirectly (e.g., if road density affects species detectability); in either case, road density influences data gaps. Similarly, habitat degradation could reduce both species abundance and observer retention to continue sampling at a site, creating an annual data gap that is MAR or MNAR (Table 1). In these cases, there are systematic differences in the biodiversity quantity of interest between sampled and non-sampled sites or times (Table 1).

To borrow from an infamous quote, if we regard data gaps as “unknowns”, then MAR can be thought of as “known unknowns” while MNAR are “unknown unknowns”. The “known” needed for MAR is knowledge and availability of data on the shared covariates affecting sampling and biodiversity. If complete data for shared covariates are available and included in the analysis, then the missing data mechanism is MAR. Hence, despite its name, MAR does not mean that sampling effort is randomly distributed in the landscape. Rather, it means that the covariates affecting sampling are known and that there is available covariate data to fully explain the differences between sampled and non-sampled potential data. If any of the relevant factors affecting sampling and biodiversity are unknown, or not modelled, the missing data mechanism becomes MNAR (Table 1). Hence, decisions of the analyst can determine whether a data gap is MNAR or MAR (discussed more fully in section III). MNAR may also arise when missingness is dependent on the value of biodiversity itself, i.e., if sampling effort directly depends on species occurrence or abundance.

Statistical tests can only partly indicate which missing data class is most likely (Little, 1988). Analysis of relationships between data availability and observed covariates can point towards MAR if some relationships are significant. But a lack of any association, or an incomplete explanation of data gaps, could reflect MCAR or MNAR. Because MNAR is associated with unavailable data, it cannot be directly tested. Concerns about whether missingness in the biodiversity data is directly associated with its values could be explored if there is a related variable that is fully available (Wu, 2022). We argue that MCAR is unlikely in most biodiversity data since unplanned data gaps can affect even the most structured monitoring schemes.

Table 1 Missing data mechanisms in biodiversity data, including examples and implications

Mechanism	Typical meaning	Meaning in the context of biodiversity data	Examples	Typical implications
Missing completely at random (MCAR)	Missingness is independent of observed and unobserved variables.	Sampling is independent of any covariates, or covariates that affect sampling probability are independent of those affecting biodiversity	Within-year: Weekday gaps Annual/Spatial: Gaps caused by the completion of a fixed-term project or retirement of a participant	Ignorable
Missing at random (MAR)	Missingness is associated with observed data but not any unobserved variables	Covariates that affect sampling probability are shared with those affecting biodiversity, but data are available on all these covariates	within-year: Season (day of year) Annual: Urban development Spatial: Accessibility	Ignorable if the model includes all relevant covariates
Missing not at random (MNAR)	Missingness depends on unobserved variables or the missing values itself	CS sampling varies with biodiversity value or an unknown or unavailable covariate affects sampling and biodiversity	within-year/annual/spatial: unknown factors causing variation in species activity/abundance that are also correlated with sampling effort	Non-ignorable - the missing data mechanism needs to be modelled

III. Implications of missingness for ecological questions

Missing data (i.e., data gaps) themselves do not necessarily have strong impacts on the results of biodiversity modelling, but can depend on the specific question and parameter of interest (Bartlett, Harel & Carpenter, 2015; Collins, Schafer & Kam, 2001; Little *et al.*, 2022). Viewing data gaps as a form of missing data can help decide whether a particular data gap matters. As we note above, data gaps that are MCAR do not cause bias, but data gaps in biodiversity data are unlikely to be wholly MCAR. For a data gap to be MAR rather than MNAR can depend on the ecological question being pursued by an analyst. This is because the ‘missing at random’ assumption of MAR is conditional on controlling for covariates affecting sampling probability, which means that these covariates are known, reflected in available data and included in the analysis (Fig. 2) (Conn, Thorson & Johnson, 2017; Hefley *et al.*, 2013). Different ecological questions will lead to different decisions about which variables to include in an analysis. Hence, data gaps of the same dataset might be MAR under some questions but MNAR under others. To illustrate these potential differences, we contrast two typical questions asked with biodiversity data.

(1) Understanding the roles of environmental drivers on species’ distributions

Monitoring data are often used to understand the environmental factors explaining species distribution patterns. The implications of missing data for species distribution models have been considered in terms of niche truncation. Niche truncation happens when a dataset only contains occurrence data from part of the geographic range of a species, which usually also means that the dataset only covers part of the ecological/environmental space that is suitable for the species (Chevalier *et al.*, 2022; Albert *et al.*, 2010; Guo *et al.*, 2023). These studies show that the implications of niche truncation depend on the functional form of the relationship between the associated covariate and the species response (Chevalier *et al.*, 2022) and whether occurrence data are presence/absence or presence-only (Baker *et al.*, 2022).

We begin considering the scenario when presence-absence data are available. In this case, if there is a simple linear relationship, missing data do not necessarily cause bias in the estimated effect of the covariate on biodiversity, even when missingness depends on the same covariate (Fig. 1a, c, e, g) (Collins *et al.*, 2001). For instance, we could estimate the effect of elevation on species occurrence, if it is linearly related, even if elevation is also associated with data gaps. This is because the relationship between the covariate and species occurrence can be estimated without bias using data over a restricted range of covariate values. This is shown in e.g., Fig. 1c - the same relationship is found with a full dataset (green) or a restricted dataset with data gaps (purple). Missing data can, however, cause problems when the underlying relationship between the covariate and species occurrence is non-linear. In this case, data gaps can hinder estimating the true form of the relationship (see Fig. 2i - a curved relationship in fit with the full dataset but a simple positive linear relationship with the restricted dataset). The relationship that is fit using the restricted dataset will critically depend on which portion of the covariate range is sampled. Since many ecological associations show some non-linearity, or context-dependencies such that relationships depend on the value of other variables (Spake *et al.*, 2023), we expect this issue is likely to be widespread in species distribution models.

We now consider the alternative scenario of fitting a distribution model with presence-only data. In this case, any data gaps could represent a lack of sampling or a lack of true species occurrence. This creates an inherent identifiability challenge for any model seeking to separate the processes affecting sampling from the true ecological processes affecting species distributions with presence-only data (Hefley *et al.*, 2013; Baker *et al.*, 2022). Many methods have been developed to generate pseudo-absences (Barbet-Massin *et al.*, 2012; Hertzog, Besnard & Jay-Robert, 2014), but such models are still usually more prone to biases when there are shared covariates affecting sampling and species occurrence (Baker *et al.*, 2022). More recent approaches to modelling presence-only data, by integrating them with any available presence-absence data (Fithian *et al.*, 2015), may help minimise some of these biases.

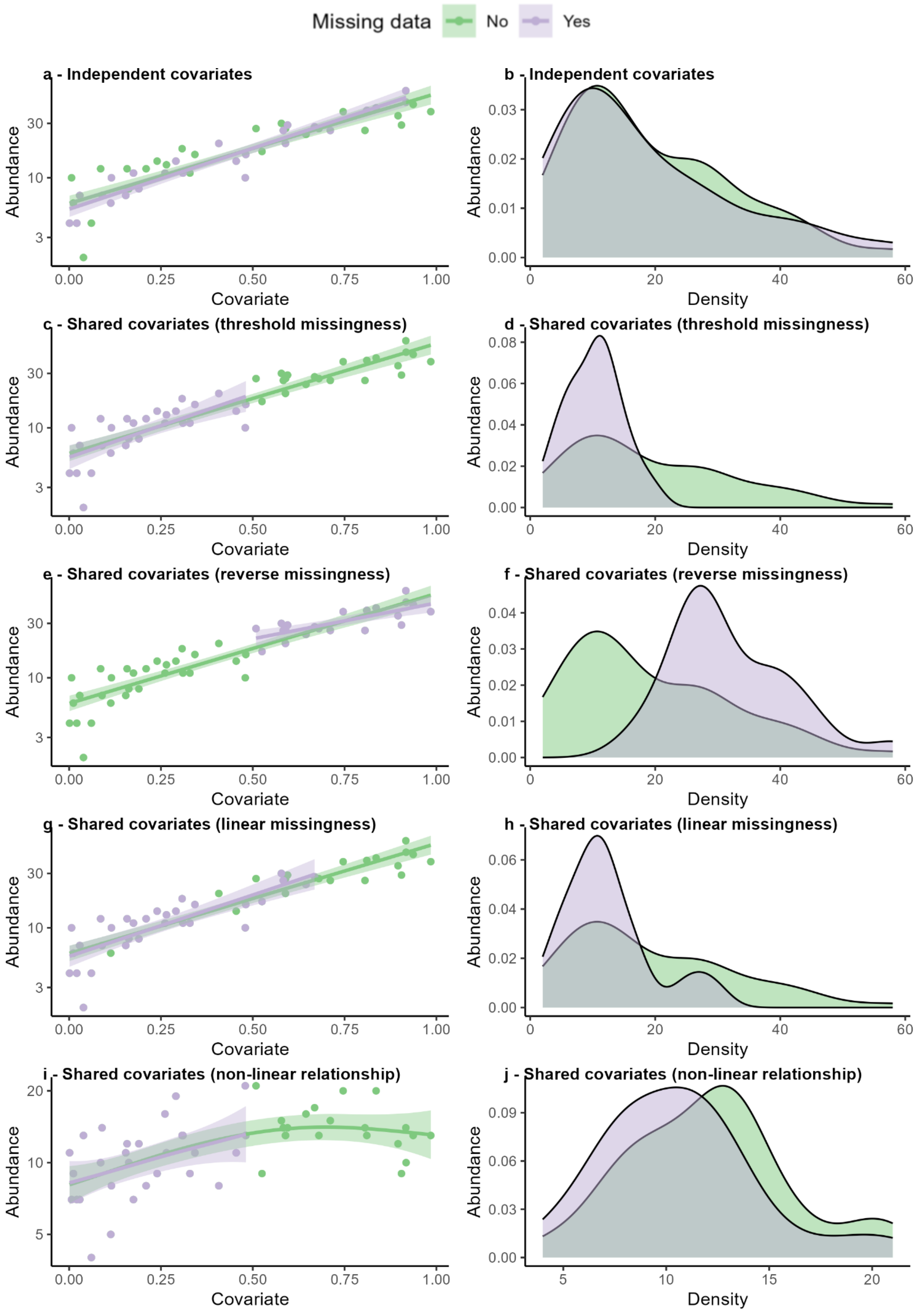

Fig. 2 The impacts of different missing data mechanisms on regression (left) and sample distributions (right).

We use a hypothetical dataset to highlight different missing data mechanisms. In (a) and (b), the covariate affecting sampling probability is *independent* from the covariate affecting species abundance. In this case, both the estimated effect of the covariate (e.g., in a linear regression, shown in a by the solid line) and the sample distribution (b) are similar in a dataset with (purple) and without (green) missing data. (i.e., missingness is MCAR). In (c) and (d), the covariate affecting sampling probability is the *same as or correlated with* the covariate affecting species abundance - in this case, data are missing when the covariate is above average (i.e., threshold missingness). The estimated effect of the covariate is the same in the dataset with and without missing values (shown in c) but the sampling distribution is different (d). In (e) and (f), the missingness pattern is reversed compared to (c) and (d) (i.e., data are missing when the covariate is below average), but we can similarly retrieve the same unbiased covariate effect (e) even though there is greater mean abundance in the dataset with missing values (f). In (g) and (h), the covariate affecting sampling probability is the *same as or correlated with* the covariate affecting species abundance - in this case, the probability of missing data increases with the value of the covariate (i.e., linear missingness). Again, the estimated effect of the covariate is the same (shown in g) but the sampling distribution is different (h). In (i) and (j), the covariate affecting sampling probability is the *same as or correlated with* the covariate affecting species abundance; additionally, the true relationship between the covariate and species abundance is non-linear and data are missing when the covariate is above average. The mechanism is now MNAR since the model cannot be correctly specified with the observed data.

(2) Estimating trends in species abundances

Models to estimate species' trends tend to be descriptive: spatial variation is modelled by including site identity (as a fixed or random term) while any temporal trend is modelled as a simple year effect (either as a linear, spline or a categorical term) (Amano *et al.*, 2012; Bled *et* *al.*, 2013). Drivers of the trend are not explicitly modelled when the goal is to simply estimate the mean trend over time. Broader inferences about the trend estimated by such models are based on the assumed representativeness of the sample, or prior knowledge of the inclusion probabilities of sampling units (see design weights discussed in section IV 2). Basing inference from the sampling design is the most traditional approach to surveys (Smith, 1976) and the approach typically taken by official governmental surveys (van den Brakel & Bethlehem, 2008). This approach has the advantage of avoiding complex assumptions in the statistical analysis (Buckland *et al.*, 2012) and is perhaps also easier to analyse and communicate to stakeholders and laypersons.

Simple trend models may, however, lead to biased trend estimates for biodiversity when data gaps are not MCAR. We illustrate this in a simple simulation in which site-level species trends were assumed to depend on a site-level covariate e.g., urban cover (Fig. 3). We assumed sites were sampled either with a probability affected by an independent covariate (Fig. 3 middle panel) or with a probability affected by the same site-level covariate affecting species trends (Fig. 3 right panel), a scenario already identified as a pitfall in some monitoring schemes (Buckland & Johnston, 2017). We estimated trends using a simple mixed effect model including site and year. This shows that when the site-level covariate affected both sampling effort and species' trends, the trends were biased, but site-level trends were unbiased when an independent covariate affected sampling. In real world situations, many factors will influence the trend of a species, but this toy simulation highlights the potential for bias caused by shared covariates. Since the specific covariates affecting sampling effort and biodiversity trends are not considered in the typical forms of analysis for trend modelling, trend analyses are liable to be affected by MNAR, whereas by including appropriate covariates (where possible), the data gaps become

MAR instead and trends will be unbiased. Without conditioning on the covariates involved, trend estimates might be overestimated if missing data are more common in static regions where species trends are more stable; but underestimated if missing data are more common in dynamic regions where species trends more strongly deviate from zero (Fig. 3) (Bowler *et al.*, 2022; Buckland & Johnston, 2017).

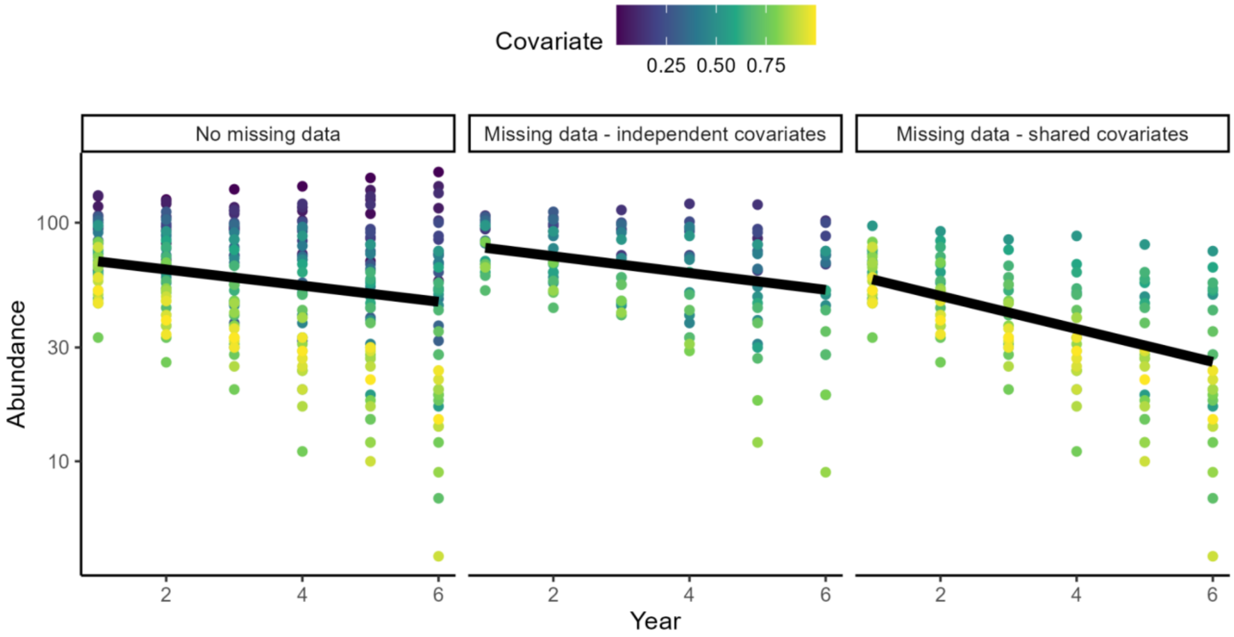

Fig. 3 The impacts of different missing data mechanisms on trend modelling

We use a hypothetical scenario in which a mean trend model is fit to datasets that vary in their missing data mechanism. We assumed a scenario of 50 sites that varied in an environmental covariate affecting species trends (trends were stable or even increasing at low values of the covariate and declining at increasingly high values of the covariate). When missing data was independent (i.e., a MCAR pattern - the covariate affecting sampling probability was a different and uncorrelated covariate), the overall mean trend (estimated by the year effect in a generalized linear mixed effect model that also included a site random effect) was similar with (middle panel) and without (left panel) missing data. By contrast, when the same covariate affected both species' trends and sampling probability, leading to less sampling in sites with low values of the covariate (notice there are fewer blue points in the right panel - a MNAR pattern), the overall mean trend was downward biased with missing data (right panel) compared to the scenario of no missing data (left panel).

IV. Missing data solutions

A broad range of methods to deal with missing data have been used in ecology (Hossie, Gobin & Murray, 2021; Nakagawa & Freckleton, 2008; Lopucki *et al.*, 2022). Many solutions are particularly relevant when data are missing in both response and predictor variables. Here, we focus on the typical scenario in biodiversity modelling of missing data only in the response variable (i.e., in the biodiversity data) since typical predictors tend to have no or few gaps (e.g., site identity or environmental data from remote sensing). We organise solutions into three groups - subsampling, weighting and imputation (Fig. 4), which have been tested to varying degrees already with both structured and unstructured biodiversity data (Table 2). Most solutions to deal with missing data are only appropriate for MCAR or MAR missingness. MNAR is the most challenging missing data mechanism to deal with in statistical modelling so, we deal with MNAR in a later section.

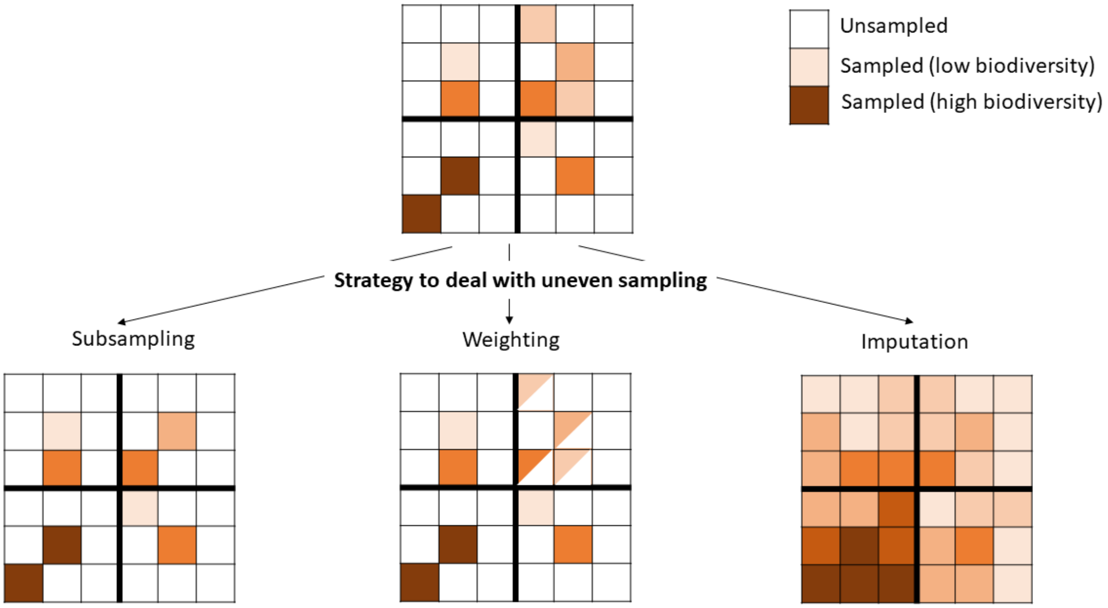
Unsampled
Sampled (low biodiversity)
Sampled (high biodiversity)
Strategy to deal with uneven sampling
Subsampling
Weighting
Imputation

Fig. 4 Visualisation of contrasting approaches to deal with data gaps.

We focus on spatial gaps to illustrate the possible approaches, but the ideas apply to other types of data gaps (Fig. 1). (top) the landscape is divided into four quarters (e.g., representing different habitats or geographic regions). One quarter (top right quarter) has been sampled more (4 sampling sites) than the others (2 sampling sites). Solutions: Random subsampling (bottom left): two sites are randomly subsampled from the oversampled quarter to create a dataset with an even sampling coverage across quarters. Weighting (bottom middle): data from the oversampled quarter is downweighted in the statistical model so data from all quarters similarly influence the modelled results. Imputation (bottom right): missing values at unsampled sites are imputed based on the spatial pattern in the data and/or environmental covariates, and summary parameters are calculated based on both predictions at sampled and unsampled sites. In subsampling and weights, the aim is to improve the representativeness of the sample for statistical inference at the population-level. In imputation, the aim is to directly predict population-level values.

Table 2 Example applications of the solutions to deal with data gaps within biodiversity data.

Type of data gaps	Typical approaches:
Within-year	Sometimes imputed e.g., spline terms to smooth over seasonal variation in sampling times during the flight period of butterflies (Dennis *et al.*, 2016)
Annual	Sometimes imputed e.g., general linear models to impute annual gaps based on mean site and year effects, optionally allowing for habitat effects e.g., used in TRIM abundance indices, (Lehikoinen *et al.*, 2016)
Spatial	Often ignored, but occasionally weighting by geographic regions (Bled *et al.*, 2013) or imputed (Breivik *et al.*, 2021), or reduced by subsampling (Johnston *et al.*, 2021).

(1) Subsampling

The 'Big Data Paradox' highlights that there can be trade-offs between dataset size and dataset quality (Bradley *et al.*, 2021; Meng, 2018). Small datasets can be preferable to large datasets, if they are more representative and less heterogeneous than a larger dataset (Bayraktarov *et al.*, 2019). Based on such thinking, some studies have proposed to 'reverse engineer' structure in biodiversity data by filtering data (Rapacciuolo, Young & Johnson, 2021). Part of this reverse engineering has attempted to deal with spatial biases; for instance, by spatially subsampling data to reduce the unevenness of sampling effort across the landscape (Steen *et al.*, 2021; Matutini *et al.*, 2021; Steen, Elphick & Tingley, 2019; Boria *et al.*, 2014; Robinson *et al.*, 2020). This has been tested on, for instance, the semi-structured data compiled by eBird (Johnston *et al.*, 2020). Some have also applied this approach to reduce temporal skews in sampling effort (Hof & Bright, 2016; Zbinden *et al.*, 2014), although not always successfully (Callcutt, Croft & Smith, 2018). Subsampling can also be used to balance the amount of data across a single or multi-dimensional environmental gradient; essentially stratified sampling of the original sample

(Meng, 2022; Nunez-Penichet *et al.*, 2022). Recent class balancing approaches have been developed to ensure that important observations, especially for rare species, are not lost during the subsampling process (Robinson *et al.*, 2020; Steen *et al.*, 2021; Gaul *et al.*, 2022).

(2) Weighting

Weighting is a common practice in survey analysis, especially in the social sciences (Li *et al.*, 2013; Seaman & White, 2013; Raghunathan, 2004). Weighting can serve different purposes, including reducing the impact of confounding variables when the goal is to estimate the causal effect of an intervention. But weighting can also be used to deal with missing data that is not MCAR. For instance, weighting can be used to reduce selection bias caused by participant nonresponse in surveys (Seaman & White, 2013), but it is less often used to account for data gaps in biodiversity data (Boyd, Powney & Pescott, 2023a; Aubry & Francesiaz, 2022).

Different types of weights have been used in the analysis of biodiversity data: (1) design weights; (2) estimated non-response weights (or sampling weights) and (3) population weights. Each form of weighing is intended to improve sample representativeness of some target population but vary in terms of whether the weights derive from the sampling design and the dimension of representativeness under consideration. Design weights are based on the study sampling design and assumed to be known with certainty, and hence are only relevant for structured schemes with a sampling design. For instance, in many national bird breeding schemes, the design weights are based on the geographic strata that underlie a random stratified study design (Buckland *et al.*, 2012). Non-response weights are used to account for unplanned missing data in structured schemes (Frair *et al.*, 2004) or variation in sampling effort in unstructured schemes (Johnston *et al.*, 2020; Hefley *et al.*, 2013), which means that are not known with certainty and must be estimated. Population weights are used to ensure the sample is representative of the full distribution/population of a species and are typically assumed to be known. Population weights are used in the calculation of supranational/international indicators (e.g., farmland or woodland bird indicators (Gregory *et al.*, 2005)) in which national estimates are combined by giving greater weight to regions that harbour a larger proportion of the species total population.

Non-response weights are usually the most difficult to include since they are not known *a priori* and need to be estimated. Predictive models (e.g., random forest models) have been used to predict the probability that a site is sampled based on the set of covariates (e.g., land cover or climate, or accessibility) available across all sampled and unsampled sites, with the inverse of these probabilities used as weights (Little *et al.*, 2022; Johnston *et al.*, 2020). Alternatively, poststratification (for categorical covariates), or more generalized calibration approaches (allowing both continuous and categorical covariates), can be used, which adjust the weight given to each data point until the joint or marginal distributions of covariate values in the observed sample matches those for the population (Boyd, Stewart & Pescott, 2023b). In both cases, weighting can cause problems when there are regions within the target population with close to zero probability of being sampled, which could lead to some data points having extremely large weights. In this case, weights may need to be redefined e.g., by coarsening the covariates used to define the weights, or by truncating weight values so that extreme weights are not produced (Battaglia, Hoaglin & Frankel, 2009). Poststratification can be preceded by multi-level regression (for so-called “Mr P” analysis) for partial pooling of information across strata before poststratification of the model predictions, which may be especially useful when some strata contain few data points (Gelman, 2007).

The most appropriate approach is likely to be question- and taxon-specific, varying with how much the species range extends across the region of interest. For example, it would usually not be important to upweight under-sampled regions where a species is rare, or even absent, when estimating trends in its total population size. If, however, the goal is to estimate trends in the average site-level population trend, then it would be important to up-weight data from under-sampled regions, even from where the species is rare. For instance, in the UK bat monitoring scheme, data are weighted to allow for the different sampling rates across England, Scotland and Wales in proportion to the ratio of non-upland area to number of sites surveyed for the relevant country (Bat Conservation Trust, 2023). However, this weighting is not applied to range restricted species, such as the serotine bat, *Eptesicus serotinus* that is only found in southern England.

(3) Imputation

Imputation involves replacing missing values in a dataset with plausible estimates. A range of imputation procedures have been developed, which can fill gaps in both response and predictor variables (Carpenter & Kenward, 2012). Imputation is probably the most flexible and widely used approach to account for missing data across ecology and beyond. In biodiversity modelling, missing values are more often concentrated in the response variable (i.e., the biodiversity value), hence imputation here can be equated with making model predictions at unsampled sites and times.

Imputation is already in use in biodiversity trend monitoring, especially to account for within-year and annual data gaps (Table 3). Early approaches used chain indices or route regression (Ter Braak *et al.*, 1992) or the Underhill index, using an expectation-maximisation algorithm (Underhill & Prysjones, 1994) designed for waterbirds (Rehfisch *et al.*, 2003). A range of further model-based approaches have been developed that fill data gaps using mean effects of site and year, e.g., to fill annual gaps using TRIM/birdSTATs, commonly used for bird indices (Lehikoinen *et al.*, 2016); or using splines e.g., to fill seasonal gaps in butterfly data (Schmucki *et al.*, 2016; Dennis *et al.*, 2016) or using ecological covariates (Dakki *et al.*, 2021). A Bayesian framework is especially useful for dealing with missing values in the response since they are naturally imputed with a full probability distribution during model fitting. Bayesian occupancy-detection models have been used to analyse opportunistic species observations from citizen science, with annual data gaps imputed before the predicted annual proportion of occupied sites is calculated (Outhwaite *et al.*, 2019). The flexibility of Bayesian models means they could also incorporate expert knowledge as priors as a way to help fill data gaps (Johnson *et al.*, 2023).

While imputation is already used to deal with annual and within-year gaps, it has been less often used to deal with spatial gaps when the focus is mean trend modelling of species' abundances or occurrences. An exception is studies of changes in species' range sizes, which use distribution models to predict the full distribution of a species at multiple time points, before change is assessed (Grattarola, Bowler & Keil, 2023). Monitoring schemes with large spatial coverage are also beginning to use distribution or abundance models to predict spatio-temporal patterns of abundance change across whole countries (e.g., eBird maps and BTO maps). In these cases, regression models fit to the available data make predictions at unsampled sites based on the effects of environmental covariates and/or spatial structure (Bush *et al.*, 2017; Ver Hoef *et al.*, 2021; Breivik *et al.*, 2021). Geostatistical methods also offer a range of interpolation methods for spatial data, including kriging, which are especially useful when there is a strong spatial pattern in the data (Ballesteros-Mejia *et al.*, 2013; Kreft & Jetz, 2007; Lin *et al.*, 2008).

V. Pro and cons of each solution

Table 3 Summary of the pros and cons of each approach to deal with missing data in biodiversity monitoring

Solution	Pros	Cons
Subsampling	- arguably the simplest approach, especially for spatial gaps - already a routine feature of many species distribution modelling protocols - aligns with rarefaction approaches used in community ecology	- could mean excluding a large amount of data, which may be unacceptable for citizen science and engaging/retaining volunteers - most protocols focus on a single dimension (e.g., filtering by geographic region) - more complex to implement when gaps are multi-dimensional or temporally varying
Weighting	- standard practise to deal with sample unrepresentativeness in other disciplines, especially social sciences	- poorly understood in ecology - diverse range of possible weighting techniques (Valliant, 2020; Boyd *et al.*, 2023b) but little ecological guidance available to help selection
Imputation	- suitable approach if missing data are within the environmental covariates as well as within the biodiversity response - offers the promise to generate the continuous space-time data cubes of the Essential Biodiversity Variable framework (Kissling *et al.*, 2018; Jetz *et al.*, 2019).	- becomes inefficient as missingness increases, e.g., when the number of unsampled locations/times is large - requires a good understanding of the ecological system to predict the missing biodiversity values

All of the approaches have the potential to reduce the bias in parameter estimates but differ in complexity, scope and typical practice (Table 3) (Little *et al.*, 2022; Collins *et al.*, 2001). Moreover, while we separated the methods into three categories for convenience, their distinctions are not absolute. For instance, subsampling essentially assigns those population units included in the subsample a weight of 1 and the remainder a weight of 0. Often, but not always, the reduction in bias due to application of the above solutions comes at a cost of increasing parameter uncertainty: the classic bias-variance trade-off (Hefley *et al.*, 2013). This is because subsampling directly reduces the sample size; weighting reduces the effective sample size; and imputation adds uncertainties via predictions at unsampled points. But this trade-off does not always apply; for instance, poststratification can lead to the dual benefits of reduced bias and increased precision depending on the choice of covariates (Little & Vartivarian, 2005).

Covariates used to account for data gaps are often called 'auxiliary variables' (Little *et al.*, 2022), which are typically not of central interest to the scientific questions but are included in one or more of the analysis steps for subsampling, weighting or imputing. The general recommendation from the missing data theory and survey sampling literature is to be generous when deciding which covariates to use to adjust for data gaps, considering covariates relating to the missingness (i.e., sampling effort in the context of biodiversity data gaps) to reduce bias and those related to the biodiversity outcome to reduce the variance (Collins *et al.*, 2001; Caughey *et al.*, 2020). It is worth noting, however, that selecting auxiliary variables on a purely correlative basis can increase bias in some circumstances (Thoemmes & Rose, 2014), and a safer strategy is to select them on theoretical grounds (Mohan & Pearl, 2021). When auxiliary variables are related to both the biodiversity outcome and the pattern of missingness, weighting approaches can reduce bias and improve precision (Little & Vartivarian, 2005). The success of any of the solutions, hence, critically depends on the choice of auxiliary variables (Little *et al.*, 2022). A recent study testing the use of weighting approaches to account for spatial biases in a reasonably well-understood ecological system found that the selected auxiliary variables had only limited success in mitigating bias (Boyd *et al.*, 2023b).

We illustrate some of these challenges and the application of each potential solution with a toy example of an abundance dataset with missing values (Fig. 5). We simulated a landscape in which a covariate (let's say representing 'habitat quality') affected both species abundance and the likelihood of a site being sampled. The analysis aimed to estimate the mean abundance of the species across all sites in the landscape. We varied the total fraction of sites that were sampled and the degree of knowledge available on the covariate affecting sampling/species (modelled as the correlation between the covariate involved in the data generation process and the covariate available to the modeller). We compared subsampling, weighting and imputation, which all used the available covariate data for adjustment. For subsampling, we subsampled one site at random at each habitat quality value. For weighing, we compared two approaches: fitting a weighted regression model using model-robust sandwich variance estimators or using a poststratification approach. For imputation, we fit a Bayesian model using JAGS in which NA values were inserted to represent the missing response data.

The results show that all methods do better than a naive approach that did not attempt to account for missingness in the estimation of the mean abundance (Fig. 5). Subsampling performed the worst, while weighting and imputation performed similarly. Poststratification tends to perform less well with a lower sampling fraction i.e., when the number of missing values was high (Fig. 5A), because the sample did not always contain all the habitat quality values found in the population and the weighing could not account for entirely unsampled regions. All models performed less well at the available covariate became a weaker proxy of the true driving covariate (Fig. 5B). In further simulations, we found that imputation performed poorer when there were additional covariates affecting species abundance and these covariates were not modelled, highlighting the importance of understanding the ecological system for imputation (Fig. S1). We do not intend this simulation to be exhaustive - rather to highlight the potential ways in which the availability of data and degree of knowledge about the factors causing bias and the availability of covariate data affects any attempts to account for missing data.

We point the reader towards some useful R packages and functions in the Supporting Information (Table S1).

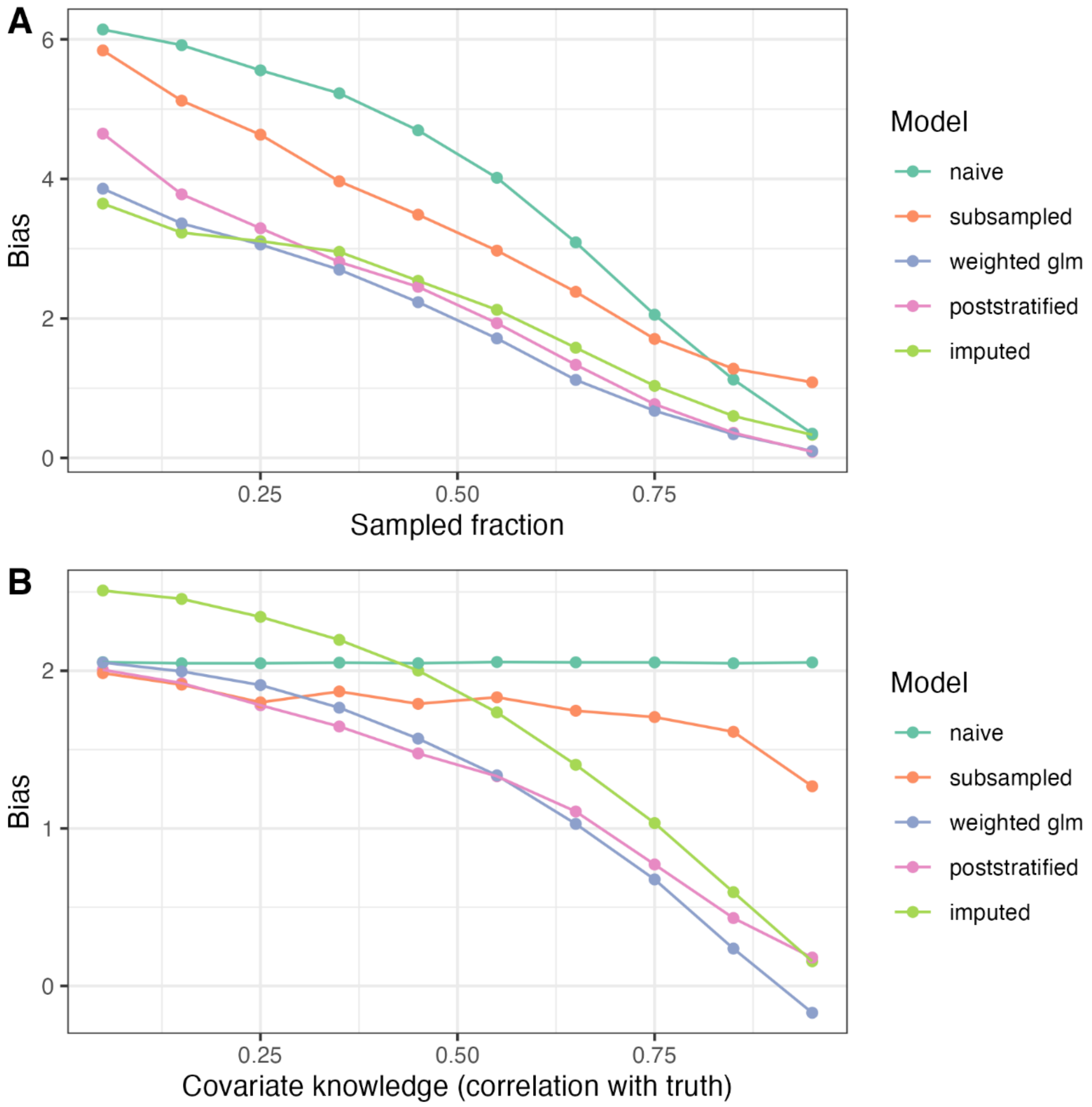

Fig. 5 The ability of missing data solutions to adjust for bias in biodiversity data.

We assumed a landscape of 400 cells and that a covariate affected both species abundance and the likelihood of a cell being sampled. In A: we vary the fraction of the cells that were sampled. In B: we vary the correlation between the true covariate and the covariate available for analysis, as measure of the available knowledge (correlation of 1 = perfect covariate and knowledge). The models to estimate the parameter of interest (mean abundance) were: naive (no correction); subsampled (cells were subsampled along the covariate gradient), weighted (two methods: weighted glm using the svyglm function, and weighted by poststratification, using postStratify, both in the survey package) and imputed (using JAGS to impute NAs in the response). Points show the mean bias (difference between model prediction and truth) across 100 independent runs.

VI. Dealing with Missing Not at Random

Dealing with Missing Not at Random (MNAR) is more challenging than dealing with the other data mechanisms (Little & Rubin, 2019). In this case, missingness is directly associated with unavailable data, which could be either the missing biodiversity values or missing covariate data that are not known to be important or are not measured/measurable. This makes MNAR especially difficult to diagnose (but see Conn et al. (2017) for suggestions) and model, since auxiliary variables are not available. MNAR can arise through a number of mechanisms in biodiversity monitoring data.

MNAR can be an outcome of preferential sampling - more intense sampling effort where the species is expected (Diggle, Menezes & Su, 2010; McClure & Rolek, 2023) - which leads to more missing values in places where the species is rare or absent. Preferential sampling can arise, for instance, if observers visit a location to specifically observe a species that others have observed there before (Laney *et al.*, 2021; Pennino *et al.*, 2019). Preferential sampling can also be a planned sampling strategy (Alessi *et al.*, 2023). For rare species, preferential sampling can be optimal when the goal is to estimate species detection probability and account for imperfect detection, since sufficient observations of the species can only be achieved by sampling where they are more common (Specht *et al.*, 2017). Similarly, it can be optimal to expend greater sampling effort where the species is common if the goal is to estimate trends in the total population size, since regions where the species is scarce are less important for the overall trend.

For organisms associated with specific habitats, such as wetland species or colonial seabirds, dedicated structured monitoring schemes target their habitats (McClure & Rolek, 2023). In such schemes, missing data outside of these core habitats are not considered part of the target population.

Typical approaches to modelling data allowing for MNAR are selection models (Heckman, 1979) and pattern-mixture models (Herzog and Rubin, 1983). Both model the joint distribution of the data and the data availability, but differ in how these processes are decomposed. Both also require making strong assumptions about the missing data mechanism, but can be useful to explore the consequences of plausible options as a sensitivity analysis (Little, 1995). In the ecological literature, preferential sampling has been modelled using marked point process models, which jointly model the sampling intensity (the points), the biodiversity value at those points (the marks) and the dependence between them (Conn *et al.*, 2017; Pennino *et al.*, 2019; Laxton *et al.*, 2023). Another approach to inference in a NMAR scenario is to use instrumental variables i.e., variables that affect the probability of sampling/data availability but are independent of the biodiversity variable of interest (Tchetgen & Wirth, 2017; Bailey, 2023). The challenge, however, is to identify such variables.

VII. General guidelines for dealing with biodiversity data gaps

Our review highlights the potential value of 'missing data thinking' when analysing biodiversity data. We argue that MCAR data gaps are unlikely in most biodiversity data contexts, which means that researchers will need to consider whether and how they deal with data gaps in their analysis. While it is premature to make very specific guidelines, we summarise here some of the considerations needed when dealing with data gaps in biodiversity data at different stages of data collection, analysis and reporting.

(1) Study design

For new monitoring schemes, planned data gaps that deviate from MCAR (i.e., a random sample) can be seen as opportunities rather than challenges since solutions are available to deal with missing data. Intentionally missing some data has been proposed for ethical or practical reasons in some study designs e.g., (Noble & Nakagawa, 2021; Herrera, 2019). In citizen science, planned data gaps could help increase uptake and avoid participant fatigue, especially caused by collecting difficult data. For instance, the UK Breeding Bird Survey includes an ‘upland rovers’ component in which the standard protocol is modified to allow for fewer visits to remote sites (Darvill *et al.*, 2020). Alternative study designs, such as wave missingness (Little & Rhemtulla, 2013) or a rotating panel design (Nielsen *et al.*, 2009) may increase the sustainability of long-term monitoring for some taxa or regions with few willing participants. But such an approach has to balance the cost of increased study design complexity and potential implications for the range of questions that can be addressed.

For existing monitoring schemes, data gaps may be filled, where possible, by promoting data collection in certain areas. Within citizen science projects, there is evidence that participants can be nudged to collect more data in regions identified as sampling priorities (Callaghan *et al.*, 2023; Callaghan *et al.*, 2019). Previous studies have identified sampling priorities in different ways; for instance, based on the expected influence of a data point (Callaghan *et al.*, 2019) or predictions based on species distribution models (Chiffard *et al.*, 2020). Since data collected by monitoring schemes are often collected for multiple purposes, the challenge is identifying the common set of sampling priorities.

For synthesis studies compiling data from independent studies, data mobilisation efforts may be tailored to improve sample representativeness of the target population, by expending more effort to under-sampled units. This could be informed by exploring the transferability of model predictions across spatial or temporal units based on currently available data (Spake *et al.*, 2022). Regions with high transferability may represent appropriate sampling strata to guide mobilisation efforts. Moreover, these sampling strata may inform the adjustment for data gaps in subsequent modelling of the population mean.

(2) Evaluating and reporting missingness

Developing a causal model (e.g., using a DAG) of the factors affecting sampling probability and biodiversity can be useful first step to identify auxiliary variables for adjusting data gaps – variables linked to both sampling probability and biodiversity are those creating bias (Mohan & Pearl, 2021). As far as possible, data should then be collected on the covariates that are likely to explain missingness. Statistical models can be used to test whether covariates that are associated with missingness are also associated with biodiversity patterns, though of course this is only possible in the sampled data. Unplanned missingness in structured schemes could be investigated by disseminating follow-up surveys to participants to determine their reasons for missed surveys. Follow-on data collection, e.g., with paid surveys, in regions or times of missing data could also help understand whether there are fundamental differences in biodiversity patterns between the original dataset and the extended dataset.

Missingness, and how it is dealt with, tends to be insufficiently reported in biodiversity trend analyses. Some reporting frameworks for missing data have been developed for other disciplines (Lee *et al.*, 2021) but are in their early stages in ecology (Boyd *et al.*, 2022). At a minimum, we propose that missingness can be reported in terms of the proportion of sampling units that are spatial, annual and within-year gaps, and the number of unplanned gaps for structured monitoring schemes (Fig. 1). Visualizations of the distributions of covariates in sampled and non-sampled times/sites could also effectively highlight key systematic differences.

(3) Modelling to account for data gaps

The impact of data gaps depends on multiple factors: whether the factors affecting missingness are independent of the factors affecting biodiversity and biodiversity itself; the ecological questions being asked and which covariates are available and included in the analysis. Because of this, potential impacts of missingness have to be considered for each species-question-dataset combination. A dataset *per se* is not biased. Subsampling, weighting and imputation all have the potential to reduce bias caused by data gaps. Many, but not all, solutions will navigate the bias-variance trade-off. Weighting is probably the most under-used in ecology and could be applied more often, especially to account for spatial gaps when the goal is estimating mean abundance or abundance trends. Imputation methods offer the potential to fill in spatio-temporal gaps to generate the space-time data cubes of the Essential Biodiversity Framework (Kissling *et al.*, 2018), but its success is dependent on the ability to model the variation in the biodiversity response. Since available covariates are likely to be only partly successful in reducing bias, sensitivity analysis could be help explore how different assumptions of missingness change the model coefficients and predictions, and the uncertainties of them (Little, 1995; Leurent *et al.*, 2018). For some contexts, it might be more statistically efficient and ecologically interpretable to redefine the target region of interest to a region with fewer data gaps.

VIII. Conclusions

(1) Biodiversity datasets containing information on species' occurrences and abundances are rapidly growing in size, but data gaps are not necessarily closing. Nonetheless, big biodiversity datasets are invaluable for a broad range of basic and applied questions, and increasingly for policy-relevant questions about the status and trends of biodiversity at large-scales. Heterogeneity in sampling efforts - whether by citizen scientists or scientists - creates different types of data gaps in the available data. Such data gaps are among the biggest hindrances to making use of these growing data sources for large-scale inferences.

(2) We show how 'missing data thinking' can help decide whether a data gap is problematic in a given context and provides directions on possible solutions. We show that an important determinant of bias is whether factors affecting sampling effort are correlated with those affecting biodiversity: shared covariates affecting sampling effort and biodiversity have the potential to lead to biased analyses if not taken into account.

(3) Multiple approaches are available to account for missing data but they depend on knowledge and availability of covariates associated with missingness. A lack of training for ecologists in commonly employed approaches in other disciplines has meant there are few standard practices in ecology to deal with gaps. We highlight multiple methods that are ripe for comparison across different ecological problems.

(4) At the same, statistical solutions can only go so far, closing data gaps with more coordinated data collection across monitoring stakeholders is also important as we move forwards.

IX. Acknowledgements

This work was supported by the Terrestrial Surveillance Development and Analysis partnership of the UK Centre for Ecology & Hydrology, British Trust for Ornithology and the Joint Nature Conservation Committee (JNCC) and by the Natural Environment Research Council award number NE/R016429/1 as part of the UK-SCAPE programme delivering National Capability. RJB was also supported by the NERC Exploring the Frontiers award number NE/X010384/1 "Biodiversity indicators from nonprobability samples: Interdisciplinary learning for science and society". We also thank the task group of the Terrestrial Evidence Partnership of Partnerships (hosted by JNCC) for discussion during an outline workshop on topic of the paper.

X. Supporting Information

Table S1 Selected R tools that can help with missing data problems and their potential application for use in biodiversity research.

Fig. S1 Extended analysis of missing data solutions when additional covariates affect the biodiversity response.

R script for the example solution simulations (Fig. 5) are here: https://github.com/bowlerbear/dataGaps

IX. 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. *Ecography* **33**(6), 1028-1037.

ALESSI, N., BONARI, G., ZANNINI, P., JIMENEZ-ALFARO, B., AGRILLO, E., ATTORRE, F., CANULLO, R., CASELLA, L., CERVELLINI, M., CHELLI, S., DI MUSCIANO, M., GUARINO, R., MARTELLOS, S., MASSIMI, M., VENANZONI, R., ZERBE, S. & CHIARUCCI, A. (2023). Probabilistic and preferential sampling approaches offer integrated perspectives of Italian forest diversity. *Journal of Vegetation Science* **34**(1).

AMANO, T., LAMMING, J. D. L. & SUTHERLAND, W. J. (2016). Spatial Gaps in Global Biodiversity Information and the Role of Citizen Science. *Bioscience* **66**(5), 393-400.

AMANO, T., OKAMURA, H., CARRIZO, S. F. & SUTHERLAND, W. J. (2012). Hierarchical models for smoothed population indices: The importance of considering variations in trends of count data among sites. *Ecological Indicators* **13**(1), 243-252.

AUBRY, P. & FRANCESIAZ, C. (2022). On comparing design-based estimation versus model-based prediction to assess the abundance of biological populations. *Ecological Indicators* **144**.

BAILEY, M. A. (2023). A New Paradigm for Polling. *Harvard Data Science Review* **5**(3).

BAKER, D. J., MACLEAN, I. M. D., GOODALL, M. & GASTON, K. J. (2022). Correlations between spatial sampling biases and environmental niches affect species distribution models. *Global Ecology and Biogeography* **31**(6), 1038-1050.

BALLESTEROS-MEJIA, L., KITCHING, I. J., JETZ, W., NAGEL, P. & BECK, J. (2013). Mapping the biodiversity of tropical insects: species richness and inventory completeness of African sphingid moths. *Global Ecology and Biogeography* **22**(5), 586-595.

BARBET-MASSIN, M., JIGUET, F., ALBERT, C. H. & THUILLER, W. (2012). Selecting pseudo-absences for species distribution models: how, where and how many? *Methods in Ecology and Evolution* **3**(2), 327-338.

BARTLETT, J. W., HAREL, O. & CARPENTER, J. R. (2015). Asymptotically Unbiased Estimation of Exposure Odds Ratios in Complete Records Logistic Regression. *American Journal of Epidemiology* **182**(8), 730-736.

BAT CONSERVATION TRUST. (2023). The National Bat Monitoring Programme Annual Report 2022. Bat Conservation Trust, London. Available at www.bats.org.uk/our-work/national-bat-monitoringprogramme/reports/nbmp-annual-report.

BATTAGLIA, M. P., HOAGLIN, D. C. & FRANKEL, M. R. (2009). Practical Considerations in Raking Survey Data. *Survey Practice* **2**(5).

BAYRAKTAROV, E., EHMKE, G., O'CONNOR, J., BURNS, E. L., NGUYEN, H. A., MCRAE, L., POSSINGHAM, H. P. & LINDENMAYER, D. B. (2019). Do Big Unstructured Biodiversity Data Mean More Knowledge? *Frontiers in Ecology and Evolution* **6**.

BINLEY, A. D. & BENNETT, J. R. (2023). The data double standard. *Methods in Ecology and Evolution* **14**(6), 1389-1397.

BLED, F., SAUER, J., PARDIECK, K., DOHERTY, P. & ROYLE, J. A. (2013). Modeling Trends from North American Breeding Bird Survey Data: A Spatially Explicit Approach. *Plos One* **8**(12).

BOAKES, E. H., MCGOWAN, P. J. K., FULLER, R. A., DING, C. Q., CLARK, N. E., O'CONNOR, K. & MACE, G. M. (2010). Distorted Views of Biodiversity: Spatial and Temporal Bias in Species Occurrence Data. *Plos Biology* **8**(6).

BORIA, R. A., OLSON, L. E., GOODMAN, S. M. & ANDERSON, R. P. (2014). Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. *Ecological Modelling* **275**, 73-77.

Bowler, D. E., Callaghan, C. T., Bhandari, N., Henle, K., Barth, M. B., Koppitz, C., Klenke, R., Winter, M., Jansen, F., Bruelheide, H. & Bonn, A. (2022). Temporal trends in the spatial bias of species occurrence records. *Ecography* **2022**(8).

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

Boyd, R. J., Powney, G. D. & Pescott, O. L. (2023a). We need to talk about nonprobability samples. *Trends in Ecology & Evolution* **38**(6), 521-531.

Boyd, R. J., Stewart, G. B. & Pescott, O. L. (2023b). Descriptive inference using large, unrepresentative nonprobability samples: An 1introduction for ecologists. *EcoEvoRxiv*.

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-+.

Breivik, O. N., Aanes, F., Sovik, G., Aglen, A., Mehl, S. & Johnsen, E. (2021). Predicting abundance indices in areas without coverage with a latent spatio-temporal Gaussian model. *Ices Journal of Marine Science* **78**(6), 2031-2042.

Buckland, S. T., Baillie, S. R., Dick, J. M., Elston, D. A., Magurran, A. E., Scott, E. M., Smith, R. I., Somerfield, P. J., Studeny, A. C. & Watt, A. (2012). How should regional biodiversity be monitored? *Environmental and Ecological Statistics* **19**(4), 601-626.

Buckland, S. T. & Johnston, A. (2017). Monitoring the biodiversity of regions: Key principles and possible pitfalls. *Biological Conservation* **214**, 23-34.

Bush, A., Sollmann, R., Wilting, A., Bohmann, K., Cole, B., Balzter, H., Martius, C., Zlinszky, A., Calvignac-Spencer, S., Cobbold, C. A., Dawson, T. P., Emerson, B. C., Ferrier, S., Gilbert, M. T. P., Herold, M., Jones, L., Leendertz, F. H., Matthews, L., Millington, J. D. A., Olson, J. R., Ovaskainen, O., Raffaelli, D., Reeve, R., Rodel, M. O., Rodgers, T. W., Snape, S., Visseren-Hamakers, I., Vogler, A. P., White, P. C. L., Wooster, M. J. & Yu, D. W. (2017). Connecting Earth observation to high-throughput biodiversity data. *Nature Ecology & Evolution* **1**(7).

Callaghan, C. T., Poore, A. G. B., Major, R. E., Rowley, J. J. L. & Cornwell, W. K. (2019). Optimizing future biodiversity sampling by citizen scientists. *Proceedings of the Royal Society B-Biological Sciences* **286**(1912).

Callaghan, C. T., Thompson, M., Woods, A., Poore, A. G. B., Bowler, D. E., Samonte, F., Rowley, J. J. L., Roslan, N., Kingsford, R. T., Cornwell, W. K. & Major, R. E. (2023). Experimental evidence that behavioral nudges in citizen science projects can improve biodiversity data. *Bioscience* **73**(4), 302-313.

Callcutt, K., Croft, S. & Smith, G. C. (2018). Predicting population trends using citizen science data: do subsampling methods produce reliable estimates for mammals? *European Journal of Wildlife Research* **64**(3).

Cardinale, B. J., Gonzalez, A., Allington, G. R. H. & Loreau, M. (2018). Is local biodiversity declining or not? A summary of the debate over analysis of species richness time trends. *Biological Conservation* **219**, 175-183.

Carpenter, J. & Kenward, M. (2012). *Multiple imputaion and its application*. Wiley.

Caughey, D., Berinsky, A. J., Chatfield, S., Hartman, E., Schickler, E. & Sekhon, J. S. (2020). *Target estimation and adjustment weighting for survey nonresponse and sampling bias*. Cambridge University Press.

Chandler, M., See, L., Copas, K., Bonde, A. M. Z., Lopez, B. C., Danielsen, F., Legind, J. K., Masinde, S., Miller-Rushing, A. J., Newman, G., Rosemartin, A. & Turak, E. (2017). Contribution of citizen science towards international biodiversity monitoring. *Biological Conservation* **213**, 280-294.

Chevalier, M., Zarzo-Arias, A., Guelat, J., Mateo, R. G. & Guisan, A. (2022). Accounting for niche truncation to improve spatial and temporal predictions of species distributions. *Frontiers in Ecology and Evolution* **10**.

Chiffard, J., Marciau, C., Yoccoz, N. G., Mouillot, F., Duchateau, S., Nadeau, I., Fontanilles, P. & Besnard, A. (2020). Adaptive niche-based sampling to improve ability to find rare and elusive species: Simulations and field tests. *Methods in Ecology and Evolution* **11**(8), 899-909.

Collins, L. M., Schafer, J. L. & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods* **6**(4), 330-351.

Conn, P. B., Thorson, J. T. & Johnson, D. S. (2017). Confronting preferential sampling when analysing population distributions: diagnosis and model-based triage. *Methods in Ecology and Evolution* **8**(11), 1535-1546.

Courter, J. R., Johnson, R. J., Stuyck, C. M., Lang, B. A. & Kaiser, E. W. (2013). Weekend bias in Citizen Science data reporting: implications for phenology studies. *International Journal of Biometeorology* **57**(5), 715-720.

Cretois, B., Simmonds, E. G., Linnell, J. D. C., van Moorter, B., Rolandsen, C. M., Solberg, E. J., Strand, O., Gundersen, V., Roer, O. & Rod, J. K. (2021). Identifying and correcting spatial bias in opportunistic citizen science data for wild ungulates in Norway. *Ecology and Evolution* **11**(21), 15191-15204.

Dakki, M., Robin, G., Suet, M., Qninba, A., El Agbani, M. A., Ouassou, A., El Hamoumi, R., Azafzaf, H., Rebah, S., Feltrup-Azafzaf, C., Hamouda, N., Ibrahim, W. A. L., Asran, H. H., Elhady, A. A., Ibrahim, H., Etayeb, K., Bouras, E., Saied, A., Glidan, A., Habib, B. M., Sayoud, M. S., Bendjedda, N., Dami, L., Deschamps, C., Gaget, E., Mondain-Monval, J. Y. & du Rau, P. D. (2021). Imputation of incomplete large-scale monitoring count data via penalized estimation. *Methods in Ecology and Evolution* **12**(6), 1031-1039.

Dambly, L. I., Jones, K. E., Boughey, K. L. & Isaac, N. J. B. (2021). Observer retention, site selection and population dynamics interact to bias abundance trends in bats. *Journal of Applied Ecology* **58**(2), 236-247.

Darvill, B., Harris, S. J., Martay, B., Wilson, M. & Gillings, S. (2020). Delivering robust population trends for Scotland’s widespread breeding birds. *Scottish Birds* **40**(4), 297-304.

Dennis, E. B., Morgan, B. J. T., Freeman, S. N., Brereton, T. M. & Roy, D. B. (2016). A Generalized Abundance Index for Seasonal Invertebrates. *Biometrics* **72**(4), 1305-1314.

Diekert, F., Munzinger, S., Schulemann-Maier, G. & Stadtler, L. (2023). Explicit incentives increase citizen science recordings. *Conservation Letters*.

Diggle, P. J., Menezes, R. & Su, T. L. (2010). Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society Series C-Applied Statistics* **59**, 191-232.

Dornelas, M., Gotelli, N. J., McGill, B., Shimadzu, H., Moyes, F., Sievers, C. & Magurran, A. E. (2014). Assemblage Time Series Reveal Biodiversity Change but Not Systematic Loss. *Science* **344**(6181), 296-299.

Evans, D. M. & Day, K. R. (2002). Hunting disturbance on a large shallow lake: the effectiveness of waterfowl refuges. *Ibis* **144**(1), 2-8.

Fink, D., Auer, T., Johnston, A., Ruiz-Gutierrez, V., Hochachka, W. M. & Kelling, S. (2020). Modeling avian full annual cycle distribution and population trends with citizen science data. *Ecological Applications* **30**(3).

Fithian, W., Elith, J., Hastie, T. & Keith, D. A. (2015). Bias correction in species distribution models: pooling survey and collection data for multiple species. *Methods in Ecology and Evolution* **6**(4), 424-438.

Forister, M. L., Black, S. H., Elphick, C. S., Grames, E. M., Halsch, C. A., Schultz, C. B. & Wagner, D. L. (2023). Missing the bigger picture: Why insect monitoring programs are limited in their ability to document the effects of habitat loss. *Conservation Letters* **16**(3).

Frair, J. L., Nielsen, S. E., Merrill, E. H., Lele, S. R., Boyce, M. S., Munro, R. H. M., Stenhouse, G. B. & Beyer, H. L. (2004). Removing GPS collar bias in habitat selection studies. *Journal of Applied Ecology* **41**(2), 201-212.

Fraisl, D., Campbell, J., See, L., Wehn, U., Wardlaw, J., Gold, M., Moorthy, I., Arias, R., Piera, J., Oliver, J. L., Maso, J., Penker, M. & Fritz, S. (2020). Mapping citizen science contributions to the UN sustainable development goals. *Sustainability Science* **15**(6), 1735-1751.

Fretwell, P. T., Scofield, P. & Phillips, R. A. (2017). Using super-high resolution satellite imagery to census threatened albatrosses. *Ibis* **159**(3), 481-490.

Garcia-Rosello, E., Guisande, C., Manjarres-Hernandez, A., Gonzalez-Dacosta, J., Heine, J., Pelayo-Villamil, P., Gonzalez-Vilas, L., Vari, R. P., Vaamonde, A., Granado-Lorencio, C. & Lobo, J. M. (2015). Can we derive macroecological patterns from primary Global Biodiversity Information Facility data? *Global Ecology and Biogeography* **24**(3), 335-347.

Gaul, W., Sadykova, D., White, H. J., Leon-Sanchez, L., Caplat, P., Emmerson, M. C. & Yearsley, J. M. (2022). Modelling the distribution of rare invertebrates by correcting class imbalance and spatial bias. *Diversity and Distributions* **28**(10), 2171-2186.

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

Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science* **22**(2), 153-164.

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.

Grattarola, F., Bowler, D. E. & Keil, P. (2023). Integrating presence-only and presence-absence data to model changes in species geographic ranges: An example in the Neotropics. *Journal of Biogeography*.

Gregory, R. D., van Strien, A., Vorisek, P., Meyling, A. W. G., Noble, D. G., Foppen, R. P. B. & Gibbons, D. W. (2005). Developing indicators for European birds. *Philosophical Transactions of the Royal Society B-Biological Sciences* **360**(1454), 269-288.

Guo, Q. F., Chen, A. P., Crockett, E. T. H., Atkins, J. W., Chen, X. W. & Fei, S. L. (2023). Integrating gradient with scale in ecological and evolutionary studies. *Ecology* **104**(4).

Hefley, T. J., Tyre, A. J., Baasch, D. M. & Blankenship, E. E. (2013). Nondetection sampling bias in marked presence-only data. *Ecology and Evolution* **3**(16), 5225-5236.

Herrera, C. M. (2019). Complex long-term dynamics of pollinator abundance in undisturbed Mediterranean montane habitats over two decades. *Ecological Monographs* **89**(1).

Hertzog, L. R., Besnard, A. & Jay-Robert, P. (2014). Field validation shows bias-corrected pseudo-absence selection is the best method for predictive species-distribution modelling. *Diversity and Distributions* **20**(12), 1403-1413.

Hof, A. R. & Bright, P. W. (2016). Quantifying the long-term decline of the West European hedgehog in England by subsampling citizen-science datasets. *European Journal of Wildlife Research* **62**(4), 407-413.

Hossie, T. J., Gobin, J. & Murray, D. L. (2021). Confronting Missing Ecological Data in the Age of Pandemic Lockdown. *Frontiers in Ecology and Evolution* **9**.

Isaac, N. J. B. & Pocock, M. J. O. (2015). Bias and information in biological records. *Biological Journal of the Linnean Society* **115**(3), 522-531.

Jetz, W., McGeoch, M. A., Guralnick, R., Ferrier, S., Beck, J., Costello, M., Fernandez, M., Geller, G. N., Keil, P., Merow, C., Meyer, C., Muller-Karger, F. E., Pereira, H. M., Regan, E. C., Schmeller, D. S. &

Turak, E. (2019). Essential biodiversity variables for mapping and monitoring species populations. *Nature Ecology & Evolution* **3**(4), 539-551.

Johnson, T. F., Isaac, N. J. B., Paviolo, A. & González-Suárez, M. (2023). Socioeconomic factors predict population changes of large carnivores better than climate change or habitat loss. . 2023 Jan 24;14(1):74. doi: . PMID: 36693827; PMCID: PMC9873912. *Nat Commun* **14**(1), 74.

Johnston, A., Hochachka, W. M., Strimas-Mackey, M. E., Gutierrez, V. R., Robinson, O. J., Miller, E. T., Auer, T., Kelling, S. T. & Fink, D. (2021). Analytical guidelines to increase the value of community science data: An example using eBird data to estimate species distributions. *Diversity and Distributions* **27**(7), 1265-1277.

Johnston, A., Moran, N., Musgrove, A., Fink, D. & Baillie, S. R. (2020). Estimating species distributions from spatially biased citizen science data. *Ecological Modelling* **422**.

Kissling, W. D., Ahumada, J. A., Bowser, A., Fernandez, M., Fernandez, N., Garcia, E. A., Guralnick, R. P., Isaac, N. J. B., Kelling, S., Los, W., McRae, L., Mihoub, J. B., Obst, M., Santamaria, M., Skidmore, A. K., Williams, K. J., Agosti, D., Amariles, D., Arvanitidis, C., Bastin, L., De Leo, F., Egloff, W., Elith, J., Hobern, D., Martin, D., Pereira, H. M., Pesole, G., Peterseil, J., Saarenmaa, H., Schigel, D., Schmeller, D. S., Segata, N., Turak, E., Uhlir, P. F., Wee, B. & Hardisty, A. R. (2018). Building essential biodiversity variables (EBVs) of species distribution and abundance at a global scale. *Biological Reviews* **93**(1), 600-625.

Kreft, H. & Jetz, W. (2007). Global patterns and determinants of vascular plant diversity. *Proceedings of the National Academy of Sciences of the United States of America* **104**(14), 5925-5930.

La Sorte, F. A. & Somveille, M. (2020). Survey completeness of a global citizen-science database of bird occurrence. *Ecography* **43**(1), 34-43.

Laney, J. A., Hallman, T. A., Curtis, J. R. & Robinson, W. D. (2021). The influence of rare birds on observer effort and subsequent rarity discovery in the American birdwatching community. *Peerj* **9**.

Laxton, M. R., de Rivera, O. R., Soriano-Redondo, A. & Illian, J. B. (2023). Balancing structural complexity with ecological insight in Spatio-temporal species distribution models. *Methods in Ecology and Evolution* **14**(1), 162-172.

Lee, K. J., Tilling, K. M., Cornish, R. P., Little, R. J. A., Bell, M. L., Goetghebeur, E., Hogan, J. W., Carpenter, J. R. & Initiative, S. (2021). Framework for the treatment and reporting of missing data in observational studies: The Treatment And Reporting of Missing data in Observational Studies framework. *Journal of Clinical Epidemiology* **134**, 79-88.

Lehikoinen, A., Foppen, R. P. B., Heldbjerg, H., Lindstrom, A., van Manen, W., Piirainen, S., van Turnhout, C. A. M. & Butchart, S. H. M. (2016). Large-scale climatic drivers of regional winter bird population trends. *Diversity and Distributions* **22**(11), 1163-1173.

Leurent, B., Gomes, M., Faria, R., Morris, S., Grieve, R. & Carpenter, J. R. (2018). Sensitivity Analysis for Not-at-Random Missing Data in Trial-Based Cost-Effectiveness Analysis: A Tutorial. *Pharmacoeconomics* **36**(8), 889-901.

Li, L. L., Shen, C. Y., Li, X. C. & Robins, J. M. (2013). On weighting approaches for missing data. *Statistical Methods in Medical Research* **22**(1), 14-30.

Lin, Y. P., Yeh, M. S., Deng, D. P. & Wang, Y. C. (2008). Geostatistical approaches and optimal additional sampling schemes for spatial patterns and future sampling of bird diversity. *Global Ecology and Biogeography* **17**(2), 175-188.

Little, R. J., Carpenter, J. R., Lee, K. J. & Initiative, S. (2022). A Comparison of Three Popular Methods for Handling Missing Data: Complete-Case Analysis, Inverse Probability Weighting, and Multiple Imputation. *Sociological Methods & Research*.

Little, R. J. & Rubin, D. B. (2019). *Statistical Analysis with Missing Data (3rd Edition)*. Wiley.

Little, R. J. & Vartivarian, S. (2005). Does Weighting for Nonresponse Increase the Variance of Survey Means? . *Survey Methodology* **31**(2), 161-168.

Little, R. J. A. (1988). A Test of Missing Completely at Random for Multivariate Data with Missing Values. *Journal of the American Statistical Association* **83**(404), 1198-1202.

Little, R. J. A. (1995). Modeling the Drop-out Mechanism in Repeated-Measures Studies. *Journal of the American Statistical Association* **90**(431), 1112-1121.

Little, T. D. & Rhemtulla, M. (2013). Planned Missing Data Designs for Developmental Researchers. *Child Development Perspectives* **7**(4), 199-204.

Lopucki, R., Kiersztyn, A., Pitucha, G. & Kitowski, I. (2022). Handling missing data in ecological studies: Ignoring gaps in the dataset can distort the inference. *Ecological Modelling* **468**.

Mandeville, C. P., Nilsen, E. B. & Finstad, A. G. (2022). Spatial distribution of biodiversity citizen science in a natural area depends on area accessibility and differs from other recreational area use. *Ecological Solutions and Evidence* **3**(4).

Marsh, D. M. & Cosentino, B. J. (2019). Causes and consequences of non-random drop-outs for citizen science projects: lessons from the North American amphibian monitoring program. *Freshwater Science* **38**(2), 292-302.

Matutini, F., Baudry, J., Pain, G., Sineau, M. & Pithon, J. (2021). How citizen science could improve species distribution models and their independent assessment. *Ecology and Evolution* **11**(7), 3028-3039.

McClure, C. J. W. & Rolek, B. W. (2023). Pitfalls arising from site selection bias in population monitoring defy simple heuristics. *Methods in Ecology and Evolution* **14**(6), 1489-1499.

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.

Meng, X. L. (2022). Comments on "Statistical inference with non-probability survey samples" - Miniaturizing data defect correlation: A versatile strategy for handling non-probability samples. *Survey Methodology* **48**(2), 339-360.

Mohan, K. & Pearl, J. (2021). Graphical Models for Processing Missing Data. *Journal of the American Statistical Association* **116**(534), 1023-1037.

Nakagawa, S. & Freckleton, R. P. (2008). Missing inaction: the dangers of ignoring missing data. *Trends in Ecology & Evolution* **23**(11), 592-596.

Nielsen, S. E., Haughland, D. L., Bayne, E. & Schieck, J. (2009). Capacity of large-scale, long-term biodiversity monitoring programmes to detect trends in species prevalence. *Biodiversity and Conservation* **18**(11), 2961-2978.

Noble, D. W. A. & Nakagawa, S. (2021). Planned missing data designs and methods: Options for strengthening inference, increasing research efficiency and improving animal welfare in ecological and evolutionary research. *Evolutionary Applications* **14**(8), 1958-1968.

Nunez-Penichet, C., Cobos, M. E., Soberon, J., Gueta, T., Barve, N., Barve, V., Navarro-Siguenza, A. G. & Peterson, A. T. (2022). Selection of sampling sites for biodiversity inventory: Effects of environmental and geographical considerations. *Methods in Ecology and Evolution* **13**(7), 1595-1607.

Outhwaite, C. L., Powney, G. D., August, T. A., Chandler, R. E., Rorke, S., Pescott, O. L., Harvey, M., Roy, H. E., Fox, R., Roy, D. B., Alexander, K., Ball, S., Bantock, T., Barber, T., Beckmann, B. C., Cook, T., Flanagan, J., Fowles, A., Hammon, P., Harvey, P., Hepper, D., Hubble, D., Kramer, J., Lee, P., MacAdam, C., Morris, R., Norris, A., Palmer, S., Plant, C. W., Simkin, J., Stubbs, A., Sutton, P., Telfer, M., Wallace, I. & Isaac, N. J. B. (2019). Annual estimates of occupancy for bryophytes, lichens and invertebrates in the UK, 1970-2015. *Scientific Data* **6**.

Pennino, M. G., Paradinas, I., Illian, J. B., Munoz, F., Bellido, J. M., Lopez-Quilez, A. & Conesa, D. (2019). Accounting for preferential sampling in species distribution models. *Ecology and Evolution* **9**(1), 653-663.

Pocock, M. J., Tweddle, J. C., Savage, J., Robinson, L. D. & Roy, H. E. (2017). The diversity and evolution of ecological and environmental citizen science. *Plos One* **12**(4).

RAGHUNATHAN, T. E. (2004). What do we do with missing data? Some options for analysis of incomplete data. *Annual Review of Public Health* **25**, 99-117.

RAPACCIUOLO, G., YOUNG, A. & JOHNSON, R. (2021). Deriving indicators of biodiversity change from unstructured community-contributed data. *Oikos* **130**(8), 1225-1239.

REHFISCH, M. M., AUSTIN, G. E., ARMITAGE, M. J. S., ATKINSON, P. W., HOLLOWAY, S. J., MUSGROVE, A. J. & POLLITT, M. S. (2003). Numbers of wintering waterbirds in Great Britain and the Isle of Man (1994/1995-1998/1999): II. Coastal waders (Charadrii). *Biological Conservation* **112**(3), 329-341.

ROBINSON, O. J., RUIZ-GUTIERREZ, V., REYNOLDS, M. D., GOLET, G. H., STRIMAS-MACKEY, M. & FINK, D. (2020). Integrating citizen science data with expert surveys increases accuracy and spatial extent of species distribution models. *Diversity and Distributions* **26**(8), 976-986.

RUBIN, D. B. (1976). Inference and Missing Data. *Biometrika* **63**(3), 581-590.

SCHMUCKI, R., PE'ER, G., ROY, D. B., STEFANESCU, C., VAN SWAAY, C. A. M., OLIVER, T. H., KUUSSAARI, M., VAN STRIEN, A. J., RIES, L., SETTELE, J., MUSCHE, M., CARNICER, J., SCHWEIGER, O., BRERETON, T. M., HARPKE, A., HELIOLA, J., KUHN, E. & JULLIARD, R. (2016). A regionally informed abundance index for supporting integrative analyses across butterfly monitoring schemes. *Journal of Applied Ecology* **53**(2), 501-510.

SEAMAN, S. R. & WHITE, I. R. (2013). Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research* **22**(3), 278-295.

SMITH, T. M. F. (1976). The Foundations of Survey Sampling *Journal of the Royal Statistical Societiy Series A* **139**, 183-204.

SPAKE, R., BOWLER, D. E., CALLAGHAN, C. T., BLOWES, S. A., DONCASTER, C. P., ANTAO, L. H., NAKAGAWA, S., MCELREATH, R. & CHASE, J. M. (2023). Understanding 'it depends' in ecology: a guide to hypothesising, visualising and interpreting statistical interactions. *Biological Reviews* **98**(4), 983-1002.

SPAKE, R., O'DEA, R. E., NAKAGAWA, S., DONCASTER, C. P., RYO, M., CALLAGHAN, C. T. & BULLOCK, J. M. (2022). Improving quantitative synthesis to achieve generality in ecology. *Nature Ecology & Evolution* **6**(12), 1818-1828.

SPECHT, H. M., REICH, H. T., IANNARILLI, F., EDWARDS, M. R., STAPLETON, S. P., WEEGMAN, M. D., JOHNSON, M. K., YOHANNES, B. J. & ARNOLD, T. W. (2017). Occupancy surveys with conditional replicates: An alternative sampling design for rare species. *Methods in Ecology and Evolution* **8**(12), 1725-1734.

STEEN, V. A., ELPHICK, C. S. & TINGLEY, M. W. (2019). An evaluation of stringent filtering to improve species distribution models from citizen science data. *Diversity and Distributions* **25**(12), 1857-1869.

STEEN, V. A., TINGLEY, M. W., PATON, P. W. C. & ELPHICK, C. S. (2021). Spatial thinning and class balancing: Key choices lead to variation in the performance of species distribution models with citizen science data. *Methods in Ecology and Evolution* **12**(2), 216-226.

SULLIVAN, B. L., PHILLIPS, T., DAYER, A. A., WOOD, C. L., FARNSWORTH, A., ILIFF, M. J., DAVIES, I. J., WIGGINS, A., FINK, D., HOCHACHKA, W. M., RODEWALD, A. D., ROSENBERG, K. V., BONNEY, R. & KELLING, S. (2017). Using open access observational data for conservation action: A case study for birds. *Biological Conservation* **208**, 5-14.

TCHETGEN, E. J. T. & WIRTH, K. E. (2017). A general instrumental variable framework for regression analysis with outcome missing not at random. *Biometrics* **73**(4), 1123-1131.

TER BRAAK, C. J. F., VAN STRIEN, A. J., MEIJER, R. & VERSTRAEL, T. J. (1992). *Analysis of monitoring data with many missing values: which method?*

THOEMMES, F. & ROSE, N. (2014). A Cautious Note on Auxiliary Variables That Can Increase Bias in Missing Data Problems. *Multivariate Behavioral Research* **49**(5), 443-459.

TROUDET, J., GRANDCOLAS, P., BLIN, A., VIGNES-LEBBE, R. & LEGENDRE, F. (2017). Taxonomic bias in biodiversity data and societal preferences. *Scientific Reports* **7**.

TULLOCH, A. I. T., MUSTIN, K., POSSINGHAM, H. P., SZABO, J. K. & WILSON, K. A. (2013). To boldly go where no volunteer has gone before: predicting volunteer activity to prioritize surveys at the landscape scale. *Diversity and Distributions* **19**(4), 465-480.

UNDERHILL, L. G. & PRYSJONES, R. P. (1994). Index Numbers for Waterbird Populations .1. Review and Methodology. *Journal of Applied Ecology* **31**(3), 463-480.

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).

VALLIANT, R. (2020). Comparing Alternatives for Estimation from Nonprobability Samples. *Journal of Survey Statistics and Methodology* **8**(2), 231-263.

VAN DEN BRAKEL, J. A. & BETHLEHEM, J. (2008). *Model-based estimation for official statistics*. Statistics Netherlands, Voorburg/Heerlen.

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.

VER HOEF, J. M., JOHNSON, D., ANGLISS, R. & HIGHAM, M. (2021). Species density models from opportunistic citizen science data. *Methods in Ecology and Evolution* **12**(10), 1911-1925.

WU, C. B. (2022). Statistical inference with non-probability survey samples. *Survey Methodology* **48**(2), 283-311.

ZBINDEN, N., KERY, M., HAFLIGER, G., SCHMID, H. & KELLER, V. (2014). A resampling-based method for effort correction in abundance trend analyses from opportunistic biological records. *Bird Study* **61**(4), 506-517.

ZHANG, W. Y., SHELDON, B., GRENYER, R. & GASTON, K. J. (2021). Habitat change and biased sampling influence estimation of diversity trends. *Current Biology* **31**(16), 3656-+.

ZIMNEY, A. & SMART, T. (2022). Effects of incomplete sampling and standardization on indices of abundance from a fishery- independent trawl survey off the Atlantic coast of the southeastern United States. *Fishery Bulletin* **120**(3-4), 252-267.

Supporting Information

Table S1 Selected R tools that can help with missing data problems and their potential application for use in biodiversity research.

R packages	Applications	Useful functions
Exploring missing data		
naniar	visualizing/exploring the missing data pattern	*mcar_test* – Little's missing completely at random (MCAR) test *vis_miss* – plot the missing data for all variables
occAssess	measure of the potential for bias in taxonomic, temporal, spatial, and environmental dimensions	*assessEnvBias* - assess whether data are sampled from a representative portion of environmental space in the spatial domain of interest *assessSpatialBias* – assess whether data resemble a random distribution in the geographic space of interest for inference *assessSpatialCov* – assess whether a representative portion of the spatial domain of interest has been sampled and whether the same portion of geographic space has been sampled over time
sampbias	a Bayesian approach to estimate how sampling rates vary as a function of proximity to one or multiple bias factors	*calculate_bias* - calculating the bias effect of sampling bias due to geographic structures, such as the vicinity to cities, airports, rivers and roads
Subsampling		
base	Base R functions	*sample* - sample data with predefined inclusion probabilities specified with the prob argument
sampling	draw random samples using different sampling schemes	*balancedcluster* – selects a balanced cluster sample according to defined auxiliary variables *strata* - stratified sampling with unequal probabilities.

spatialEco	spatial data manipulation and modelling	*stratified.random* - creates a stratified random sample of an sp class object *stratified.distance* - draws a minimum, and optional maximum constrained, distance sub-sampling
spThin	Spatial thinning of species occurence records	*thin* - returns a dataset with the maximum number of records for a given thinning distance
terra	spatial data manipulation and processing	*spatSample* – sample a SpatRaster, SpatVector or SpatExtent objcy
Imputation		
agTrend	modelling regional trends with missing data	*mcmc.aggregate* - a zero-inflated, nonparametric model with a definable observation model, augmenting missing values before calculating regional abundances
INLA/ inlabru	fitting Bayesian models, especially useful for spatial models via its spatial mesh	*inla/bru* - fit a Bayesian model using Integrated Nested Laplace approximation *predict* – draw predictions from the fitted model, where the prediction data frame can be a SpatialPointsDataFrame object
LORI	imputation of missing count data	*lori* – impute missing count data using a large covariate set, including interactions, with a LASSO penalty
mice	multiple imputation by chained equations	*mice* – multiple imputation method that will generate plausible values for any missing data – in the response and in any covariates
Rjags JAGS nimble	fitting Bayesian models allowing for missing values in the response	*Jags/runMCMC* - fitting Bayesian models allowing for imputation of missing values in the response during model fitting (options available for missing values in covariates too)
rtrim	functions to calculate annual indices and trends of abundances	fit a GLM imputing missing values based on mean site and year effects, with optional covariates
Weighting		

survey srvyr	range of functions for analysis of data from complex surveys, including fitting models with weights	*Svyglm* – generalized linear models with survey weights *postStratify* – function for post-stratification to match the joint distribution of the variables of the population	
svrep	Analysis of replicate/boostrapped survey weights	*svyby_repwts* – compare estimates from different sets of weights	
twang	functions to estimate propensity scores and weights	*ps* - gradient boosted trees to predict non-response from covariates *bal.table* – compare covariate values between sample and population	

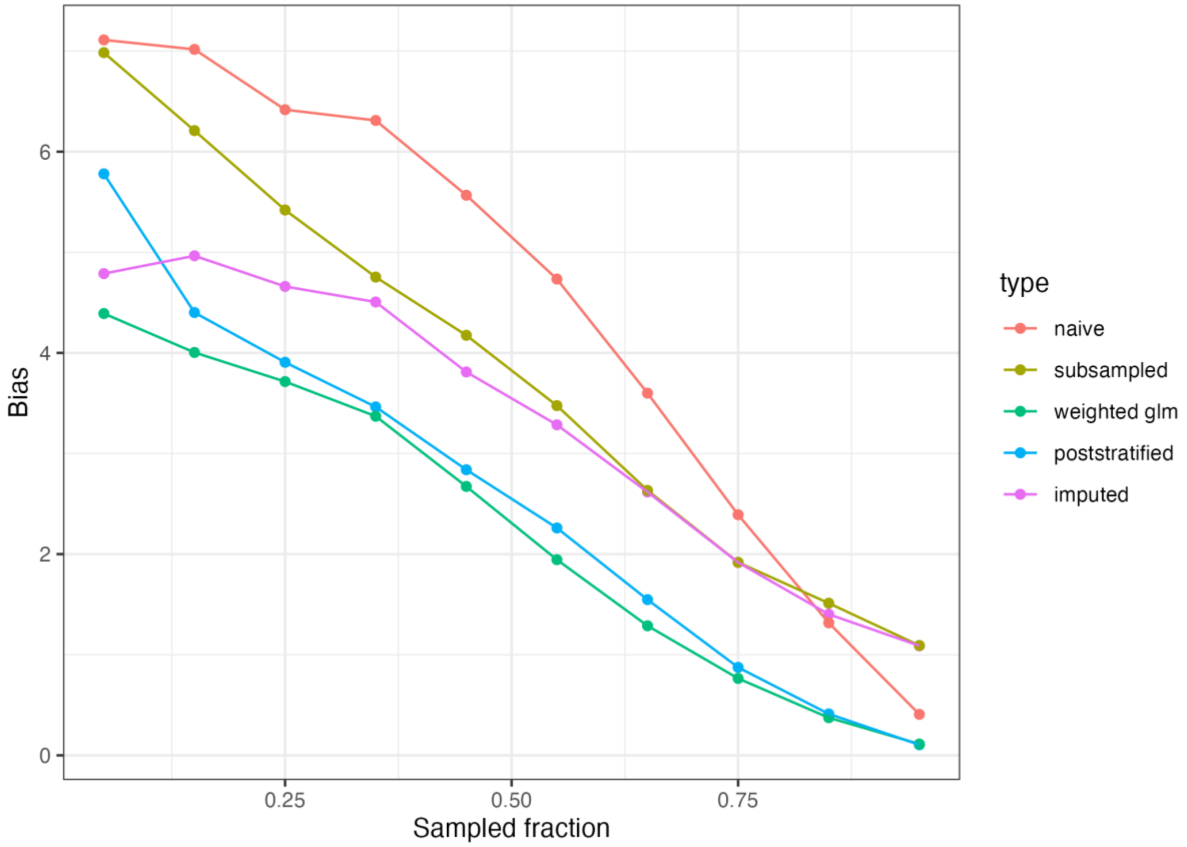

Fig. S1 The ability of missing data solutions to adjust for bias in biodiversity data.

We assumed a landscape of 400 cells and that a covariate affected both species abundance and the likelihood of a cell being sampled. We vary the fraction of the cells that were sampled. In contrast to Fig 5A (main text), we assumed that the species abundance was affected by an additional covariate that did not affect sampling; this variable was not included in any of the analysis. The models to estimate the parameter of interest (mean abundance) were: naive (no correction); subsampled (cells were subsampled along the covariate gradient to reduce the sampling bias), weighted (two methods: weighted glm using the svyglm function, and weighted by poststratification, using postStratify, both in the survey package) and imputed (using JAGS to impute NAs in the response). Points show the mean bias (difference between model prediction and truth) across 100 independent runs.