Trends in
Ecology & Evolution

CellPress
OPEN ACCESS

Opinion

We need to talk about nonprobability samples

Robin J. Boyd [1,*,@], Gary D. Powney, [1,@] and Oliver L. Pescott [1,@]

In most circumstances, probability sampling is the only way to ensure unbiased inference about population quantities where a complete census is not possible. As we enter the era of 'big data', however, *non*probability samples, whose sampling mechanisms are unknown, are undergoing a renaissance. We explain why the use of nonprobability samples can lead to spurious conclusions, and why seemingly large nonprobability samples can be (effectively) very small. We also review some recent controversies surrounding the use of nonprobability samples in biodiversity monitoring. These points notwithstanding, we argue that nonprobability samples can be useful, provided that their limitations are assessed, mitigated where possible and clearly communicated. Ecologists can learn much from other disciplines on each of these fronts.

Highlights

As the data revolution gathers pace, researchers are increasingly relying on nonprobability samples from meta-databases, citizen science and other sources to monitor biodiversity.

The use of nonprobability samples can lead to biased inference, and seemingly large nonprobability samples can actually have very low information content.

A number of recent high-profile disagreements in the biodiversity literature stem from the use of such samples, and the inadequate communication of their potential weaknesses.

Nonprobability samples can be useful for the purpose of monitoring biodiversity, provided that their limitations are assessed, mitigated where possible, and the almost inevitable remaining issues clearly communicated.

Monitoring the biodiversity crisis

There is a scientific consensus that the sixth mass extinction of life on earth is underway [1]. To understand the scale of the problem, data on the state of biodiversity, and how it has changed over time, are needed. Collecting and analysing such data is known as biodiversity monitoring and is an active area of research. As we will argue, however, biodiversity monitoring often rests on shaky statistical foundations.

There is no census of life on earth

Monitoring biodiversity is typically a matter of **descriptive statistical inference** (see Glossary). There is an implied finite population, which comprises all observation units of interest. These units might be, say, patches of land across a landscape. The researcher wants to infer something about those population units; for example, the average species richness. Putting measurement error to one side, it is simple to calculate this quantity if each patch of land has been sampled. In many cases, however, it is not possible to census all population units. In such situations, researchers rely on a sample and use sample-based **estimators** of the population quantities of interest. In the above example, it would be typical to use the sample mean as an estimator of the population mean.

Probability samples, nonprobability samples and estimator bias

Broadly speaking, statisticians define two types of sample: probability samples and nonprobability samples. In a probability sample, the probability that each population unit was included in the sample is known. The simplest type of probability sampling is simple random sampling (SRS), in which each population unit has an equal chance of selection [2]. In a nonprobability sample, the sampling mechanism, and therefore the chance that each population unit was sampled, is not known *a priori*.

Many sample-based estimators of population quantities – including sample means and proportions – are known to be unbiased under SRS [2]. A key property of SRS is that, as sample size increases, they are likely to be representative of the population (note the distinction between the variable

[1]UK Centre for Ecology and Hydrology, Benson Lane, Crowmarsh Gifford, Oxfordshire, OX10 8BB, UK

*Correspondence: robboy@ceh.ac.uk (R.J. Boyd).
[@]Twitter: @roboyd91 (R.J. Boyd), @GaryPowney (G.D. Powney), and @sacrevert (O.L. Pescott).

© 2023 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

representativeness of a single SRS, and the long-run unbiasedness of such samples over many hypothetical realisations). By representative, we mean that there is little to no correlation between an indicator variable, taking the value 1 if the population unit is in the sample and 0 otherwise, and the variable of interest [3,4]. If this correlation is ~0, the values of the variable of interest in the sample are similar to those in the population, and, consequently, sample averages, proportions and so forth are similar to their population equivalents.

In addition to SRS, it is relatively straightforward to construct unbiased estimators for other types of probability sample [2]. Recall that in a probability sample, the probability that each population unit was sampled is known by design. Again, where sample size is not small, these probabilities can be used to correct for inbuilt unrepresentativeness. For example, rather than estimating a population total using a sample total, one would instead use the weighted total, where the weights are equal to the inverse of the sample selection probabilities [2]. Weights of this type are known as **design weights**, and using them to construct unbiased estimators is known as **design-based inference**.

Matters are more complicated for nonprobability samples. Nonprobability samples are often unrepresentative of the population, but estimators of population quantities are unlikely to be fully adjustable using inclusion probabilities because these are not precisely known. Methods have been developed to mitigate unrepresentativeness in nonprobability samples, which we review later (see 'How can we do better?'), but it is typically very difficult to know how well this has been achieved [5].

Quantity does not necessarily imply quality

While it is challenging to construct unbiased estimators for nonprobability samples, they are frequently used for inferential research on pragmatic grounds [6]. Probability samples, even relatively small ones, are often very difficult to collect. By contrast, nonprobability samples, even large ones, are relatively easy to come by. Hence, researchers often justify the use of nonprobability samples on the basis of data quantity.

Unfortunately, quantity of data is no substitute for representativeness. Meng [4] derived a formula for the 'effective' size of a sample (Box 1). The effective size of a sample is equivalent to the size of the SRS that would yield an estimate of the population average with the same mean squared error (MSE), assuming that the sample average is used as the estimator. An important implication of the formula is that, where a sample is even slightly unrepresentative – that is, there is a correlation between sample membership and the variable of interest – the effective sample size becomes much smaller than the actual sample size (Box 1). The formula also shows that as the sampling rate (sample size divided by population size) increases, bias (i.e., nonsampling error, a.k.a. systematic or irreducible error) comes to dominate random sampling error.

The formula for calculating effective sample size yields startling results. In his original paper, Meng [4] analysed a dataset of size n = 2 300 000 on voting intentions for the 2016 US election. He found that the correlation between sample membership and respondents' intention to vote for Trump was a modest −0.005. This seemingly miniscule correlation led to an effective sample size of ~400 (a 99.98% reduction). More recently, Bradley and colleagues [7] used the same formula to analyse several surveys of coronavirus disease 2019 (COVID-19) vaccine uptake in the USA. They showed that the largest (but nonrandom) survey, with 250 000 responses per week, can produce estimates with the same error as a random sample of less than ten. Summarising, the authors stated, '[o]ur central message is that data quality matters more than data quantity, and that compensating the former with the latter is a mathematically provable losing proposition.'

Glossary

Descriptive statistical inference: the process of estimating some quantity describing a population from a sample of that population. This can be broadly contrasted with other inferential goals, such as prediction or causation.

Design-based inference: a mode of inference from sample to population. Design-based inference is most straightforward with probability samples, because sample inclusion probabilities can be used to adjust for selection bias. Design-based inference can be traced back to the work of Jerzy Neyman in the early- and mid-20th century.

Design variables: variables that influence, or are thought to influence, the probability that a population unit is sampled.

Design weights: often the inverse of sample inclusion probabilities. Design weights are used to correct for selection bias (as well as issues such as nonresponse) in design-based inference.

Estimator: a rule for calculating an estimate of a population quantity from a sample. An oft-cited example is the sample mean, which is an estimator of the population mean.

Estimator bias: a systematic deviation of a sample-based estimate from its corresponding population quantity. In a technical sense, this can imply different processes depending on the context, but it is always the irreducible, or 'nonsampling', component of error; that is, it cannot be reduced simply by adding more of the same 'type' of data.

Model-based inference: a mode of inference from sample to population. In model-based inference, a model thought to describe the variable of interest in the population is constructed, and inferences are drawn from this model. Model-based inference can be traced back to the work of Ronald Fisher in the early 20th century.

Representativeness: a common definition of sample representativeness is the correlation between an indicator variable, taking the value 1 if a population unit is in the sample and 0 otherwise, and the outcome variable of interest. Larger correlations equal lower representativeness. A representative sample has little selection bias, and vice versa. In some areas, representativeness is assessed using 'R-indicators', which are measures of the variability of sample inclusion probabilities.

CellPress
OPEN ACCESS

Box 1. Selection bias and data 'bigness'

Meng [4] derived a formula relating selection bias to the accuracy of the sample average $\overline{Y}_n$ as an estimator of the population average $\overline{Y}_N$. As it is not possible to provide the full derivation here, we simply present two relevant equations. The first shows that the difference between the sample average and the population average is:

$$(\overline{Y}_n) - (\overline{Y}_N) = \rho(R, Y)\sqrt{\frac{1-f}{f}}\,\sigma_Y \qquad [I]$$

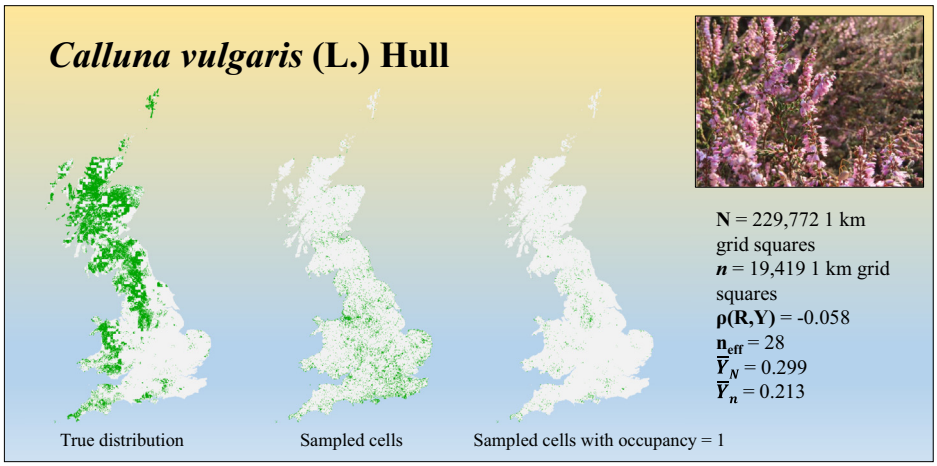
where $\rho(R, Y)$ is the (population) correlation between Y and an indicator variable taking the value 1 if the population unit is in the sample and 0 otherwise, f is the sampling rate (n/N), and the final term σ_Y is the population standard deviation of Y. $\rho(R, Y)$ indicates the direction and magnitude of the selection bias: when $\rho(R, Y) > 0$, larger values of Y are more likely to be in the sample than in the population and vice versa. Where $\rho(R, Y) = 0$, this term cancels the others and there is no error.

The second equation gives the 'effective' size n_{eff} of a sample, defined as the size of SRS that would produce an estimate of the population average with the same MSE (on average). n_{eff} may be expressed as:

$$n_{eff} = \frac{f}{(1-f)}\,\frac{1}{E\left[\rho(R, Y)^2\right]} = \frac{n}{1-f}\,\frac{1}{E\left[\rho(R, Y)^2\right]N} \qquad [II]$$

$E[\rho(R, Y)^2]$ is the expectation of the square of $\rho(R, Y)$ for a given data selection mechanism. $E[\rho(R, Y)^2]$ is expressed as an expectation because $\rho(R, Y)$ has many possible realisations for a given mechanism. However, where N is large, the variance in $\rho(R, Y)$ is typically negligible so $E[\rho(R, Y)^2]$ can be substituted by $\rho(R, Y)^2$ [4].

An important implication of this formula is that, where $\rho(R, Y)^2$ deviates even slightly from 0, n_{eff}/n (i.e., the relative effective sample size) decreases with N. In biodiversity monitoring, N is typically very large, so this reduction can be substantial. Thus, for nonprobability samples where $\rho(R,Y) \neq 0$, big data can be (effectively) very small. Figure I demonstrates the impact of selection bias on the effective sample size of a nonprobability biodiversity dataset.



Trends in Ecology & Evolution

Figure I. The effects of selection bias on effective sample size. Here, the population comprises 229 772 1 km land-containing grid cells in Britain.19 419 cells were sampled by volunteers, who submitted records of vascular plants to iRecord[i] via the smartphone app, the website, or indirectly via the iSpot initiative[ii] from 2000 to 2019. The variable of interest is the occupancy {1,0} of common heather *Calluna vulgaris*. True occupancy was estimated from the union of the UKCEH Land Cover Map [40], 1 km Atlas data (2000–2019; [36]), and the UK Countryside Survey 2007 plot data [41], constrained using the plant's 10 km distribution 2000–2019 [36]. This example assumes no measurement error.

Figure I highlights an interesting feature of Meng's formulae as applied to biodiversity monitoring: the area of the patches of land – their resolution – affects n, N and hence the sampling rate f = n/N. In our example, working at a coarser resolution, for example, 10 km^2, would increase f. This would be likely to reduce the error in our occupancy estimate and bring n_{eff} closer to n [36].

Selection bias: selection bias induces a correlation between an indicator variable, taking the value 1 if a population unit is in the sample and 0 otherwise, and the variable of interest. Hence, selection bias results in unrepresentative samples. If not dealt with, selection bias leads to **estimator bias**.

Figure I in Box 1 illustrates the effect of **selection bias** on our ability to accurately estimate mean occupancy of the common heather *Calluna vulgaris* in Britain. The selection bias, or the correlation between sample membership and occupancy, is –0.058. The sample-based estimate of mean occupancy is 0.213, which is a substantial underestimate of the population mean, 0.299. The effective sample size, 28, is 99.86% smaller than the actual sample size of 19 419!

The estimate of mean occupancy is not just wrong, but precisely wrong. The apparent sample size is large, which means that the normal approximation of the 95% confidence interval for the estimate of mean occupancy is narrow. Consequently, as we show in Box 2, it has virtually no chance of covering the true value. This is the big data paradox: 'the more the data, the surer we are to fool ourselves' [4].

Use and misuse of nonprobability samples in biodiversity monitoring

As we enter the era of 'big data', ecologists have access to more (and larger) nonprobability samples than ever before. Examples include digitised museum and herbarium collections [8], distribution data collected for species atlases, citizen science data, and newer types of data from various sensors (e.g., acoustic and radar; [9]). Much of these data are available through data aggregators and meta-databases such as GBIF (gbif.org) and BioTIME [10]. Probability samples, too, may be held in data aggregators, but these become nonprobability samples when combined with additional data. Given the various challenges associated with inference from nonprobability samples, it is not surprising that the increasing availability of these for research has led to some high-profile disagreements in the biodiversity literature (Table 1).

The disagreements in Table 1 arguably all relate to issues of sample representativeness, with the possible exception of the reply to [12]. Soroye and colleagues [12] used an occupancy-detection model to estimate changes in the range sizes of bees in North America and Europe. Occupancy-detection models require data on detections and nondetections, and the published criticism related to how nondetections were inferred. We, too, are sceptical about the models of Soroye *et al.* [12], but suggest that the most pernicious issue is likely to be the lack of representativeness in their data, not the precise way that nondetections were inferred (cf. [13]). We should remember that any model-based estimates, whether of a data selection mechanism (e.g., an observation process) or of an ecological state variable, can suffer from bias. Occupancy-detection models use information from repeat visits to the same area to estimate the probability that a species is detected given that it is present, a form of measurement error. These estimates could themselves be biased if the sites or visit types are not representative of the typical data-generating process [14]. We find that this is rarely, if ever, investigated by those applying occupancy-detection models to nonprobability samples, despite the fact that this is also a model of a real-world process.

The distinction between measurement error and representativeness can also be illustrated using Meng's formula for effective sample size (Box 1). The formula shows that, for a given selection bias, effective sample size is lower in the presence of measurement error [7]. This is not to say that removing measurement error would bring effective sample size back to parity with actual sample size. In our example (Box 1), we assume no measurement error, but a small amount of selection bias causes the effective sample size to be 99.86% lower than the actual sample size. This clearly demonstrates that dealing with measurement error alone does not fix issues caused by a lack of representativeness.

How can we do better?

Given such high-profile disagreements, the potentially wasted research time [15,16] and the clear issues with small effective sample sizes and unrepresentativeness, one might reasonably argue

CellPress
OPEN ACCESS

Box 2. The big data paradox

In Box 1, we demonstrated that the effective size n_{eff} of a biased sample, defined as the size of the SRS that would give an estimate of the population mean with the same MSE on average, is much smaller than its apparent size n. That is not to say that the estimate from the large, biased sample is as 'good' for inference as that from a SRS of size n_{eff}. The confidence interval for the estimate from the biased sample of size n will be misleadingly small relative to that for the SRS of size n_{eff} because $n_{eff} < n$. We have already seen in Box 1 that, where there is selection bias, the confidence intervals are centred on the wrong answer, which means that the tighter confidence interval for the larger sample is more likely to miss the truth. This is the big data paradox: 'the bigger the data, the surer we are to fool ourselves' [4].

To demonstrate the big data paradox, we return to the example from Box 1. We simulated 1000 SRSs of size n_{eff} (28), and 1000 samples of size n (19 419) with $\rho(R, Y) = -0.058$. For each sample, we calculated a 95% confidence interval for the mean occupancy estimate using the normal approximation. 94.4% of the SRS confidence intervals actually covered the true value, whereas none of the biased samples' confidence intervals covered the true value. That is, the large but biased samples gave us false confidence in our estimate, which turned out to be precisely wrong. The 95% 'confidence' interval is truly 0%. Figure I, which shows the distributions of the estimates of mean occupancy and squared error from the simulated samples, demonstrates this clearly.

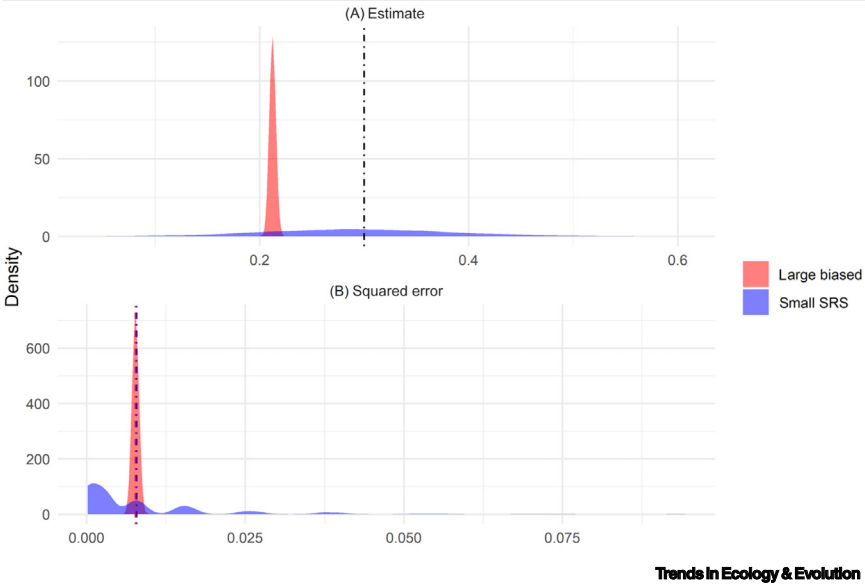


Figure I. Density plots showing the distributions of two statistics from 1000 simulated SRSs of size $n_{eff} = 28$ (small SRS) and 1000 of size n = 19 419 with $\rho(R, Y) = -0.058$ (large biased). Panel A shows the estimates of mean occupancy; the broken line denotes the true value, 0.3. Panel B shows the squared differences between the sample-based estimates of mean occupancy and the population mean. It also shows the MSE from each set of simulations (broken lines), which are very similar so hard to distinguish. These converge to the same value with increasing simulation number. Abbreviations: MSE, mean squared error; SRS, simple random sampling.

The interested reader should consult Meng [4], who demonstrated the big data paradox analytically. His analysis is based on the z score. Under the normal approximation, the z score denotes the half width of the confidence interval, in units of standard deviations, needed to obtain a given level of coverage. Meng shows that the z score needed to cover the true value is approximated by $\sqrt{n/n_{eff}}$, which is typically much larger than conventional values (e.g., 1.96 for 95% coverage). In our example, $\sqrt{n/n_{eff}} = 26.34$.

All the code and data needed to reproduce the analyses here and in Box 1 can be found at https://github.com/robboyd/selectionBiasEffects.

that ecologists should stop using nonprobability samples for the purpose of monitoring biodiversity. Indeed, some researchers essentially argue from this standpoint, and it is not unusual to see such samples written off as 'unscientific' (e.g., [17,18]). We do not fully accept

Table 1. Nine recent examples of high-profile biodiversity research papers with responses.[a] We have chosen examples based on our reading rather than a systematic search for disagreements. Issues with which respondents voiced concerns have been divided by us into three main sampling domains: geography, the environment, and taxonomy/others [11]. The final column ('Issues mentioned…') categorises the original paper as to whether the issues of sample representativeness were clearly acknowledged in either the title, abstract, or the paper's main body. The categories are: Y (Yes), N (No), and P (Partially). 'Partial' recognition can either be due to the fact that unrepresentativeness was mentioned for some domains but not others, or because of a lack of recognition that a bias mitigation strategy was likely to have weaknesses, as highlighted by the given response. Note that we do not review any potential weaknesses in the 'Responses' here, as we simply intend to demonstrate that there have been disagreements concerning sample coverage

Original paper	Response	Geographic domain	Environmental domain	Taxonomic/other domain	Issues mentioned in title/abstract/ main body?
Bruelheide *et al.* [42]	Christensen *et al.* [43]	–	Change in the biotope (environmental domain) focus between the survey periods generated incomplete data. The method to handle these gaps while assessing trends was not sufficient, leading to a systematic underestimate of declines.	–	N/P/P
Crossley *et al.* [44]	Welti *et al.* [45] (W) Desquilbet *et al.* [46] (D)	–	Several datasets used to assess species trends over time are from experimental sites (for example, a site where a target species was being systematically removed) or have inconsistent sampling methods over time.	Both W and D, note that non-insect taxa were included in the analysis despite the taxonomic domain of interest being restricted to insects.	N/Y/Y
Hallmann *et al.* [47]	Saunders [48] (S) Vereecken *et al.* [49] (V)	S notes a lack of repeat sampling at the same locations, potentially limits the ability to understand the true extent of declines.	S also notes that sites were restricted to small nature reserves.	V raise concerns about this paper (and Lister and Garcia [50]), where biomass is used as a proxy for declines in biodiversity and the services they provide. V show that the relationship between biomass and various biodiversity metrics/indicators vary with habitat type.	Y/P/Y
Leung *et al.* [51]	Murali *et al.* [52]	-	Murali *et al.* [52] note environmental bias in the populations included in the Living Planet Index (LPI), where populations inside protected areas are significantly over-represented. Therefore, the declines revealed in the LPI are likely to be worse at a global scale (the target domain of the LPI).	Murali *et al.* [52] also note that if the most extreme increasing vertebrate populations are excluded then the original LPI results remain broadly similar.	N/N/Y
WWF [53]	Leung *et al.* [51]	–	–	LPI estimates mean decline >50% for vertebrates since 1970. Leung *et al.* show that this is driven by <3% of the vertebrate populations included in the LPI; if excluded the mean trend becomes positive.	P

Table 1. (continued)

Original paper	Response	Geographic domain	Environmental domain	Taxonomic/other domain	Issues mentioned in title/abstract/ main body?
Newbold *et al.* [54]	Martin *et al.* [55]	–	Concerns about apparent underestimated losses in the biodiversity intactness index, likely driven by significantly differing levels of human impact on the baseline 'primary vegetation' sites used in the model to represent pristine condition.	–	N/N/P
Sánchez-Bayo and Wyckhuys [56]	Saunders [48] Simmons *et al.* [57]	Title states 'worldwide decline', replies highlight concerns around strong European and North American bias in the studies included in the review.	–	Many insect groups are completely absent from the analysis, for example, cockroaches, termites, and many fly and beetle families. Also the choice of search term (declin*) for the review could lead to bias in the subset of results, given the focus was 'all long-term insect surveys conducted over the past 40 years…'	N/N/P
Soroye *et al.* [12]	Guzman *et al.* [13]	Methodological issue where absence was inferred despite no evidence the location was visited, yielding biased estimates of decline. Furthermore, there was a large reduction in site visits between the two major time periods, particularly in North America and Southern Europe.	–	Species modelled across the entire geographic scope of the study, effectively assuming North American species may have been present at European sites, but had simply gone undetected, and vice versa for European bees at North American sites.	N/N/P
van Klink *et al.* [58]	Jähnig *et al.* [59] (J) Audisio *et al.* [60] (A) Scholl *et al.* [61] (S) Murray-Stoker and Murray-Stoker [62] (M-S) Desquilbet *et al.* [63] (D)	D highlights geographic bias in the modelled data with 76% of the studies covering the USA and Europe, despite a global geographic domain of interest. J also note the restricted nonrandom geographic representation of the study. A total of five datasets cover Africa, South America, and large parts of Asia.	S notes that many of the datasets included in the analysis are from studies examining insect responses to recent perturbation, and therefore are more likely to reflect population recovery.	J, A, S, M-S, all note that an increase in insect abundance should not be interpreted as a positive ecosystem response. This could be due to an increase in common, pollution-tolerant species, while overall richness declines. The domain of interest is insects (the title stating 'insect' abundance). However, D show that crustacean, mollusc, and worm data were included in the analysis. They also note that stress-tolerant species were over-represented.	N/N/P

[a]The LPI is a report: the P here refers to the main body of the report.

this rather harsh view, although we have some sympathy with it. Elsewhere, statisticians working on survey sampling consider nonprobability samples as almost a fact of life rather than an indictment and have embraced the resulting inferential challenges [5]. We suggest that many of the potential pitfalls of nonprobability samples could be avoided by formal assessments of

the risk-of-bias (RoB), clear and honest communication, and mitigation (to the extent possible) of the unrepresentativeness that is typical of such samples in ecology. Much can be learned from other disciplines in each of these areas (see Outstanding questions).

Formal risk-of-bias assessments

The field of medical research has arguably led the way in terms of assessing and documenting the types of issues we have raised: qualitative RoB assessments, for example, are typically required for primary studies and evidence synthesis in this area (Table 2). These often focus on the impact of bias on causal inference (e.g., in studies on medical interventions; [19]), but the principle can equally be applied to sampling, where the potential for bias relates to systematic differences between sample and population, rather than between treatment and control [20]. Simons and colleagues [21], for example, proposed the inclusion of a 'constraints on generality' statement within all primary psychological research papers, with the aim of highlighting the extent to which experimental findings on particular groups are likely to be generalisable. Boyd and colleagues [11] developed a similar initiative for biodiversity time trends, encouraging researchers to complete a formal 'Risk-Of-Bias In Temporal Trends' (ROBITT) assessment when publishing such descriptive studies (see also [22] for a broader discussion of generality in ecology). Table 2 lists examples of similar RoB tools across disciplines, including information on uptake.

Mitigation

Statisticians working on survey sampling problems have developed methods to adjust for unrepresentativeness in nonprobability samples. These methods generally come under the banner of '**model-based inference**', which may be contrasted with the design-based approach often used for probability samples (model-based approaches can also be used with probability samples, and may be preferable to design-based inference in some cases; [23]). The key distinction between these modes of inference is the way in which the population is treated: in design-based inference, the population values are treated as fixed (the random element is the sampling); in model-based

Table 2. Examples of qualitative RoB tools for different types of primary study-level assessments across disciplines[a]

Tool	Field	Study/data type	Details	Community promotion	Refs	Citations[b]
Cochrane RoB tool	Medical research	Randomised controlled trials of medical interventions	Used to qualitatively stratify meta-analyses according to RoB	Assessment of RoB is regarded as an essential component of a systematic review in this area	V1: [64]; V2: [19]	>40 000 [19]
CEE Critical Appraisal Tool	Environmental management	Experimental/quasi-experimental studies	For qualitative assessment of RoB across different study designs used in the research area	To assist environmental evidence synthesisers conduct critical appraisal	V0.3 (prototype): [65]	2
Constraints On Generality (COG) tool	Psychology	Any inferential study	Engenders clear definition of the statistical population of interest and assesses external validity	Requested by some psychology journals as a part of good practice in open science [66]	[21]	624
PROBAST	Medical research	Predictive modelling studies of diagnoses and prognoses	Used at either the primary research or systematic review stage	Endorsed and recommended by journals in the area	[67,68]	1136
ROBINS-I	Medical research	Nonrandomized studies of medical interventions	Compares data to that of a hypothetical randomised trial	Endorsed and recommended by journals in the area	[69]	6787
ROBITT	Ecology	Descriptive inference, especially temporal trends	Assessment of potential representativeness across relevant study domains	None to date	[11]	2

[a]Abbreviation: RoB, risk of bias; V, version.
[b]Citation numbers estimated by Google Scholar (1 September 2022) where not otherwise attributed.

inference, the population values are treated as a realisation of a stochastic data-generating process [24,25]. If this data-generating process, the model, can be recovered from the sample, then it can be used to draw inferences about the population [6,26,27].

Of course, there is a risk that a model constructed from a sample will not extrapolate well to nonsampled population units. To increase the chances that the model is representative of nonsampled units, **design variables**, thought to explain the sample selection process, can be included as covariates [26]. Alternatively, models can be used to estimate sample inclusion probabilities; these can be used to construct design weights, enabling the researcher to proceed in a similar way to design-based inference [28]. This sometimes is referred to as 'quasi-randomisation' [6] and is similar to propensity score weighting in causal inference [20].

Some attempts have been made to correct for the unrepresentative nature of big biodiversity datasets. In the Living Planet Index – an indicator of vertebrate population trends – the contribution of each biogeographic region to the global trend is weighted by an estimate of regional species richness [29]. Johnston and colleagues [30] estimated the probability that birds were searched for across spatial units in Great Britain; they then used these estimates to weight the likelihood function when fitting a regression model predicting species' occupancy. Maxent, a popular species distribution modelling algorithm, includes an option to 'factorBiasOut' [31]. The user provides information on sampling effort as input, and this is used to adjust the estimates of species' habitat suitability. These contributions are a good start for biodiversity science, but they need to become routine practice.

It should also be remembered that bias mitigation is not guaranteed to succeed, and that 'success' can be very difficult to evaluate [5,6]. For example, Boyd and colleagues [32] found that species distribution models (SDMs) fitted using a standard method to correct for sampling effort, the target group approach [33], were often rated as 'poor' by taxon experts. In some circumstances, statistical correction procedures have even been shown to make SDM predictions worse [34].

Communication

Even where bias mitigation strategies successfully reduce the RoB, it is unlikely that they will completely eliminate it: the remaining risk should be clearly communicated to readers and data users [35]. Issues of representativeness should also be conveyed in paper titles and abstracts, or at the least these should not be actively misleading through the omission of key qualifications (Table 1; see Outstanding questions). An example of good practice would be to specify the geographic and taxonomic extents of a study in its title, rather than letting the reader assume that the conclusions are more widely applicable. Researchers who understand a dataset, and the phenomenon being studied, should be able to integrate both bias-related and ecological factors when discussing results (e.g., see [36]). The RoB could also be communicated graphically, as in health research [37]. Similar approaches have recently been proposed for use with ecological indicators and species trends [38].

Concluding remarks

The word science ultimately derives from the Latin verb *scire*, 'to know', but do the types of biodiversity studies and approaches discussed here really yield accurate knowledge? It is difficult to know unless, as readers, we are presented with honest appraisals of the potential RoB and appropriate sensitivity analyses. As the statistician John Tukey opined towards the end of his career, '[t]he combination of some data and an aching desire for an answer does not ensure

Outstanding questions

Is the use of nonprobability samples cost-effective?
It could be argued that we are in a vicious cycle: if researchers continue to assert that it is possible to monitor biodiversity accurately using highly unrepresentative samples, then probability sampling is less likely to be funded.

Will new types of biodiversity data, for example, from acoustic, radar, and other sensors, make issues of representativeness better or worse?
These sensors have the potential to produce unprecedented quantities of data, but are likely to suffer from issues of representativeness. For example, acoustic monitoring stations could still be placed in an unrepresentative set of locations, and radar can only detect species above a particular size.

How can we incentivise communication of the RoB and unrepresentativeness in scientific papers?
High-impact journals seek eye-catching results, sometimes at the expense of rigour and clarity. If understanding the true state of our knowledge about reality is the goal, then the communication of such uncertainties has to improve.

What other methods for making inferences from nonprobability samples exist, and how reliable are they with real data?
Disciplines such as political science, epidemiology, and applied statistics have investigated methods for making inferences from nonprobability samples (or analogous methods for causal estimands). How many are likely to be suitable for ecology, and what should users report to fully communicate their likely final accuracy to the reader?

What other opportunities does the honest representation of uncertainty in biodiversity science present?
For example, expert assessments of the representativeness of aggregated biodiversity data, or other nonprobability samples, may provide additional opportunities for taxon and field-based experts to contribute to the ongoing data science revolution.

that a reasonable answer can be extracted from a given body of data' [39]. Papers, their titles, and their abstracts should all represent this uncertainty, and journals should not give authors a free pass on these in the pursuit of 'impact' and publicity. This will likely require a step change in the incentive structure for scientific publishing and funding ([15]; see Outstanding questions).

Acknowledgments

The authors thank Xiao-Li Meng, Andrea Stephens, Robin Walls, Nick Isaac, and three anonymous reviewers whose comments all improved the paper. We also thank Robin Hutchinson for facilitating access to the *Calluna* iRecord data, and the Botanical Society of Britain and Ireland for providing access to their *Calluna* Atlas data. All authors were supported by the NERC award number NE/R016429/1 as part of the UK Status, Change and Projections of the Environment (UK-SCAPE) programme delivering National Capability.

Declaration of interests

No interests are declared.

Resources

[i] www.irecord.org.uk

[ii] www.ispotnature.org/

References

1. Ceballos, G. and Ehrlich, P. (2018) The misunderstood sixth mass extinction. *Science* 360, 1080–1081
2. Lohr, S. (2022) *Sampling: design and analysis* (3rd edn), CRC Press
3. Groves, R.M. (2006) Nonresponse rates and nonresponse bias in household surveys. *Public Opin. Q.* 70, 646–675
4. 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. *Ann. Appl. Stat.* 12, 685–726
5. Rao, J.N.K. (2021) On making valid inferences by integrating data from surveys and other sources. *Sankhya B* 83, 242–272
6. Elliott, M.R. and Valliant, R. (2017) Inference for nonprobability samples. *Stat. Sci.* 32, 249–264
7. Bradley, V.C. *et al.* (2021) Unrepresentative big surveys significantly overestimated US vaccine uptake. *Nature* 600, 695–700
8. Boyd, R.J. *et al.* (2022) Inferring trends in pollinator distributions across the Neotropics from publicly available data remains challenging despite mobilization efforts. *Divers. Distrib.* 28, 1404–1415
9. van Klink, R. *et al.* (2022) Emerging technologies revolutionise insect ecology and monitoring. *Trends Ecol. Evol.* 37, 872–885
10. Dornelas, M. *et al.* (2018) BioTIME: a database of biodiversity time series for the Anthropocene. *Glob. Ecol. Biogeogr.* 27, 760–786
11. Boyd, R.J. *et al.* (2022) ROBITT: a tool for assessing the risk-of-bias in studies of temporal trends in ecology. *Methods Ecol. Evol.* 2022, 1–11
12. Soroye, P. *et al.* (2020) Among bumble bees across continents. *Science* 367, 685–688
13. Guzman, L.M. *et al.* (2021) Using historical data to estimate bumble bee occurrence : variable trends across species provide little support for community-level declines. *Biol. Conserv.* 257, 109141
14. Royle, J.A. (2006) Site occupancy models with heterogeneous detection probabilities. *Biometrics* 62, 97–102
15. Van Calster, B. *et al.* (2021) Methodology over metrics: current scientific standards are a disservice to patients and society. *J. Clin. Epidemiol.* 138, 219–226
16. Grainger, M.J. *et al.* (2020) Evidence synthesis for tackling research waste. *Nat. Ecol. Evol.* 4, 495–497
17. MacDonald, T. (2012) Spatial sampling designs for long-term ecological monitoring. In *Design and Analysis of Long-term Ecological Monitoring Studies*, pp. 101–125, Cambridge University Press
18. Anderson, D. (2008) *Model Based Inference in the Life Sciences* (1st edn), Springer, US
19. Sterne, J.A.C. *et al.* (2019) RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ* 366, 1–8
20. Mercer, A.W. *et al.* (2017) Theory and practice in nonprobability surveys. *Public Opin. Q.* 81, 250–279
21. Simons, D.J. *et al.* (2017) Constraints on Generality (COG): a proposed addition to all empirical papers. *Perspect. Psychol. Sci.* 12, 1123–1128
22. Spake, R. *et al.* (2022) Improving quantitative synthesis to achieve generality in ecology. *Nat. Ecol. Evol.* 6, 1818–1828
23. Royall, R.M. and Cumberland, W.G. (1981) An empirical study of the ratio estimator and estimators of its variance. *J. Am. Stat. Assoc.* 76, 66–77
24. Dumelle, M. *et al.* (2022) A comparison of design-based and model-based approaches for finite population spatial sampling and inference. *Methods Ecol. Evol.* 2022, 1–12
25. Little, R.J.A. (1982) Models for nonresponse in sample surveys. *J. Am. Stat. Assoc.* 77, 237–250
26. Sterba, S.K. (2009) Alternative model-based and design-based frameworks for inference from samples to populations: from polarization to integration. *Multivar. Behav. Res.* 44, 711–740
27. Williams, B.K. and Brown, E.D. (2019) Sampling and analysis frameworks for inference in ecology. *Methods Ecol. Evol.* 10, 1832–1842
28. Valliant, R. (2020) Comparing alternatives for estimation from nonprobability samples. *J. Surv. Stat. Methodol.* 8, 231–263
29. McRae, L. *et al.* (2017) The diversity-weighted living planet index: controlling for taxonomic bias in a global biodiversity indicator. *PLoS One* 12, 1–20
30. Johnston, A. *et al.* (2020) Estimating species distributions from spatially biased citizen science data. *Ecol. Model.* 422, 108927
31. Dudik, M. *et al.* (2005) Correcting sample selection bias in maximum entropy density estimation. *Adv. Neural Inf. Proces. Syst.* 323–330
32. Boyd, R.J. *et al.* (2022) Causal inference and large-scale expert validation shed light on the drivers of SDM accuracy and variance. *ecoevorxiv* Published online September 9, 2022. https://doi.org/10.32942/osf.io/68ats
33. Phillips, S.J. *et al.* (2009) Sample selection bias and presence-only distribution models: Implications for background and pseudo-absence data. *Ecol. Appl.* 19, 181–197
34. Inman, R. *et al.* (2021) Comparing sample bias correction methods for species distribution modeling using virtual species. *Ecosphere* 12, e03422
35. Blastand, M. *et al.* (2020) Five rules for evidence communication. *Nature* 587, 81–98

36. Stroh, P.A. *et al.* (2023) *Plant Atlas 2020: Mapping Changes in the Distribution of the British and Irish Flora*, Princeton University Press
37. McGuinness, L.A. and Higgins, J.P.T. (2021) Risk-of-bias VISualization (robvis): an R package and Shiny web app for visualizing risk-of-bias assessments. *Res. Synth. Methods* 12, 55–61
38. Pescott, O.L. *et al.* (2022) Simple methods for improving the communication of uncertainty in species ' temporal trends. *Ecol. Indic.* 141, 109117
39. Tukey, J.W. (1980) Sunset Salvo. *Am. Stat.* 34, 23–25
40. Morton, D. *et al.* (2014) *Land Cover Map 2007 (1km percentage target class, GB) v1.2*. https://doi.org/10.5285/fdf8c8d3-5998-45a5-8431-7f5e6302fc32
41. Carey, P. *et al.* (2008) *Countryside Survey: UK Results from 2007*, National Environmental Research Council/Centre for Ecology & Hydrology
42. Bruelheide, H. *et al.* (2020) Using incomplete floristic monitoring data from habitat mapping programmes to detect species trends. *Divers. Distrib.* 26, 782–794
43. Christensen, E. *et al.* (2021) Problems in using Beals' index to detect species trends in incomplete floristic monitoring data (Reply to Bruelheide et al. (2020)). *Divers. Distrib.* 27, 1324–1327
44. Crossley, M.S. *et al.* (2020) No net insect abundance and diversity declines across US long term ecological research sites. *Nat. Ecol. Evol.* 4, 1368–1376
45. Welti, E.A.R. *et al.* (2021) Studies of insect temporal trends must account for the complex sampling histories inherent to many long-term monitoring efforts. *Nat. Ecol. Evol.* 5, 589–591
46. Desquilbet, M. *et al.* (2021) Adequate statistical modelling and data selection are essential when analysing abundance and diversity trends. *Nat. Ecol. Evol.* 5, 592–594
47. Hallmann, C.A. *et al.* (2017) More than 75 percent decline over 27 years in total flying insect biomass in protected areas. *PLoS One* 12, e0185809
48. Saunders, M.E. (2019) No simple answers for insect conservation. *Am. Sci.* 107, 148
49. Vereecken, N.J. *et al.* (2021) Insect biomass is not a consistent proxy for biodiversity metrics in wild bees. *Ecol. Indic.* 121, 107132
50. Lister, B.C. and Garcia, A. (2018) Climate-driven declines in arthropod abundance restructure a rainforest food web. *Proc. Natl. Acad. Sci. U. S. A.* 115, E10397–E10406
51. Leung, B. *et al.* (2020) Clustered versus catastrophic global vertebrate declines. *Nature* 588, 267–271
52. Murali, G. *et al.* (2022) Reply to: Emphasizing declining populations in the Living Planet Report. *Nature* 601, E25–E26
53. WWF (2018) *Living Planet Report - 2018: Aiming higher*. p. 26
54. Newbold, T. *et al.* (2016) Has land use pushed terrestrial biodiversity beyond the planetary boundary? A global assessment. *Science* 353, 288–291
55. Martin, P.A. *et al.* (2019) The biodiversity intactness index may underestimate losses. *Nat. Ecol. Evol.* 3, 862–863
56. Sánchez-Bayo, F. and Wyckhuys, K.A.G. (2019) Worldwide decline of the entomofauna: a review of its drivers. *Biol. Conserv.* 232, 8–27
57. Simmons, B.I. *et al.* (2019) Worldwide insect declines: an important message, but interpret with caution. *Ecol. Evol.* 9, 3678–3680
58. van Klink, R. *et al.* (2020) Meta-analysis reveals declines in terrestrial but increases in freshwater insect abundances. *Science* 368, 417–420
59. Jähnig, S.C. *et al.* (2021) Revisiting global trends in freshwater insect biodiversity. *Wiley Interdiscip. Rev. Water* 8, 1–5
60. Audisio, P. *et al.* (2020) *eLetter: Are freshwater species really bucking the trend of global insect decline?*
61. Scholl, E.A. *et al.* (2020) *eLetter: Are human remediation efforts responsible for increases in insect abundance?*
62. Murray-Stoker, K.M. and Murray-Stoker, D. (2020) *eLetter: Context matters for studies of insect abundance and biomass*
63. Desquilbet, M. *et al.* (2020) Comment on "Meta-analysis reveals declines in terrestrial but increases in freshwater insect abundances." *Science* 370, 1–3
64. Higgins, J.P.T. *et al.* (2011) The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ* 343, 1–9
65. Konno, K. *et al.* (2021) *Collaboration for Environmental Evidence Critical Appraisal Tool, Version 0.3 (Prototype)*
66. Lindsay, D.S. (2019) Swan Song Editorial. *Psychol. Sci.* 30, 1669–1673
67. Wolff, R.F. *et al.* (2019) PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann. Intern. Med.* 170, 51–58
68. Moons, K.G.M. *et al.* (2019) PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. *Ann. Intern. Med.* 170, W1–W33
69. Sterne, J.A. *et al.* (2016) ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ* 355, 4–10