BIOLOGICAL REVIEWS

Cambridge Philosophical Society

Biol. Rev. (2024), pp. 000–000.
doi: 10.1111/brv.13127

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] and Michael J. O. Pocock[1]

[1] *UK Centre for Ecology & Hydrology, Maclean Building, Benson Lane, Wallingford, OX10 8BB, UK*
[2] *Department of Wildlife Ecology and Conservation, Fort Lauderdale Research and Education Center, University of Florida, 3205 College Avenue, Davie, Florida 33314-7719, USA*
[3] *British Trust for Ornithology, The Nunnery, Thetford, Norfolk, IP24 2PU, UK*

ABSTRACT

Big biodiversity data sets have great potential for monitoring and research because of their large taxonomic, geographic and temporal scope. Such data sets have become especially important for assessing temporal changes in species' populations and distributions. Gaps in the available data, especially spatial and temporal gaps, often mean that the data are not representative of the target population. This hinders drawing large-scale inferences, such as about species' trends, and may lead to misplaced conservation action. Here, we conceptualise gaps in biodiversity monitoring data as a missing data problem, which provides a unifying framework for the challenges and potential solutions across different types of biodiversity data sets. We characterise the typical types of data gaps as different classes of missing data and then use missing data theory to explore the implications for questions about species' trends and factors affecting occurrences/abundances. 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 species. But a data set *per se* is not biased. The outcome depends on the ecological question and statistical approach, which determine choices around which sources of variation are taken into account. We argue that typical approaches to long-term species trend modelling using monitoring data are especially susceptible to data gaps since such models do not tend to account for the factors driving missingness. To identify general solutions to this problem, 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 techniques 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 this review to outline the necessary considerations when dealing with data gaps at different stages of the data collection and analysis workflow.

Key words: biodiversity change, citizen science, ecological modelling, macroecology, spatial bias.

CONTENTS

[*] Author for correspondence (E-mail: diana.e.bowler@gmail.com).

Biological Reviews (2024) 000–000 © 2024 The Author(s). *Biological Reviews* published by John Wiley & Sons Ltd on behalf of Cambridge Philosophical Society.
This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

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 (GBIF)] (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 contain gaps (Boakes *et al.*, 2010). Data gaps are not necessarily a problem; indeed, most ecological 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). Many ecologists have raised concerns about the impacts of bias due to data gaps on estimated spatial or temporal biodiversity patterns (Bayraktarov *et al.*, 2019; Valdez *et al.*, 2023). For instance, biases could mean that species' trends are over- or under-estimated, leading to ill-informed decisions about which species should be conservation priorities and misplaced direction of conservation action. 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 are affected by the original motivations for, and constraints on, data-collection, reporting and mobilisation activities. There are, however, typical patterns in data availability that indicate common causes of data gaps. Spatial patterns in the data available from citizen science, which form the majority of monitoring data (Chandler *et al.*, 2017), have been especially well studied. Citizen-science programs have varying degrees of standardisation in 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, leading to data gaps in remote 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; for instance, towards regions undergoing less habitat 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 data sets. Missing data are a widespread problem crossing disciplines, with a large body of literature on their implications and possible solutions (Little & Rubin, 2019; Carpenter & Kenward, 2012; Carpenter & Smuk, 2021). We expect that aligning the generalised problem of missing data, conceptualised within missing data theory, to the problem of biodiversity data gaps 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. The general problems and potential solutions could be applied to animal or plant data, or terrestrial or marine systems; however, the implications of ignoring data gaps, and the ability to account for problematic data gaps, will vary according to the causal factors at play in both sampling probability and biodiversity patterns and the availability of data to model them. We show that bias is not a property of a data set; rather, bias is a property of the use of a data set for

1469185x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/brv.13127 by Centre For Ecology & Hydrology, Wiley Online Library on [09/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

a specific question and target population that are imposed by the data analyst. We review some commonly used solutions for 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 otherwise have been sampled. Together, spatial and annual gaps determine the spatial and temporal coverage of a data set. 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 can be used to estimate detection probabilities. Biodiversity data sets can also have taxonomic gaps (Troudet *et al.*, 2017) – this is outside the scope of this review since we are primarily interested in the implications of data gaps for species-level questions asked by monitoring schemes. However, some of the approaches we discuss later for dealing with spatio-temporal gaps have been applied to account for taxonomic gaps (e.g. weighting in the Living Planet Index, McRae, Deinet & Freeman, 2017) and missing data thinking could be extended to these types of gaps.

Considering why these gaps arise can help understand their likely impact, for instance, on species long-term trend estimation. Data gaps can be found in all types of monitoring data including highly structured monitoring schemes with a standardised protocol, such as many national bird survey schemes, as well as unstructured/opportunistic monitoring data that are typically an aggregation of heterogeneous observations. 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. 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 can still occur in these schemes, for instance due to a 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 and monitoring schemes, there is no large-scale planned spatial sampling design. 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 spatial sampling design since they contain data that were independently collected as part of separate studies, without any coordination in their efforts.

Drivers of data gaps may differ across data sets because of differences in sampling objectives and constraints, but

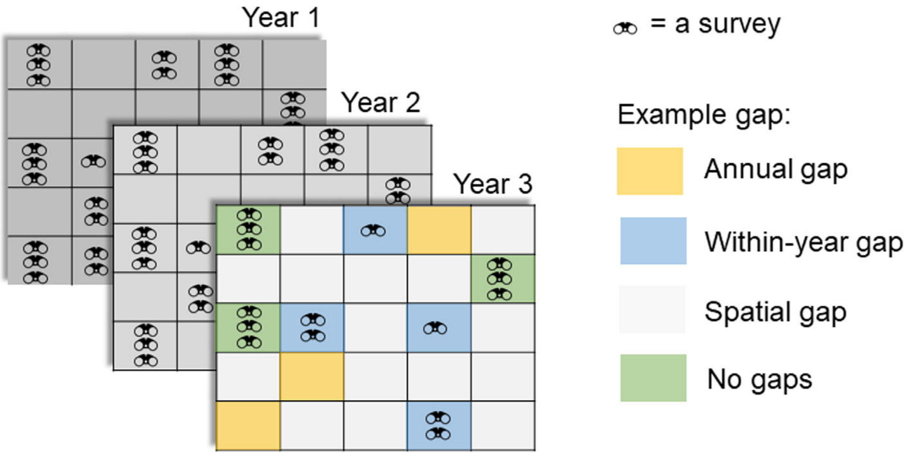


Fig. 1. Different types of data gaps within biodiversity data. We imagine a scenario where there are multiple survey visits across sites and years. Visits can be in response to a protocol ("structured" data) or opportunistic ("unstructured"), and repeat visits can be made by the same or different recorders. Data gaps, or more generally uneven data availability, can arise due to (*a*) within-year gaps (e.g. blue squares, 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 squares, i.e. some sites that are usually sampled are entirely unvisited in some years) or (*c*) spatial gaps (e.g. light grey squares, 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 squares).

1469185x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/brv.13127 by Centre For Ecology & Hydrology, Wiley Online Library on [09/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

similar gaps are often found within monitoring schemes 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 species and for surveying, e.g. agricultural land (Tulloch *et al.*, 2013; Dambly *et al.*, 2021; Marsh & Cosentino, 2019). At large scales, gaps can also be associated with socio-economic variables such as metrics of economic activity that might be associated with lower sampling and data-compilation effort (Meyer *et al.*, 2015). 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 (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, for example missing surveys for butterflies are more common at the 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 or missingness (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) 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 sampling, and causing missingness, are independent of those affecting species (Table 1). Under MCAR missingness, the observed data are effectively a random sample of the whole population, and the distribution of values of the biodiversity variable of interest are similar in sampled and non-sampled sites or times. For instance, if sampling 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, because many volunteers only have the necessary spare time to sample at the weekends (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 species occurrence or abundance (Table 1). In this case, missing data could reduce the precision of parameter estimates through reduced sample size but would not increase the bias.

When the factors affecting sampling effort are the same as, or correlated with, those factors affecting species, the missing data are not MCAR and can either be Missing at Random (MAR) or Missing Not at Random (MNAR). In both cases, there are systematic differences in the distribution of values of the biodiversity variable of interest between sampled and non-sampled sites or times (Table 1). For instance, if road density affects both sampling probability and species abundance, then any spatial gaps associated with roads are not MCAR. Similarly, habitat degradation at a site could reduce both species abundance and participant retention in a citizen science scheme, creating an annual data gap that is not MCAR (Table 1), because the missing values are lower than the sampled values.

Table 1. Missing data classes in biodiversity data, including examples and implications.

Missing data class	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: Weekdays Annual/spatial: 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 and included in the analysis	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 themselves	Sampling varies with biodiversity values or an unknown or unavailable covariate affects sampling and biodiversity	Within-year/annual/spatial: Unknown factors causing variation in species activity/abundance that correlate with sampling effort	Non-ignorable – the missing data mechanism might need to be modelled

1469185x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/brv.13127 by Centre For Ecology & Hydrology, Wiley Online Library on [09/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

We can separate MAR and MNAR by borrowing from the "Rumsfeld Matrix": MAR are effectively "known unknowns" while MNAR are "unknown unknowns". The "known" for MAR is knowledge and availability of data on the shared covariates affecting sampling probability and the biodiversity variable. If data for these covariates are available, and included in the analysis, then the missing data are MAR. Hence, despite its name, MAR in biodiversity monitoring 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 are available covariate data to explain fully the differences between sampled and non-sampled sites/times. If any of the relevant factors affecting sampling and species are unknown, unavailable or not modelled, the missing data become MNAR (Table 1). Hence, decisions of the analyst, and whether to collect and model the effects of a specific variable, 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 missing values of the biodiversity variable itself, that is, if sampling effort directly depends on species occurrence or abundance.

Statistical tests can only partly help assess which form of missingness 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 tested directly. Concerns about whether missingness in the biodiversity data is directly associated with the data's values could be explored if there is a related variable with available data (Wu, 2022). We argue that MCAR is unlikely in most biodiversity data because many variables that affect sampling probability (such as road density or human population density) are also likely to affect species. Even in schemes with a planned spatial design, a similar set of variables are likely to be associated with unplanned data gaps that arise from variation in participant recruitment or drop-out. However, MCAR is still a useful concept as a null hypothesis and because it is the assumption made when no consideration is given to adjust for data gaps when using monitoring data to estimate species trends. Since gaps in biodiversity data are caused by a range of different factors, some gaps may be understandable by knowledge of the data collection process and/or with available environmental covariates, while other gaps may be harder to explain. This means that data gaps are unlikely to be entirely MAR or MNAR, but typically a mixture.

III. IMPLICATIONS OF MISSINGNESS FOR ECOLOGICAL QUESTIONS

Missing biodiversity data do not necessarily have strong impacts on the results of statistical modelling – the outcome often depends on the specific question and parameter of interest (Bartlett, Harel & Carpenter, 2015; Collins, Schafer & Kam, 2001; Little *et al*., 2022; Hughes *et al*., 2019). 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 MCAR. The "missing at random" assumption of MAR is conditional on accounting for variables affecting sampling probability within an analysis, require that these variables are known, reflected in available data and included in the analysis (Fig. 2) (Conn, Thorson & Johnson, 2017; Hefley *et al*., 2013). Because different ecological questions will lead to different decisions about which variables to collect and include in an analysis, a data gap might be MAR under some questions/analyses but MNAR under others. To illustrate these potential differences, we contrast two typical questions asked with biodiversity data.

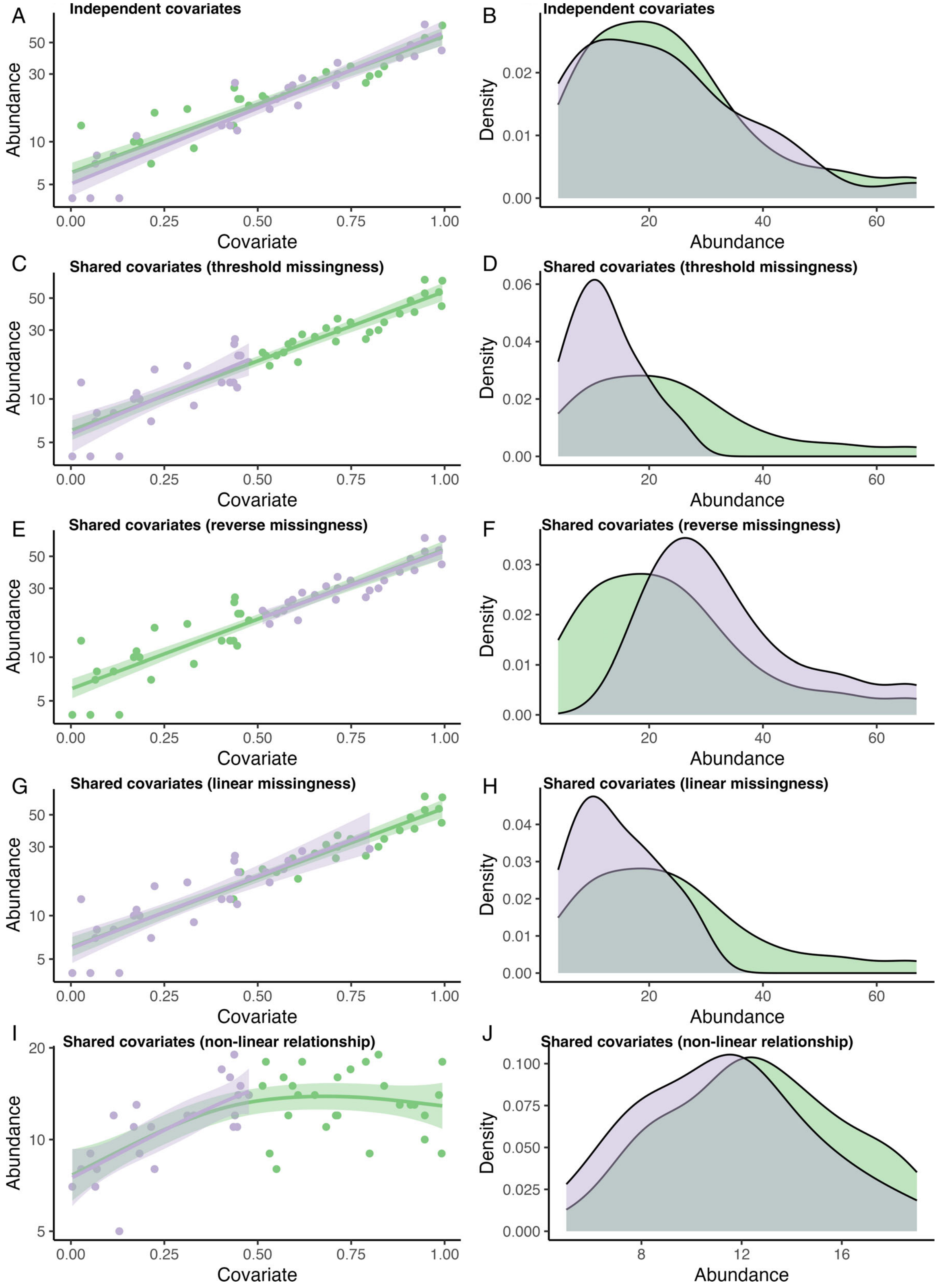
(1) Understanding the roles of environmental drivers on species' distributions

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 often considered in terms of niche truncation. Niche truncation happens when a data set only contains occurrence data from part of the geographic range of a species, which usually also means that the data set 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 the type of monitoring data available (Baker *et al*., 2022).

We begin considering the scenario when abundance data are available. In this case, if there is a simple linear relationship between an environmental covariate and species abundance, missing data do not necessarily cause bias in the estimated effect of the covariate on abundance, even when missingness depends on the same covariate (Fig. 2A, C, E, G) (Collins *et al*., 2001). For instance, we could estimate the effect of elevation on species abundance without bias, even if elevation is associated with data gaps (e.g. if we are missing data from high-elevation regions), provided elevation and abundance are linearly related. This is because the relationship between the covariate and species abundance can be estimated without bias using data over the sampled range of covariate values, as shown in Fig. 2C – the same relationship is found with a full data set (green in Fig. 2) or a restricted data set with data gaps (purple in Fig. 2). Missing data can, however, cause problems when the underlying relationship between the covariate and species abundance is non-linear. In this case, data gaps hinder estimating the true form of the relationship (see Fig. 2I – the true curved relationship is fitted with the full data set but a simple positive linear relationship is fitted with the restricted data set). The fitted relationship using the restricted data set

1469185x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/brv.13127 by Centre For Ecology & Hydrology, Wiley Online Library on [09/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

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



(Figure 2 legend continues on next page.)

1469185x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/brv.13127 by Centre For Ecology & Hydrology, Wiley Online Library on [09/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

1469185x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/brv.13127 by Centre For Ecology & Hydrology, Wiley Online Library on [09/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

et al., 2023), we expect this issue is likely to be widespread. We also note that we assumed a linear relationship on the log-scale in our example (Fig. 2), which matched the log link function of the fitted regression model, but non-linearity in other cases could also be affected by the specific link functions used in generalised linear models.

We now consider the alternative scenario of fitting a distribution model with presence-only occurrence data, typical of opportunsitic citizen science. In this case, any data gaps within a geographic region 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 sampling processes from the true ecological processes affecting species distributions (Hefley *et al.*, 2013; Baker *et al.*, 2022). Many methods have been developed to generate pseudo-absences for analysis of presence-only data (Barbet-Massin *et al.*, 2012; Hertzog, Besnard & Jay-Robert, 2014), but they are still usually more prone to biases when there are shared covariates affecting sampling probability and true species occurrence (Baker *et al.*, 2022). The target-group background method is a popular approach to generate pseudo-absences by integrating data from multiple species assumed to be surveyed by similar methods/people. With this method, the aim is to produce absence data with a similar pattern of spatial sampling bias as the presence data of the focal species (Phillips *et al.*, 2009), but its performance depends on the range of environmental preferences of the species included in the target group (Botella *et al.*, 2020). 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.

(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 function, spline function or as a factor (Amano *et al.*, 2012; Bled *et al.*, 2013). Drivers of the trend are not explicitly modelled when the goal is simply to estimate the mean long-term trend. As such, broader inferences about the estimated mean trend are based on the assumed representativeness of the sampled sites, or prior knowledge of sampling unit inclusion probabilities (see design weights discussed in Section IV.2). Relying on the representativeness of the sampling design is the most traditional approach to survey sampling (Smith, 1976) and the one typically taken by official governmental surveys using some form of random sampling (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 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, for example 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 in some monitoring schemes (Buckland & Johnston, 2017). We estimated the mean trend using a simple mixed-effect model including site and year. The results show that when an independent covariate affected sampling, the trends were unbiased, but when the site-level covariate affected both sampling and species' trends, the trends were biased (Fig. 3). In real-world situations, many factors will simultaneously influence the trend of a species, but this simple simulation highlights the potential for bias caused by shared covariates. Since the specific causal covariates driving species trends or sampling probability are not included in the commonly used descriptive trend models, trend analyses are liable to be affected by MNAR. Without conditioning on the covariates involved, trend estimates might be underestimated if missing data are more common in regions where species trends are more strongly

(Figure legend continued from previous page.)
Fig. 2. The impacts of different missing data patterns on regression (left) and sample distributions (right). We use a hypothetical data set 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 data set with (purple) and without (green) missing data [i.e. missingness is "missing completely at random" (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 (threshold missingness). The estimated effect of the covariate is the same in the data set 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 data set 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 (linear missingness rather than a theshold). 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.

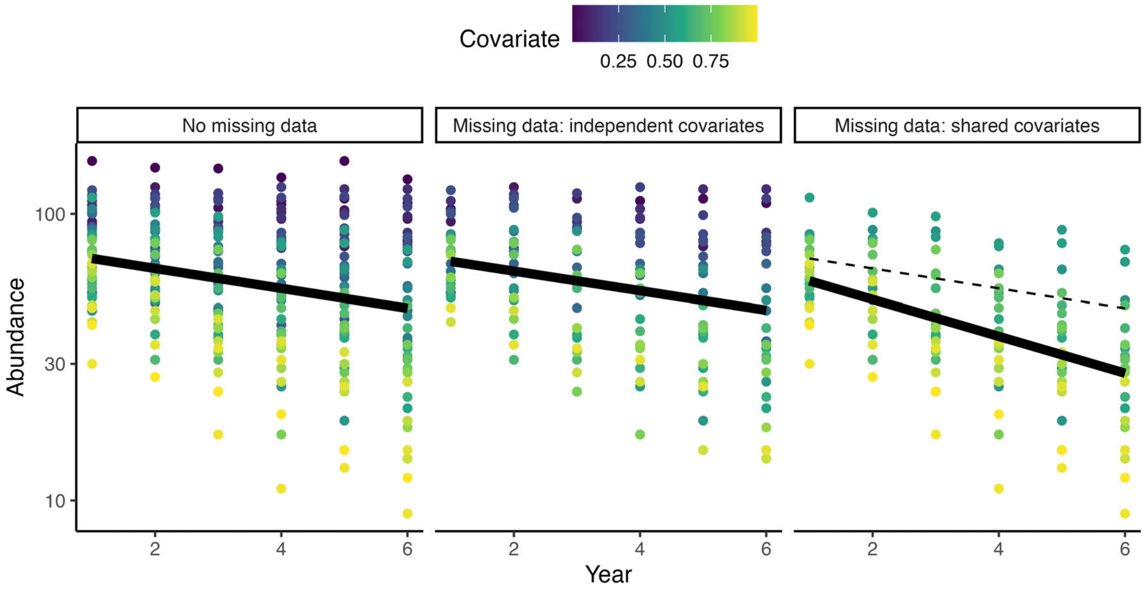


Fig. 3. The impacts of different missing data mechanisms on trend modelling. We use a hypothetical scenario in which a mean trend model is fitted to data sets that vary in their missing data mechanism. We assumed a scenario of 50 sites that varied in an environmental covariate affecting species trends (trends were stable or even increasing at low values of the covariate and declining at increasingly high values of the covariate). When the probability of sampling a site was independent of the covariate driving species trends [i.e. a “missing completely at random” (MCAR) pattern – there are fewer points in each year in the middle panel, but they are a random set of those in the left panel, that is the covariate affecting sampling probability was a different and uncorrelated covariate to the one affecting species], the overall mean trend (estimated by the year effect in a generalised 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 “missing not at random” (MNAR) pattern], the overall mean trend was downward biased with missing data (right panel) compared to the scenario of no missing data (shown by the dashed black line and in the left panel).

declining or overestimated if missing data are more common in regions where species are stable or increasing (Fig. 3) (Bowler *et al*., 2022; Buckland & Johnston, 2017).

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 predictors typically used in large-scale modelling tend to have no or few gaps (e.g. site identity or environmental data products derived 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, when data are available on the key covariates affecting sampling to be included in the analysis. MNAR is the most challenging class of missing data to deal with in statistical modelling, so we deal with MNAR separately in Section VI.

(1) Subsampling

The “Big Data Paradox” highlights that there can be trade-offs between data set size and data set quality (Bradley *et al*., 2021; Meng, 2018). Small data sets can be preferable to large data sets if they are more representative and less heterogeneous (Bayraktarov *et al*., 2019). Based on such thinking, some studies have proposed “reverse engineering” structure in biodiversity data by filtering data points (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*., 2021). Typically, subsampling is done using geographic covariates or spatial units, such as grid cells, rather than using environmental covariates that are assumed to have a causal link with either sampling or species. Some have also applied this approach to reduce the effects of temporal

1469185x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/brv.13127 by Centre For Ecology & Hydrology, Wiley Online Library on [09/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

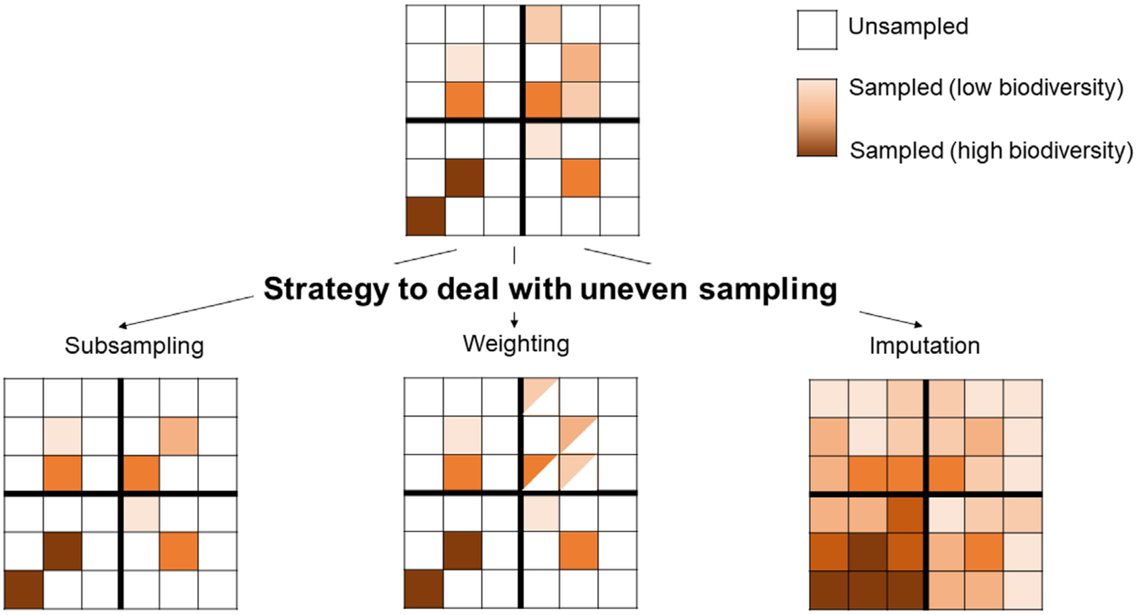

Unsampled
Sampled (low biodiversity)
Sampled (high biodiversity)
Strategy to deal with uneven sampling
Subsampling
Weighting
Imputation

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). In the top panel, the landscape is divided into four quarters (e.g. representing different habitats or geographic regions). One quarter (top right quarter) has been sampled more (four sampling sites) than the others (two sampling sites each). The bottom panel shows possible solutions. In random subsampling (bottom left), two sites are randomly subsampled from the oversampled quarter to create a data set with an even sampling coverage across quarters. In weighting (bottom middle), data from the oversampled quarter are down-weighted in the statistical model so data from all quarters similarly influence the modelled results. In 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 weighting, the aim is to improve the representativeness of the sample for statistical inference at the population level. In imputation, the aim is directly to 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; Schmucki *et al.*, 2016)
Annual	Sometimes imputed e.g. generalised linear models to impute gaps based on mean site and year effects, optionally allowing for habitat differences, 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).

changes in sampling effort (Hof & Bright, 2016; Zbinden *et al.*, 2014), although not always successfully (Callcutt, Croft & Smith, 2018). Subsampling could 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 detections 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; Valliant, Dever & Kreuter, 2018). 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 are not MCAR. For instance, weighting has been used to reduce selection bias caused by participant non-response in surveys (Seaman & White, 2013), but it is less often used to account for data gaps in biodiversity modelling (Boyd, Powney & Pescott, 2023; Aubry & Francesiaz, 2022).

Different types of weights have been used in the analysis of biodiversity data, especially to deal with spatial gaps: (1) design weights; (2) non-response weights (or sampling weights) and (3) population weights. Each form of weighting is intended to improve sample representativeness of some target population but vary in terms of whether or not the

1469185x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/brv.13127 by Centre For Ecology & Hydrology, Wiley Online Library on [09/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

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 monitoring 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 can be 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). In both cases, the non-response weights must be estimated based on the available data and hence differ from design weights since they cannot be known with certainty. Population weights are primarily used in the calculation of supranational/international indicators in which estimates from national surveys are combined (e.g. farmland or woodland bird indicators; Gregory *et al.*, 2005). For these indicators, populations weights are used to give greater weight to data from regions/countries that harbour a larger proportion of the species' total population, when calculating the overall mean.

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 a 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, post-stratification (for categorical covariates or subgroups termed strata), or more generalised 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. For instance, when estimating the occupancy change of a plant species in the UK, Boyd, Stewart & Pescott (2024) used data on elevation – a factor affecting both sampling and species occupancy – to upweight data from undersampled high-elevation regions and produce more accurate estimates of the species distribution size at different time points. 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, for example by coarsening the covariates used to define the weights so that all strata have some probably of being sampled, or by truncating weight values so that extreme weights are not produced (Battaglia, Hoaglin & Frankel, 2009). Another approach that can help deal with small sample size in some strata is so-called "Mr P" analysis (= Multilevel Regression with Post-stratification). With this approach, variables for the sampling strata are included as random effects in a multilevel regression/mixed-effect model, so that there is partial pooling of information across strata, before the model predictions for each strata are reweighted for representativeness of the target population (Gelman, 2007; Authier, Rouby & Macleod, 2021).

The most appropriate approach to using weighting is likely to be question and taxon specific, varying with how much the species range extends across the region of interest. For example, when estimating trends in the total population size of a species, it might not be important to upweight undersampled regions if those regions overlap with where a species is rare, or even absent. If, however, the goal is to estimate trends in average site-level population trends of a species, then it would be important to up-weight data from undersampled 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 data set 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), so imputation here can be equated with making model predictions at unsampled sites and times.

Imputation is already in use in species 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, designed for waterbirds (Underhill & Prysjones, 1994; 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, for example to fill annual gaps using TRIM/birdSTATs, commonly used for bird indices (Lehikoinen *et al.*, 2016); or using temporal splines, for example to fill seasonal gaps in butterfly sampling (Schmucki *et al.*, 2016; Dennis *et al.*, 2016) or using ecological covariates (Dakki *et al.*, 2021). A Bayesian framework can be especially useful for dealing with missing values in the response since they are naturally imputed with a full probability distribution during model fitting, for example with Just Another Gibbs Sampler (JAGS) or NIMBLE. For instance, Bayesian occupancy-detection models have been used to analyse opportunistic species observations from citizen science, with annual data gaps at each site imputed before the predicted annual proportion of occupied sites is calculated (Outhwaite *et al.*, 2019). The flexibility of Bayesian models means they

1469185x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/brv.13127 by Centre For Ecology & Hydrology, Wiley Online Library on [09/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

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 practice to deal with sample unrepresentativeness in other disciplines, especially social sciences	– less commonly applied in ecology – diverse range of possible weighting techniques (Valliant, 2020; Boyd *et al.*, 2024) but little guidance available for ecologists to decide which approach to use
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).	– requires a good understanding of the ecological system to predict the missing biodiversity values – inefficient when the number of unsampled sites/times is large if the goal is only to estimate mean abundance or occupancy

could also incorporate expert knowledge as priors 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 used less often to deal with spatial gaps when the focus is modelling change over time in species' abundances or occurrences. An exception is studies of changes in species' range sizes using distribution models that predict the full distribution of a species at multiple time points before change is assessed (e.g. Grattarola, Bowler & Keil, 2023). Monitoring schemes with large spatial coverage (e.g., eBird) are also beginning to use models to predict spatio-temporal patterns of abundance change across whole countries (Fink *et al.*, 2020). In these cases, statistical models of the effects of environmental covariates and/or spatial structure are used to make predictions at unsampled sites (Bush *et al.*, 2017; Ver Hoef *et al.*, 2021; Breivik *et al.*, 2021). Geostatistical methods, such as kriging, also offer a range of interpolation methods for spatial data, which are especially useful when there is a strong spatial autocorrelation (Ballesteros-Mejia *et al.*, 2013; Kreft & Jetz, 2007; Lin *et al.*, 2008).

V. PRO AND CONS OF EACH SOLUTION

All of the above-mentioned approaches have the potential to reduce the bias in parameter estimates associated with data gaps 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 included data points 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 increased parameter uncertainty: the classic bias–variance trade-off (Hefley *et al.*, 2013). This is because subsampling directly reduces the sample size; weighting can reduce the effective sample size; and imputation adds uncertainties *via* predictions at unsampled points. But this trade-off does not always apply; for instance, post-stratification 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 used to adjust for missing data. The general recommendation from the missing data theory and survey sampling literature is to be generous when deciding on auxiliary variables, considering those relating to the missingness (i.e. sampling effort in the context of biodiversity data gaps) to reduce bias, and to the biodiversity outcome to reduce variance (Collins *et al.*, 2001; Caughey *et al.*, 2020). It is worth noting, however, that selecting auxiliary variables purely on the hypothesised strength of the correlation can increase bias in some circumstances (Thoemmes & Rose, 2014), and a safer strategy is to select covariates *a priori* based on causal reasoning (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.*, 2024), suggesting that the limiting factor in accouting for bias often may be defining the right auxillary variables.

We illustrate some of these challenges and the application of each potential solution with a toy example of an abundance data set with missing values (Fig. 5). We simulated a

1469185x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/brv.13127 by Centre For Ecology & Hydrology, Wiley Online Library on [09/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

landscape in which a covariate (for example 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/the 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 for each value of the covariate. For weighting, we compared two approaches: first, fitting a generalised linear regression model with cell-specific weights (inverse of sampling probability) using model-robust variance estimators that take into account the weighting of the observations, and second, using post-stratification to weight data so the covariate distribution of the sampled data matched that of the population – in this case the unique covariate values were treated as separate sampling strata (Valliant *et al*., 2018). For imputation, we fitted a Bayesian generalised linear regression model (using JAGS) in which missing values were set as "NA" in the response and were imputed based on the estimated effect of the covariate.

The results show that all methods do better, in terms of reducing bias, than a naive approach that did not attempt to account for missingness in the estimation of the mean abundance (Fig. 5A, C). Subsampling performed the worst, and weighting the best. Post-stratification tended to perform slightly less well (i.e. led to higher bias) when the sampling fraction was low, when the number of missing values was high (Fig. 5A). This latter pattern was because the sample did not contain all the habitat quality values found in the population, meaning there were no available data to upweight in under-sampled regions. All models performed less well, shown by higher bias, as the available covariate became a weaker proxy of the true driving covariate

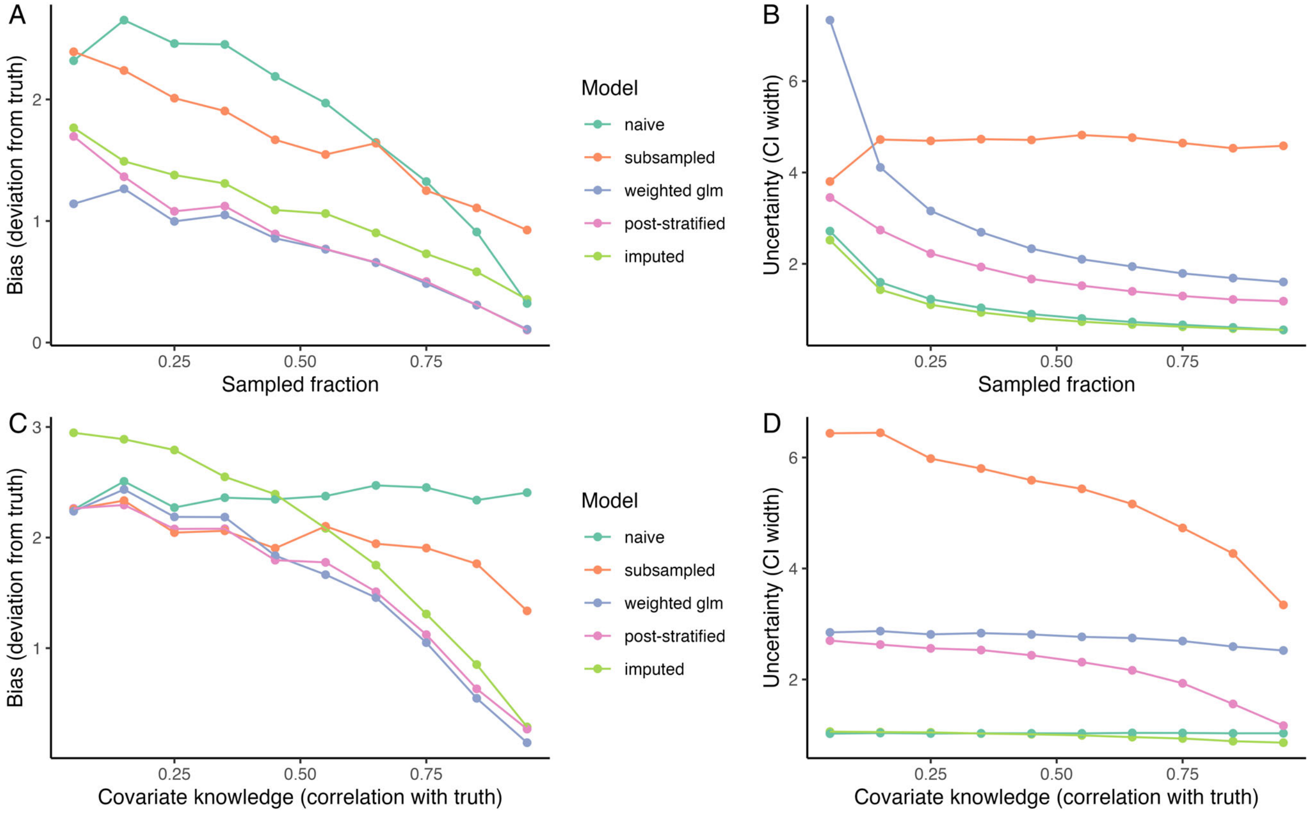


Fig. 5. The ability of missing data solutions to adjust for bias in biodiversity data. We assumed a landscape of 400 cells and that a cell-varying covariate affected both species abundance and the likelihood of a cell being sampled. In A and B, we varied the fraction of the cells that were sampled. In C and D, we varied the correlation between the true covariate and the covariate available for analysis, as measure of our knowledge (correlation of 1 = perfect covariate and knowledge). The models used to estimate the parameter of interest (mean abundance) were: naive [no correction, Poisson generalised linear model (GLM)]; subsampled (cells were subsampled along the covariate gradient), weighted (two methods: weighted GLM and post-stratification, using postStratify in the *survey* package) and imputed (using Just Another Gibbs Sampler, or JAGS, in which missing values were set as "NA" in the response). Points in A and C show the mean bias (difference between model prediction and truth – note the true mean value was 7.3) while B and D show the mean width of the confidence intervals (CI) of the mean abundance estimate across 100 independent runs. In A and B, covariate knowledge was fixed at a correlation of 0.75; while in C and D, the sampling proportion was fixed at 0.35.

1469185x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/brv.13127 by Centre For Ecology & Hydrology, Wiley Online Library on [09/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

(lower correlation with the truth; Fig. 5C), especially imputation. In terms of uncertainty of the parameter estimates (measured as the width of the confidence interval), attempts to adjust for bias usually increased the width of the confidence intervals (Fig. 5B, D) – subsampling led to the greatest increase in uncertainty (explained by reduced sample size), while weighting added intermediate levels of uncertainty. For post-stratification, the increase in uncertainty was minimal when the covariate was a good proxy (Fig. 5D), which is a pattern noted elsewhere (Little & Vartivarian, 2005). In this simple example, imputation led to a similar uncertainty – in terms of width of the confidence interval – as the naïve approach, partly because we assumed a relatively small and well-sampled system. In further simulations, we found that imputation performed less well when there were additional covariates affecting species abundance and these covariates were not modelled, highlighting the importance of understanding the ecological system for imputation (see online Supporting Information, 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 affect any attempts to account for missing data. We point the reader towards some useful R packages and functions in Table S1.

VI. DEALING WITH MISSING NOT AT RANDOM

Dealing with MNAR is more challenging than dealing with the other classes of missing data (Little & Rubin, 2019). In this case, missingness is directly associated with unavailable data, which could be either because sampling is affected by the missing biodiversity values or important covariates that are not known to be important and/or are not measured or measurable. This makes MNAR especially difficult to diagnose [but see Conn *et al.* (2017) for suggestions] and model, since possible auxiliary variables to adjust for the data gaps are not available. MNAR can arise through several 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) – leading to more missing values in places where the species is rare or absent. Preferential sampling can arise, for instance, if observers visit a location specifically to 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 chosen 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 found (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 (Little, 1993). 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 used to explore the consequences of plausible missing data mechanisms as sensitivity analyses (Little, 1995). In the ecological literature, approaches to deal with preferential sampling have also involved jointly modelling the sampling intensity, the biodiversity value at sampled points and the dependence between them, such as using marked point process models (Conn *et al.*, 2017; Pennino *et al.*, 2019; Laxton *et al.*, 2023). Meta-analyses often face similar MNAR problems, caused by publication bias when data are missing according to values of the data itself. In meta-analysis, similar sensitivity analyses, including selection models and the trim-and-fill method, have been proposed to test the robustness of model predictions to possible assumptions about missing data (Maier, VanderWeele & Mathur, 2022; Sutton *et al.*, 2000). Another approach to inference in a MNAR 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); however, the challenge 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 because at least some of the known factors affecting sampling probability, especially accessibility, urban land cover and human population density, overlap with those affecting species. This 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

1469185x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/brv.13127 by Centre For Ecology & Hydrology, Wiley Online Library on [09/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

opportunities rather than challenges since solutions are available to deal with missing data, provided that sampling inclusion probabilities are known. Indeed, planned data gaps are already used in schemes with a spatially stratified sampling design, often in relation to sampling probabilities of different geographic regions. In other fields, beyond monitoring, intentionally missing data has been proposed for ethical or practical reasons (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, with the long-term aim of increasing spatial coverage of the data (Darvill *et al.*, 2020). Alternative study designs, such as wave missingness or a rotating panel design (Nielsen *et al.*, 2009; Little & Rhemtulla, 2013) explicitly incorporate planned data gaps (e.g. years when a site is not planned to be surveyed) and may similarly 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 regions that represent sampling priorities – either because they lack data or because they are dissimilar to sampled regions. 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.*, 2019, 2023). 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). A similar targeting of effort may be used in synthesis studies that compile data from independent studies. In this case, efforts to mobilise data may be targeted towards under-represented sites/times, or those with the most uncertain predictions according to models of the existing data.

(2) Evaluating and reporting missingness

Developing a causal model of the factors affecting sampling probability and species [e.g. using a directed acyclic graph (DAG) to visualise the hypothesised causal links] can be a useful first step to identify the covariates linked to both sampling probability and species occurrence or abundance (Mohan & Pearl, 2021; Hughes *et al.*, 2019). As far as possible, data should then be collected on these covariates. Statistical models can be used to test whether covariates that are associated with missingness are also associated with the species, although of course the latter is only possible in the sampled data. Unplanned missingness in structured monitoring schemes could be investigated by disseminating follow-up surveys to participants to determine their reasons for missed surveys. Follow-on data collection, for example with paid surveys, or targeted citizen science, in regions or times of missing data could also help understand whether there are fundamental differences in species occurrence/abundance between the original data set and the extended data set.

Missingness, and how it is dealt with, is often not clearly reported in species trend analyses. Some reporting frameworks for missing data have been developed for other disciplines (Lee *et al.*, 2021). Such frameworks are in their early stages in ecology, but an approach has been proposed recently (Boyd *et al.*, 2022) that builds on the "risk of bias" tools used in other fields, especially in systematic reviews in medicine (Babic *et al.*, 2019). 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). But also important are summaries or visualisations of the distributions of environmental covariate values in sampled and non-sampled times/sites to highlight potentially important differences between the sample and target population of inference.

(3) Modelling to account for data gaps

The impact of data gaps will depend on multiple factors: their frequency and contiguity; how well data gaps are understood; whether the factors affecting missingness are independent of the factors affecting species and species abundance itself; the ecological questions being asked and which covariates are available and included in the analysis. Because of this, the potential impacts of missingness and possible solutions should be considered for each species–question–data set combination. A data set *per se* is not biased. Subsampling, weighting and imputation all have potential to reduce bias caused by data gaps. 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 overall means or mean trends in abundance or occupancies. Imputation methods offer the potential to fill in spatio-temporal gaps to generate the space–time data cubes underlying the Essential Biodiversity Framework (Kissling *et al.*, 2018), but their success depends on the ability to model variation in the biodiversity response. If bias is expected to be strong, but the causes are not fully known or relevant covariate data not fully available to adjust for it, the broader implications that can be drawn from a model of the data become difficult to communicate. Sensitivity analysis could help explore how different assumptions of the missingness would affect model interpretation and the robustness of conclusions (Little, 1995; Leurent *et al.*, 2018). Alternatively, it might be sensible to redefine the target region of interest to a region with fewer data gaps so that the sampled data are more representative of the target population. If this is not possible due to wide data gaps, a final option might be to revise the generality of the study question to make explicit the limits of information within the sampled data.

1469185x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/brv.13127 by Centre For Ecology & Hydrology, Wiley Online Library on [09/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

VIII. CONCLUSIONS

(1) Biodiversity data sets containing information on species' occurrences and abundances are rapidly growing in size, but data gaps are not necessarily closing. Nonetheless, big biodiversity data sets 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 volunteer citizen scientists or contracted surveyors – 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 about biodiversity patterns.

(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 species: shared covariates affecting sampling effort and species occurrence or abundance 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 time, statistical solutions can only go so far, closing data gaps with more coordinated data collection across stakeholders in biodiversity and environmental monitoring is also important to advance predictions of the state of, and trends in, biodiversity.

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. R.J.B. 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 the topic of this review.

X. DATA AVAILABILITY STATEMENT

R script for the example solution simulations (Figs 5 and S1) can be found at: https://github.com/bowlerbear/dataGaps.

XI. 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., *et al.* (2023). Probabilistic and preferential sampling approaches offer integrated perspectives of Italian forest diversity. *Journal of Vegetation Science* **34**(1), e13175.

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**, 109394.

Authier, M., Rouby, E. & Macleod, K. (2021). Estimating cetacean bycatch from non-representative samples (I): a simulation study with regularized multilevel regression and post-stratification. *Frontiers in Marine Science* **8**, 719956.

Babic, A., Tokalic, R., Cunha, J. A. S., Novak, I., Suto, J., Vidak, M., Miosic, I., Vuka, I., Pericic, T. P. & Puljak, L. (2019). Assessments of attrition bias in Cochrane systematic reviews are highly inconsistent and thus hindering trial comparability. *BMC Medical Research Methodology* **19**(1), 76.

Bailey, M. A. (2023). A new paradigm for polling. *Harvard Data Science Review* **5**(3). https://doi.org/10.1162/99608f92.9898eede.

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), 1–10. https://doi.org/10.29115/SP-2009-0019.

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**, 239.

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), e81867.

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), e1000385.

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.

Botella, C., Joly, A., Monestiez, P., Bonnet, P. & Munoz, F. (2020). Bias in presence-only niche models related to sampling effort and species niches: lessons for background point selection. *PLoS One* **15**(5), e0232078.

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), e06219.

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.

1469185x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/brv.13127 by Centre For Ecology & Hydrology, Wiley Online Library on [09/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

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

Boyd, R. J., Stewart, G. B. & Pescott, O. L. (2024). Descriptive inference using large, unrepresentative nonprobability samples: an introduction for ecologists. *Ecology* **105**(2), e4214.

Bradley, V. C., Kuriwaki, S., Isakov, M., Sejdinovic, D., Meng, X. L. & Flaxman, S. (2021). Unrepresentative big surveys significantly overestimated US vaccine uptake. *Nature* **600**(7890), 695–700.

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., *et al.* (2017). Connecting earth observation to high-throughput biodiversity data. *Nature Ecology & Evolution* **1**, 0176.

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), 20191487.

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), 28.

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, Chichester.

Carpenter, J. R. & Smuk, M. (2021). Missing data: a statistical framework for practice. *Biometrical Journal* **63**(5), 915–947.

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, Cambridge.

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**, 944116.

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., *et al.* (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* **16**(5), e12973.

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). https://doi.org/10.1002/eap.2056.

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), e12951.

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* **50**(9), 1561–1575.

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), e3982.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica* **1**, 153–161.

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), e01338.

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.

1469185x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/brv.13127 by Centre For Ecology & Hydrology, Wiley Online Library on [09/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

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**, 669477.

Hughes, R. A., Heron, J., Sterne, J. A. C. & Tilling, K. (2019). Accounting for missing data in statistical analyses: multiple imputation is not always the answer. *International Journal of Epidemiology* **48**(4), 1294–1304.

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., 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. *Nature Communications* **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**, 108927.

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., et al. (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. *Peer J* **9**, e10713.

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. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association* **88**(421), 125–134.

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*. https://doi.org/10.1177/00491241221113873.

Little, R. J. & Rubin, D. B. (2019). *Statistical Analysis with Missing Data*, Third Edition Edition. Wiley, Hoboken, New Jersey.

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**, 109964.

Maier, M., VanderWeele, T. J. & Mathur, M. B. (2022). Using selection models to assess sensitivity to publication bias: a tutorial and call for more routine use. *Campbell Systematic Reviews* **18**(3), e1256.

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), e12185.

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.

McRae, L., Deinet, S. & Freeman, R. (2017). The diversity-weighted living planet index: controlling for taxonomic bias in a global biodiversity indicator. *PLoS One* **12**(1), e0169156.

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.

Meyer, C., Kreft, H., Guralnick, R. & Jetz, W. (2015). Global priorities for an effective information basis of biodiversity distributions. *Nature Communications* **6**, 8221.

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., et al. (2019). Annual estimates of occupancy for bryophytes, lichens and invertebrates in the UK, 1970–2015. *Scientific Data* **6**, 259.

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.

Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J. & Ferrier, S. (2009). Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications* **19**(1), 181–197.

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), e0172579.

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.,

1469185x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/brv.13127 by Centre For Ecology & Hydrology, Wiley Online Library on [09/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

et al. (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: a review. *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.

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.

Sutton, A. J., Song, F., Gilbody, S. M. & Abrams, K. R. (2000). Modelling publication bias in meta-analysis: a review. *Statistical Methods in Medical Research* **9**(5), 421–445.

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? In *Bird Numbers 1992. Distribution, Monitoring and Ecological Aspects* (eds E. J. M. Hagemeijer and T. J. Verstrael), pp. 663–673. Proceedings of the 12th International Conference of IBCC and EOAC, Noordwijkerhout, The Netherlands. Statistics Netherlands, Voorburg/Heerlen & SOVON, Beek-Ubbergen, The Hague.

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**, 9132.

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. I. 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), e06604.

Valliant, R. (2020). Comparing alternatives for estimation from nonprobability samples. *Journal of Survey Statistics and Methodology* **8**(2), 231–263.

Valliant, R., Dever, J. A. & Kreuter, F. (2018). *Practical Tools for Designing and Weighing Survey Samples*. Springer, Cham.

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), 3656–3662.

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.

XII. SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Fig. S1. The ability of missing data solutions to adjust for bias in biodiversity data: extended analysis with additional covariates affecting the biodiversity response.

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

(*Received 20 October 2023; revised 23 July 2024; accepted 25 July 2024*)

1469185x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/brv.13127 by Centre For Ecology & Hydrology, Wiley Online Library on [09/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License