Received: 22 April 2024 | Accepted: 20 December 2024

DOI: 10.1111/2041-210X.14492

Methods in Ecology and Evolution

RESEARCH ARTICLE

Using causal diagrams and superpopulation models to correct geographic biases in biodiversity monitoring data

Robin J. Boyd[1] | **Marc Botham**[1] | **Emily Dennis**[2] | **Richard Fox**[2] | **Collin Harrower**[1] | **Ian Middlebrook**[2] | **David B. Roy**[1,3] | **Oliver L. Pescott**[1]

[1]UK Centre for Ecology and Hydrology, Crowmarsh Gifford, UK

[2]Butterfly Conservation, Wareham, UK

[3]Centre for Ecology and Conservation, University of Exeter, Falmouth, Cornwall, UK

Correspondence
Robin J. Boyd
Email: robboy@ceh.ac.uk

Funding information
UKCEH National Capability for UK Challenges Programme, Grant/Award Number: NE/Y006208/1; Natural Environment Research Council, Grant/Award Number: NE/R016429/1 and NE/X010384/1; Butterfly Conservation; UK Centre for Ecology & Hydrology; British Trust for Ornithology; Joint Nature Conservation Committee

Handling Editor: Nicolas Lecomte

Abstract

1. Biodiversity monitoring schemes periodically measure species' abundances and distributions at a sample of sites to understand how they have changed over time. Often, the aim is to infer change in an average sense across some wider landscape. Inference to the wider landscape is simple if the species' abundances and distributions are similar at sampled to non-sampled locations. Otherwise, the data are geographically biased, and some form of correction is desirable.
2. We combine causal diagrams with 'superpopulation models' to correct time-varying geographic biases in biodiversity monitoring data. For a given time-period, expert-derived causal diagrams are used to deduce the set of variables that explain the geographic bias, and superpopulation models adjust for these variables to produce a corrected estimate of a landscape-wide mean of for example abundance or occupancy. Estimating a time trend in the variable of interest is achieved by fitting models for multiple time-periods and, if the drivers of bias are suspected to change over time, by constructing per period causal diagrams. We test the approach using simulated data then apply it to real data from the UK Butterfly Monitoring Scheme (UKBMS).
3. If the variables that explain the geographic bias are known and measured without error, our method is unbiased. Introducing measurement error reduces the method's efficacy, but it is still an improvement on using the sample mean. When applied to data from the UKBMS, the approach gives different results to the scheme's current method, which assumes no geographic bias.
4. Where the goal is to estimate change in some variable of interest at the landscape level (e.g. biodiversity indicators), models that do not adjust for geographic bias implicitly assume it does not exist. Our approach makes the weaker assumption that there is no geographic bias conditional on the adjustment variables, so it should yield more accurate estimates of time trends in many circumstances. The method does require assumptions about the drivers of bias, but these are codified explicitly in the causal diagrams. Operationalising our approach should

be less costly than full probability sampling, which would be needed to satisfy the assumptions of conventional approaches.

KEYWORDS
directed acyclic graph, expert consultation, imputation, sampling bias, species abundance, time trend

1 | INTRODUCTION

1.1 | Monitoring biodiversity

The goal of biodiversity monitoring is to understand how some variable of interest *Y*, often when summarised across the relevant landscape or study area, has changed over time (cf. Yoccoz et al., 2001). *Y* might be a species' (relative) abundance, occupancy or some summary thereof across many species. The landscape is often delimited by national boundaries (Moussy et al., 2022), but in principle it could be any geographic area. *Y* is usually summarised across the landscape periodically by taking its mean or something similar (e.g. a categorical year effect from a model in which *Y* is the response; Brereton, Roy, et al., 2011; Powney et al., 2019; Stroh et al., 2023).

To calculate the landscape-wide mean of *Y* directly, one would need to know its distribution across the landscape as a whole (i.e. its value at every location). Generally, however, the landscape distribution of *Y* is not known, because it is not possible to collect data at every location. Instead, the usual strategy is to split the landscape into areal units, or 'sites', and to collect data at some of them. The landscape-wide mean of *Y* is then *estimated* from the sample of sites for which data are available.

1.2 | Geographic sampling biases

When estimating the landscape-wide mean of *Y* from a sample of sites, there is a risk of geographic sampling bias. A dataset is geographically biased if the distribution of *Y* across sampled sites differs from its distribution across the wider landscape. An analogous definition is the existence of a non-zero correlation between *Y* and a sample inclusion indicator *R*, which takes the value 1 at sampled sites and 0 elsewhere (Aubry et al., 2024; Boyd, Powney, & Pescott, 2023; Meng, 2018; see Table 1 for list of key variables). If the correlation is positive, then the bias is positive (i.e. the mean of *Y* is larger across sampled sites than across the landscape as a whole), and vice versa. The greater the magnitude of the correlation, the more severe the bias.

Biodiversity monitoring data are often collected in such a way that the same factors affect *R* and *Y*, which induces a correlation between the two (and therefore a bias). To give one simple example (many others are possible), *Y* might be some species' abundance, which is positively affected by habitat quality. Habitat quality might also have a positive effect on *R*, because data collectors are often volunteer naturalists (i.e. citizen scientists) and prefer to visit sites that are interesting in terms of wildlife (Bowler et al., 2022; Forister et al., 2023). In this situation, *Y* would be larger at sampled sites, which is to say that there would be a positive bias. The reverse would be true if sampling was more likely at sites where species are faring poorly: say, in built-up areas that have poor quality habitat, but which are easy for recorders to access by road (Hughes et al., 2020).

Data from structured monitoring schemes are less susceptible to geographic biases than those from most other sources (Geldmann et al., 2016), but they are not immune. Many structured monitoring schemes aim to collect data at a random sample of sites (e.g. Pescott et al., 2019; Robbins et al., 1986). If implemented properly, random sampling ensures no correlation between *R* and *Y* in expectation (i.e. on average over many possible samples; see Meng, 2018). Obtaining a true random sample is challenging, however, especially if there is a reliance on volunteers to collect the data (which is true of many monitoring schemes). The volunteers might be unwilling or unable to visit some randomly selected sites, whether because they are difficult to access, uninteresting in terms of wildlife or for some other reason (Pescott et al., 2015).

TABLE 1 Description of key variables and sets of variables.

Variable	Denoted by	Description
Species (relative) abundance	*Y*	An index of the number of individuals of the focal species present at each site or some other variable of analytic interest (e.g. occupancy)
Sample inclusion indicator	*R*	Binary variable taking the value 1 for sampled sites and 0 elsewhere
Auxiliary variables (adjustment set)	***A***	A set of variables that are not of direct analytic interest and which we assume to be measured at every site (e.g. via satellite). A good set of auxiliary variables explain the dependence between *R* and *Y*

Nominally random samples with incomplete uptake of sites, or 'nonresponse' in survey sampling parlance, are almost certain to be biased (Bailey, 2023a).

1.3 | Mitigating geographic biases

Various 'model-based' approaches have been developed to mitigate sampling biases (Buckland et al., 2000), many of which exploit the fact that some set of variables A could explain the correlation between R and Y (Lohr, 2022b; Meng, 2022). (Other model-based approaches to dealing with sampling biases work on different premises; we do not cover these here except for a brief overview in the Section 4.) If, to use the above example, R and Y are both greater at sites with good quality habitat, then the presence of that habitat will explain some of the correlation between the two. When the variables in A can be identified—a point we come back to below—and are reflected in available data, statistically 'adjusting for' them will reduce the sampling bias (Collins et al., 2001; Mohan et al., 2013).

One way to adjust for A is to fit a 'superpopulation model', which is a regression (or other type of) model for Y that includes A as covariates (Elliott & Valliant, 2017). In effect, the superpopulation model stratifies sites based on levels of A and predicts the average or expected value of Y for each stratum. If A explains the correlation between R and Y, the two should be uncorrelated within each stratum, and the predicted within-stratum means should be close to their true values (as there is no sampling bias). Averaging the within-stratum means of Y with respect to the distribution of A across all sites in the landscape yields an 'adjusted' estimate of the landscape-wide mean of Y (the target quantity). More simply, the average of the superpopulation predictions across all sites is the estimate of the landscape-wide mean—or, in practice, the average of the predictions for non-sampled sites and the observations for sampled sites (Elliott & Valliant, 2017)—and it is unbiased if A completely explains the geographic bias.

1.4 | Causal diagrams and what to adjust for

To identify the variables in A, insight can be gleaned from causal inference, where a similar challenge arises. Recall that A is the set of variables that explains the correlation between R and Y. In causal inference, where to goal is to estimate the causal effect of one variable on another, analysts must identify and adjust for the set of variables that explain the non-causal portion of the relationship between the two (Pearl et al., 2016). One way to identify these variables, which could also be used to identify the variables in A, is to construct 'causal diagrams' (Greenland et al., 1999; Thoemmes & Mohan, 2015).

Causal diagrams—not to be confused with parametric structural equation models—will not be familiar to some ecologists (but see Grace & Irvine, 2020), so we will introduce the relevant concepts using the example in Figure 1 (a full description of the graph in Figure 1 is provided in the Methods section). An arrow from one variable to another indicates a direct causal effect; that is, it indicates that the cause is part of the real-world function that determines the value of the effect (Pearl et al., 2016). In Figure 1, annual temperature (annual_temp) has a direct effect on Y, which is a species' abundance. A path consists of several variables linked by arrows regardless of the direction of those arrows. Figure 1 depicts a path linking R (site inclusion) to Y: R ← heather_grass → heather_grass_quality → Y. The existence of a path between two variables implies that they are correlated, whether by association or a causal link. Hence, Figure 1 implies a correlation between R and Y and therefore a geographic bias.

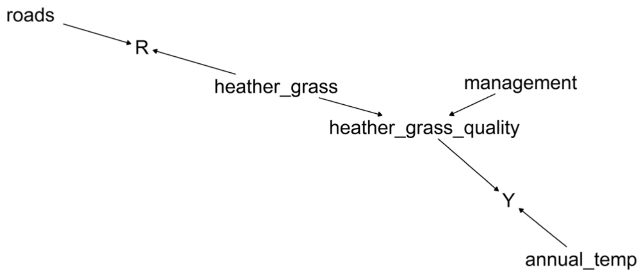


FIGURE 1 A simplistic causal diagram depicting causes and effects of sample inclusion R and a species' abundance Y.

While a path between two variables implies that they are correlated (dependent[1]), it does not imply that the dependence is unbreakable. Rather, it might be possible to "block" paths between variables, which is to say, to block the flow of association. Blocking a path is achieved by adjusting for certain variables (i.e. A), and a set of rules—the rules of d-separation—tell us which ones (Cinelli et al., 2022; Pearl et al., 2016). It is not necessary to spell out the rules of d-separation here, partly because they are described elsewhere (see the references above) and partly because they are built into software packages such as the R package dagitty (Textor et al., 2016). The important point is that the rules can be used to determine the sets of variables A that, when statistically adjusted for, will render R and Y independent (if such a set exists; Thoemmes & Mohan, 2015).

Identifying the correct set of variables in A is contingent on the causal diagram being an accurate reflection of reality, and there are broadly two ways to achieve this (which are not mutually exclusive). One is to consult domain experts, who understand the relevant system, when constructing the diagram (Grace & Irvine, 2020). The other is to develop the diagram iteratively and test the implied conditional independencies (according to the rules of d-separation) of each iteration (Pearl et al., 2016). Where data are missing on Y for most sites (i.e. where $R = 0$), it will not be possible to test many of a diagram's implied conditional independencies, in which case consulting domain experts is even more important than usual. Of course,

---

[1] For convenience, we will use the terms correlation and dependence interchangably. Strictly speaking, a nonlinear dependence might not imply a correlation.

experts cannot be expected to have a perfect understanding of most systems, and we come back to this point in several places below.

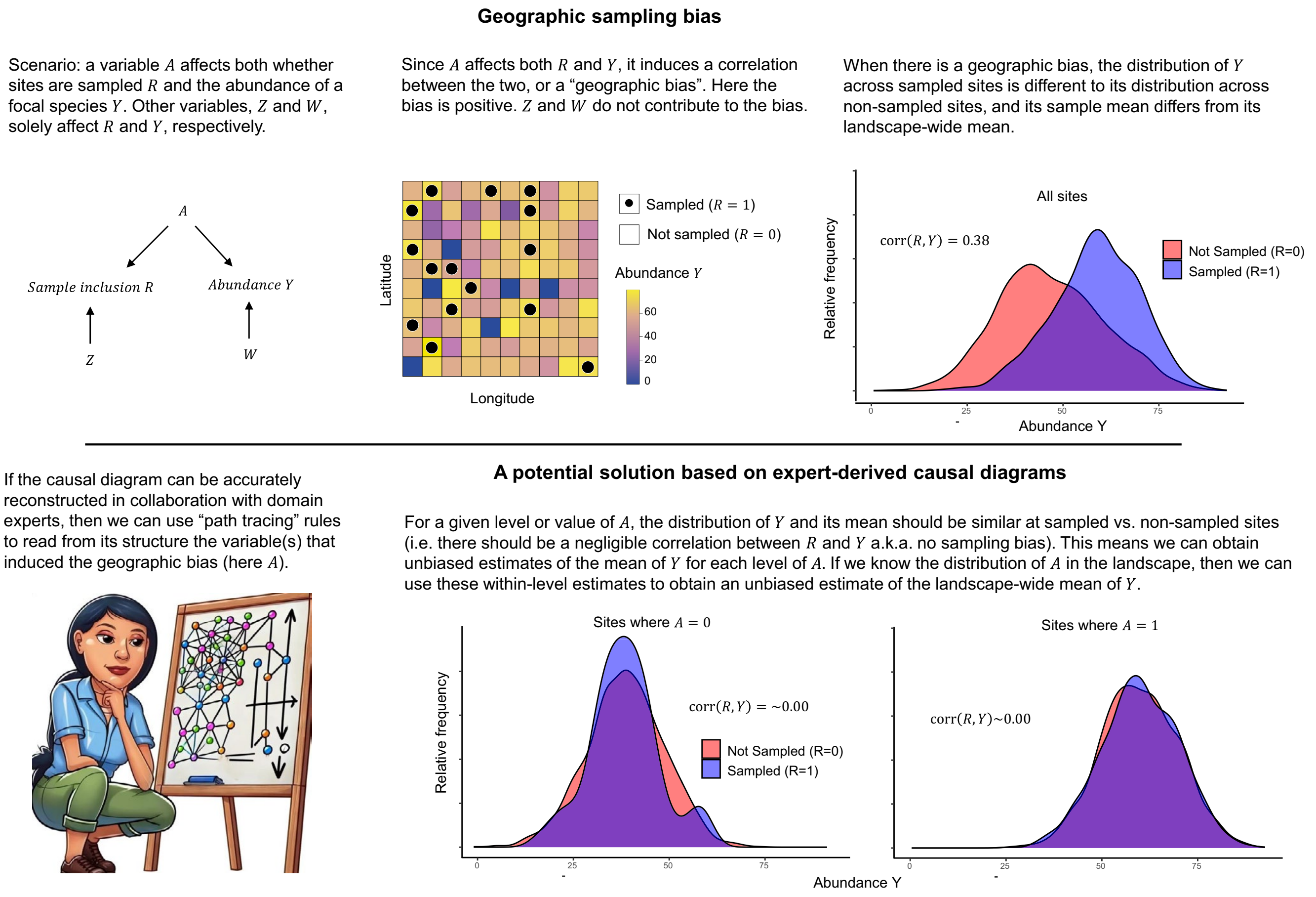
1.5 | Structure of this paper

In this paper, we demonstrate how one might combine expert knowledge with causal diagrams and superpopulation models to mitigate geographic biases in biodiversity monitoring data (see Figure 2 for an overview). We assume that a time trend in the mean of *Y* across all sites in the relevant landscape is the target quantity, and it is estimated by fitting separate superpopulation models for each of several time-periods. (An alternative approach is to estimate parameters describing the time trend in *Y* itself, which is perfectly valid but which we do not consider here.) Starting with some simple simulations, we test the abilities of superpopulation models including *A* as covariates to recover the landscape-wide mean of *Y* from perfect and imperfect (i.e. subjected to measurement error) data on *A*. Next, we apply our method to empirical data from the UK Butterfly Monitoring Scheme (UKBMS). The scheme estimates time trends in the mean abundances butterflies in the UK using a model that does not adjust for geographic bias, and we compare these trends to those produced using superpopulation models. In the final section, we discuss the pros and cons of our approach and how it could be improved in future.

2 | METHODS

2.1 | Demonstrating superpopulation modelling via simulation

Although superpopulation models can mitigate geographic biases in theory, reality is more complex. It is likely, for example, that data will not be available on some variable(s) in *A* at the exact location of a site. Instead, the available data might represent an average across some larger area in which the site is situated. Alternatively, data might be available on *A* at the precise location of the site but not for the relevant time-period. In these cases, we would expect the data to correlate with the true variable(s) in *A*, but not to be a perfect proxy. To test the implications of including imperfect data on *A* in a superpopulation model, as well as some other methodological



Scenario: a variable A affects both whether sites are sampled R and the abundance of a focal species Y. Other variables, Z and W, solely affect R and Y, respectively.

Since A affects both R and Y, it induces a correlation between the two, or a “geographic bias”. Here the bias is positive. Z and W do not contribute to the bias.

When there is a geographic bias, the distribution of Y across sampled sites is different to its distribution across non-sampled sites, and its sample mean differs from its landscape-wide mean.

If the causal diagram can be accurately reconstructed in collaboration with domain experts, then we can use “path tracing” rules to read from its structure the variable(s) that induced the geographic bias (here A).

For a given level or value of A, the distribution of Y and its mean should be similar at sampled vs. non-sampled sites (i.e. there should be a negligible correlation between R and Y a.k.a. no sampling bias). This means we can obtain unbiased estimates of the mean of Y for each level of A. If we know the distribution of A in the landscape, then we can use these within-level estimates to obtain an unbiased estimate of the landscape-wide mean of Y.

FIGURE 2 Schematic depicting a simple scenario that illustrates how causal diagrams can be used to identify and correct sampling biases in biodiversity monitoring data. In this example, the target parameter is the mean of a variable *Y* (e.g. species abundance) across a predefined set of sites, a fraction of which have been sampled.

choices, we conducted a simple simulation study. (see Supporting Information S3 for an additional simulation that includes two time-periods and the estimation of a trend).

Our simulation was based on the causal diagram in Figure 1. The diagram depicts the causes of some species' abundance *Y* and sample inclusion *R* across 250,000 sites in a simple virtual landscape. It implies that the quantity of good quality heather grassland (heather_grass_quality) and annual temperature at each site directly affect *Y*. The quantity of good quality heather grassland at a site is determined by the total quantity of heather grassland (of any quality; heather_grass) and the proportion that is well managed (management). *R* is greater at sites with more heather grassland, because recorders know that it is a favourable habitat for the focal species, which they hope to see. The presence of major roads near a site makes it more accessible so it also affects *R*. According to the rules of d-separation, the causal diagram in Figure 1 implies that *R* is independent of *Y* given heather grassland (i.e. there is only one variable in *A*).

To simulate the data, we used the simulateSEM function in the R package dagitty (Textor et al., 2016). With the exception of *R*, all variables are standard normal (i.e. mean of 0 and unit variance). Other than causes of *R* (see below), each variable explains ~25% of the variation in its direct descendent. *R* is a binary variable (0 if the data is missing and 1 otherwise), so we could not simulate it in the same way as the others. Instead, following Thoemmes and Rose (2014), we simulated a latent standard normal variable and discretized it by assigning all values above the first percentile the value 1 and the remainder the value 0 (i.e. we assume data on *Y* are missing for 99% of sites). A consequence of this strategy is that each cause of *R* explains less than 25% of its variance, because discretizing the latent normal variable attenuates the correlation between the two. Strictly speaking, the latent normal variable should be included in the causal diagram in Figure 1, but it does not affect its implied conditional independencies, so we omit it for simplicity.

Having simulated the data, we estimated the mean of *Y* across all sites using the data at sites where $R = 1$ (i.e. the sample). We used six estimators, which are listed in Table 2. The first is the sample mean, and the remainder are superpopulation models—in this case, linear regressions—including different covariates. The first superpopulation model includes *A* (heather_grass) as a covariate. The second includes *A* and an additional cause of *Y* (annual_temp), which should increase precision (Cinelli et al., 2022). The remainder of the superpopulation models include correlates of *A* (with different strengths of correlation) to reflect the fact that the available data are unlikely to be error-free.

We evaluated the superpopulation models' performances in terms of their estimation error. The procedure involved simulating 1000 datasets based the causal diagram in Figure 1 and calculating the difference between the true, landscape-wide mean of *Y* and the superpopulation model estimate for each one. The 1000 datasets differed slightly from one another, as there is a random component to the simulateSEM function. Averaging the estimation error across the 1000 estimates of mean *Y* for each method gives an estimate of its estimator bias.

TABLE 2 Six analytical approaches to estimating the mean of *Y* (species abundance) across all sites in the simulated landscape. The adjustment set is the set of covariates included in the superpopulation model.

Estimator	Adjustment set	Details
1	NA	Sample mean
2	Heather_grass	Superpopulation model
3	Heather_grass, annual_temp	Superpopulation model
4	Correlate of heather_grass ($\rho = 0.5$)	Superpopulation model
5	Correlate of heather_grass ($\rho = 0.7$)	Superpopulation model
6	Correlate of heather_grass ($\rho = 0.9$)	Superpopulation model

2.2 | UKBMS case study

2.2.1 | The scheme

To demonstrate how causal diagrams and superpopulation models might be used to correct biases in a real dataset, we applied them to data from the UK Butterfly Monitoring Scheme (UKBMS). The UKBMS has been running since 1976 (Pollard & Yates, 1996). Data are collected by volunteers, who walk transects at a network of sites in the UK and count the butterflies they see within an imaginary 5-m box when weather permits (Pollard, 1977). The volunteers are free to decide where to establish transects and generally do so in good quality semi-natural habitat, where butterflies are most abundant (Brereton, Roy, et al., 2011). Recorders are asked to walk UKBMS transects at least once in each of the 26 weeks from April 1 to September 29 to cover the main flight periods of UK butterflies. On average, however, ~20 weeks are sampled due to poor weather conditions, recorder availability and so forth.

In 2009, the Wider Countryside Butterfly Survey (WCBS) was established and incorporated in the UKBMS. The primary motivation for the WCBS was to increase coverage of habitats that were poorly represented by the UKBMS (Brereton, Cruickshanks, et al., 2011). Hence, rather than being chosen by the volunteers, WCBS sites are located within randomly selected 1 km grid squares in the UK. Importantly, however, volunteers are not willing/able to visit all WCBS squares: they prefer to visit and to re-visit sites that are accessible or where they are likely to see species that interest them. Consequently, WCBS squares are not truly random. WCBS transects also differ from traditional UKBMS transects in that volunteers are asked to walk them a minimum of twice in July–August (although some do more) at least 10 days apart.

2.2.2 | Existing analytical method

The current method used to analyse the UKBMS (including WCBS) data has three steps (Dennis, Morgan, Freeman, Brereton, &

Roy, 2016). The first is to fit a GAM to estimate normalised seasonal abundance curves for each species and year (these do not vary geographically; Dennis et al., 2013). The second is to estimate annual site indices of relative abundance using the fitted GAMs. The total observed counts are scaled by the proportion of the seasonal abundance curve that was surveyed to provide an estimate of the expected total had the entire season been surveyed. Finally, a Poisson GLM with site and year effects is fitted to the annual site indices, and the estimated year effects are used as annual indices of relative abundance. In the final GLM, each site/year index is weighted in proportion to the fraction of the 26-week flight curve that was sampled (rather than imputed). Consequently, the WCBS sites, which are sampled as little as twice per year, should be heavily downweighted.

2.2.3 | New analytical method

We modified the current UKBMS framework by replacing the final GLM with superpopulation models that adjust for *A*. Steps one and two, which produce seasonally adjusted annual site indices of relative abundance, remain the same and the superpopulation models are fitted to these indices. Taxon and dataset experts provided the information needed to construct causal diagrams and identify the variables in *A* (more details below).

For demonstrative purposes, we focused on two species of butterfly, whose geographic distributions and ecologies are very different: the meadow brown (*Maniola jurtina*) and the small pearl-bordered fritillary (*Boloria selene*). The causes and effects of *Y*, that is relative abundance, differ between the two species. As we are only working with one dataset, collected by the UKBMS (including WCBS sites), the causes and effects of *R* (site inclusion) do not differ between species.

We asked two taxon and UKBMS experts to provide information on the causes and effects of *R* and *Y* via the forms in Supporting Information S1 and S2. Both experts (DBR and IM) are authors on this paper. To constrain their answers, we asked the experts to select causes and effects from a list of 49 land cover (i.e. habitat), bioclimatic, geological and other variables on which data are available (see Supporting Information S3 for details of the data). Of course, this strategy risks omitting important variables—a point we come back to in the Discussion. The experts provided their feedback independently.

Rather than the transect-level, we asked the experts to think about causes and effects of *R* and *Y* at the coarser resolution of 1 × 1 km, which was necessary for two reasons. First, it is the finest resolution at which data on some variables in *A* are available. The second reason is more complex. Recall that the superpopulation model predictions of *Y* must be averaged across all sites in the landscape. It would be conceptually challenging the split the landscape into imaginary transects, which are the 'true' sites, but splitting it into 1 × 1 km grid squares is simple. In the few cases where multiple UKBMS transects fell within the same grid square, we averaged the site indices across those transects in a given year to obtain a grid square-level data point.

The information provided by the experts enabled us to construct causal diagrams and, using the rules of d-separation, to identify the variables in *A* (according to each expert). In total, we produced four causal diagrams: one for each species and expert. For any one diagram, there may be multiple sets of variables *A* that d-separate *R* and *Y*. We selected the 'minima' (i.e. smallest) set that only included variables on which data are available (listed in Table 3).

Both experts indicated that some land cover (habitat) types have negative effects on *R* and *Y*, but we did not include these in the causal diagrams. Our logic is that these land cover types are only detrimental to the species in the sense that they are not the right habitat (indeed, this is how it was phrased by expert two [IM]). The land cover data are expressed as the proportion of each 1 km grid square covered by each land cover type (Morton et al., 2022), which means that a large proportion of one necessarily means a small proportion of another (i.e. the land cover types are not independent). That is, a high proportion of favourable habitat would imply a low proportion of unfavourable habitat, and vice versa, so it is not necessary to include unfavourable land cover types in the causal diagrams.

Having identified two sets of variables *A* (one from each expert) for each species, the next step was to fit superpopulation models including these variables as covariates. We fitted one model per species, expert (i.e. *A*) and year. As the data are (normalised) counts, and for consistency with the current UKBMS method, we used Poisson GLMs with a log link function (Dennis, Morgan, Freeman, Brereton, Roy, Ecology, et al., 2016):

$$Y_i \sim \text{Pois}(\lambda_i)$$

$$\log(\lambda_i) = \boldsymbol{A}_i \boldsymbol{\beta},$$

where λ_i is the expected count at grid square i, $\boldsymbol{A}_i$ is the vector of adjustment variables *A* at grid square i, and $\boldsymbol{\beta}$ is the vector of coefficients to be estimated. A prediction of the expected count for grid square i is

$$\hat{\lambda}_i = e^{\boldsymbol{A}_i \hat{\boldsymbol{\beta}}},$$

where $\hat{\boldsymbol{\beta}}$ is the iteratively reweighted least squares estimate of $\boldsymbol{\beta}$. If s is the set of sampled grid squares and $\bar{s}$ is the set of non-sampled grid squares, the superpopulation model estimate of the landscape-wide mean of *Y* is (Elliott & Valliant, 2017)

TABLE 3 Minimal adjustment sets derived from the experts' knowledge for each species. The adjustment set need not include all predictors of the species' abundances.

Expert	Species	Minimal adjustment set
1	Meadow brown	Calcareous grassland
1	Small pearl-bordered fritillary	Broadleaved woodland
2	Meadow brown	Calcareous grassland and elevation
2	Small pearl-bordered fritillary	Broadleaved woodland, calcareous bedrock and heather

$$\hat{Y}_N = \frac{\sum_{i \in s} y_i + \sum_{i \in \bar{s}} \hat{\lambda}_i}{N}.$$

That is, it is the sum of the observed counts at sampled sites and the predicted counts at non-sampled sites divided by the total number of sites in the landscape N. To measure change over time, we used the slope from a linear regression of $\hat{\hat{Y}}_N$ on year.

For simplicity, we treated the model covariates (i.e. A) as constant over time. This was a natural choice for elevation and calcareous bedrock, which are genuinely time-invariant (at least since 1976). Our approach might seem less appropriate for the three land cover classes that featured in at least one of the models (Table 3), but inspection of the available land cover maps shows that there has been little change in the recent past (85% of 25 × 25 m grid squares in Great Britain were the same land cover class in 2015 as in 1990; Supporting Information S3). Full details of the auxiliary data are provided in Supporting Information S3.

2.2.4 | Comparing trends from the current and new methods

To enable comparisons between the year effects from the current UKBMS models and the annual estimates of mean abundance from the superpopulation models, we present all outputs on the log10 scale and relative to the values in the final year of the time-series (2022).

2.2.5 | Variance estimation

For both the superpopulation models and the simpler GLM used by the UKBMS, we estimated the sampling variance in each year's index of relative abundance by bootstrapping across sites. We created 1000 bootstrap samples by resampling the site indices with replacement within each year (since different sites were sampled in different years) and fitted a model to each sample. We present percentile (95%) confidence intervals for each year from the bootstrap distribution of the annual indices (year effects from the simple GLM and estimates of mean Y from the superpopulation models). In principle, one could also use bootstrapping to estimate the variance of the species' seasonal flight curves. Since we are primarily interested in the effects of geographic bias on the recoverability of mean abundance in each time-period, and for consistency with the UKBMS, we did not go to these lengths here, but we acknowledge that the uncertainty intervals could be larger (for both approaches) if we had.

3 | RESULTS

3.1 | Simulations

Figure 3 shows the estimation error—the difference between the true, landscape-wide mean of Y and the estimate thereof—of each analytical method listed in Table 1. The sample mean is highly biased (mean error, i.e. bias, of ~0.35 in units of standard deviations of Y). Using a superpopulation in which A is the sole covariate (A comprises just one variable in this example), the estimate of mean Y across all sites is unbiased (scenario 2; mean error ~0). The estimate is also unbiased if A plus an additional cause of Y are included as covariates (scenario 3), and it has slightly lower variance than the estimate from scenario 2. Including a correlate of A as the covariate reduces the bias relative to the sample mean but does not eliminate it (scenarios 4–6; see Table 2). The stronger the correlation between A and its proxy, the closer the bias to that from scenario 2, where A itself was included as a covariate.

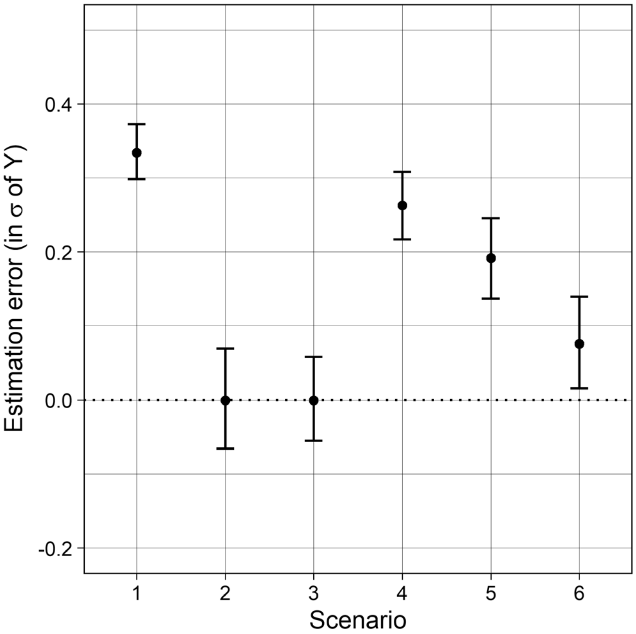


FIGURE 3 Estimation error (difference between the true, landscape-wide mean of Y and the estimate thereof) under each of the six scenarios in Table 2. The dots represent the mean error across 1000 simulated datasets, and the error bars depict the 2.5th and 97.5th percentiles.

3.2 | UKBMS case study

3.2.1 | Identifying the variables in *A*

The experts provided different perspectives on the causes and effects of R (UKBMS site inclusion including WCBS sites) and Y (the abundances of the meadow brown and small pearl-bordered fritillary). Their feedback is provided in full in Supporting Information S1 and S2. Figure 4 shows the causal diagram reflecting expert two's knowledge of the causes and effects of R and the Y in the case of the small pearl-bordered fritillary.

A major difference between the two experts' feedback is that expert two indicated positive effects of several land cover classes (habitats) on Y conditional on the way that they are managed. That is, the habitats support high abundances of the species, but only if they are managed appropriately. We codified these conditional effects in

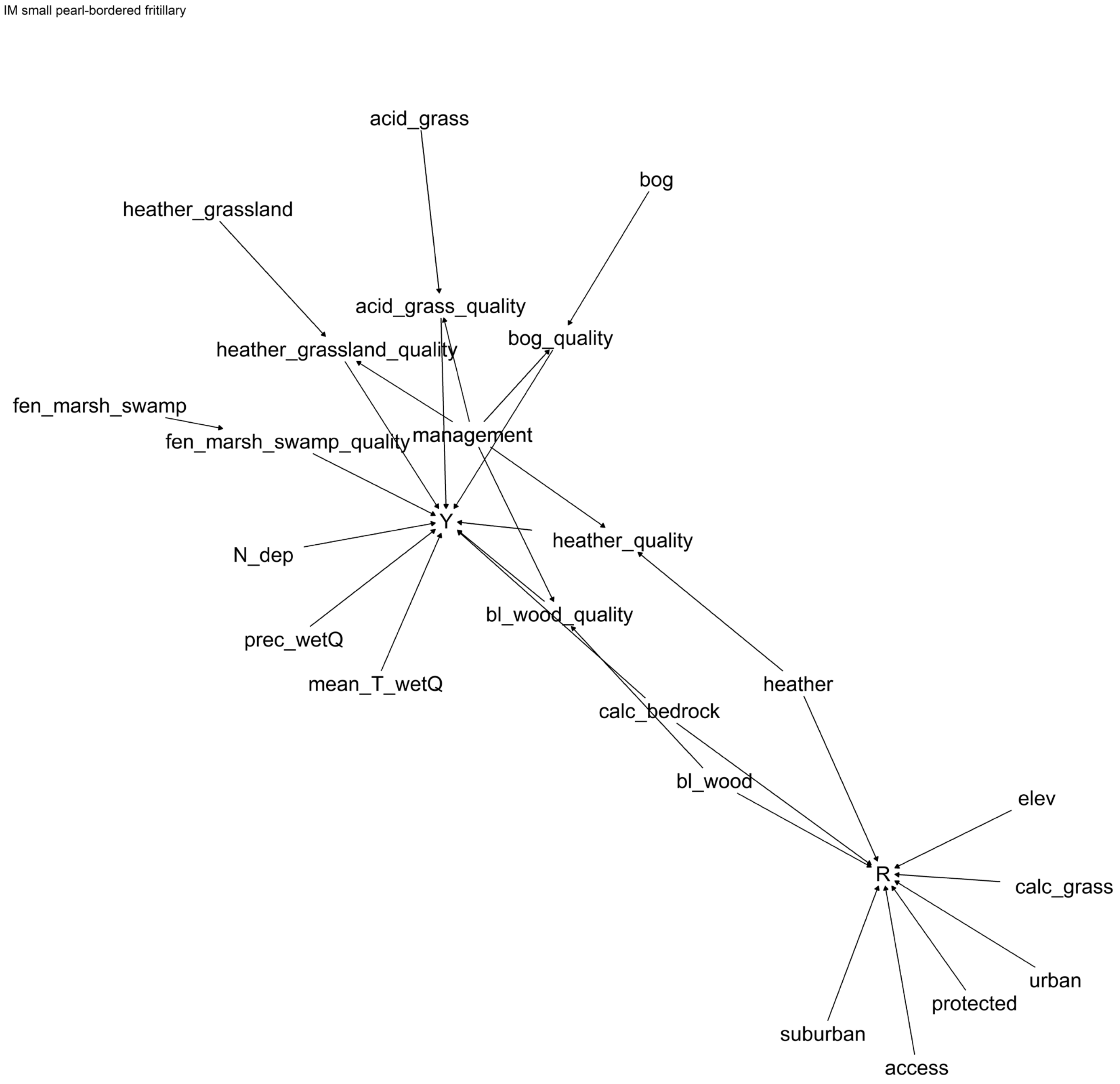


FIGURE 4 Causal diagram depicting causes and effects of sample inclusion in the UK Butterfly Monitoring Scheme (*R*) and the abundance of the small pearl-bordered fritillary (*Y*). The diagram reflects the knowledge of a taxon and UKBMS expert (expert two). N_dep is nitrogen deposition, mean_T_wet_Q is the mean temperature of the wettest quarter of the year, elev is elevation, bl_wood is broadleaved woodland, protected is the proportion of each grid square that is some form of protected area and variables suffixed by '_quality' denote the proportion of each grid square that is both the relevant land cover type (indicated by the rest of its name) and managed appropriately for the species. The other variables are self-explanatory.

the causal diagrams by introducing new, *unobserved* variables. One is 'management', which indicates whether the habitat is managed appropriately for the species, and the others are variables that denote the proportion of each site that is both the relevant land cover class and managed appropriately (i.e. 'good' quality). For example, the expert indicated that heather grassland has a positive effect on the abundance of the small pearl-bordered fritillary conditional on appropriate management. In the causal diagram, the effect of heather grassland on *Y* is heather grassland → quality heather grassland → *Y*, where management also has a direct effect on quality heather grassland (Figure 4). Note that we did not specifically ask the experts to provide information on conditional effects, which is why expert one did not.

From each of the four causal diagrams—one for each species and expert—we identified the minimal adjustment set *A* (the variables on which *R* and *Y* are hopefully conditionally independent) using the rules of d-separation (Table 3). For each diagram, it is possible to d-separate *R* and *Y* without having to adjust for any of the unobserved variables introduced to depict the conditional effects of land cover classes on *Y* (i.e. management or 'quality' habitats). For both species, the variables in *A* differ between experts, which meant that two superpopulation models had to be fitted for each one.

3.2.2 | Current versus new model outputs

We compared the superpopulation models' outputs to those of the current UKBMS GLM in terms of the annual estimates and the long-term trend, which is the coefficient from a regression of annual index on year (Figures 5 and 6). For both species, the outputs of the superpopulation models are markedly different to those of the UKBMS method. For the meadow brown, the two superpopulation models (one for each expert) agree that there has been a decline in abundance, whereas the UKBMS method suggests that abundance has been stable over time. For the small pearl-bordered fritillary, the current UKBMS method suggests a decline in abundance over time, whereas the two superpopulation models indicate a slight increase.

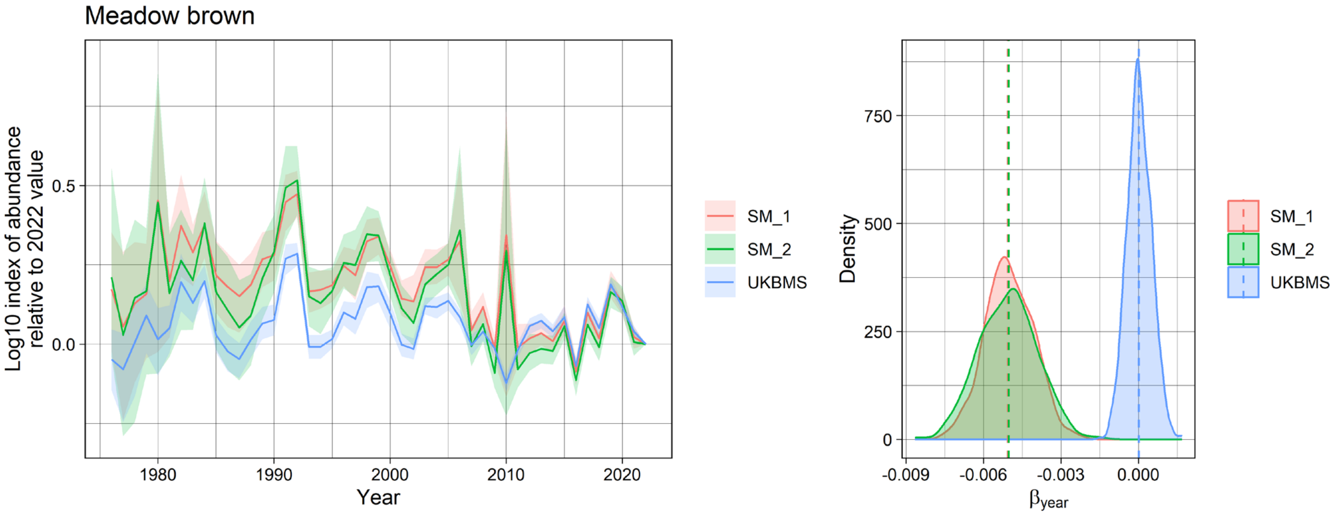


FIGURE 5 Time trends in the annual indices of abundance for the meadow brown produced by the current UKBMS GLM (UKBMS), the superpopulation model based on expert one's knowledge (SM_1) and the superpopulation model based on expert two's knowledge (SM_2). The left panel shows the time-series, and the right panel shows the estimated trends, which are the coefficients from a regression of index value on year. The uncertainty in both panels was derived by bootstrapping across sites. The vertical lines in the right-hand panel represent the median slopes.

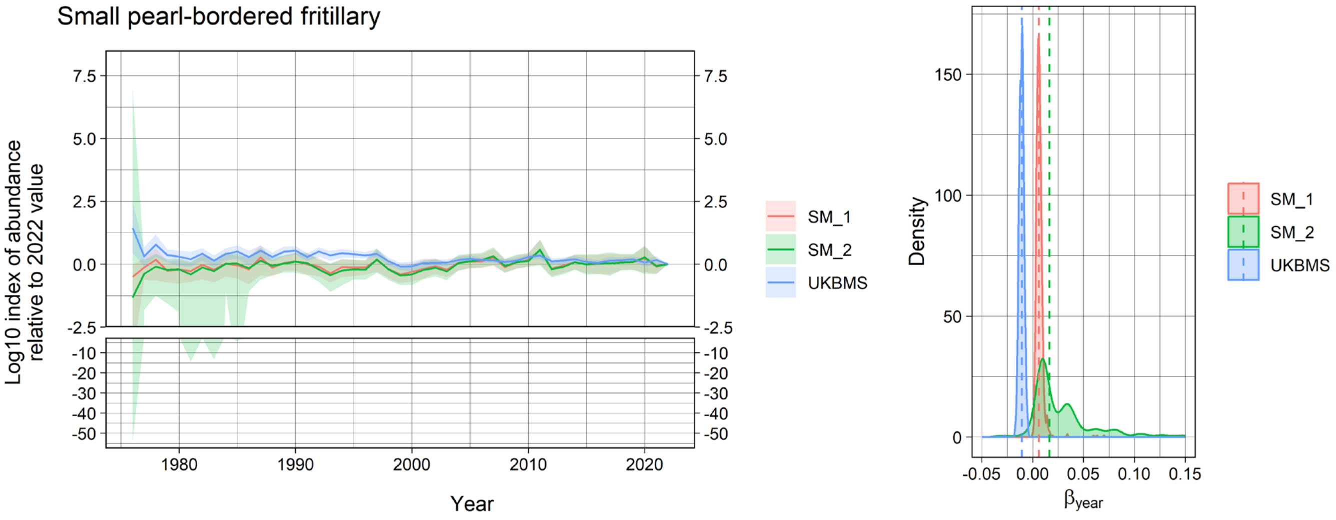


FIGURE 6 Time trends in the annual indices of abundance for the small pearl-bordered fritillary produced by the current UKBMS GLM (UKBMS), the superpopulation model based on expert one's knowledge (SM_1) and the superpopulation model based on expert two's knowledge (SM_2). The left panel shows the time-series, and the right panel shows the estimated trends, which are the coefficients from a regression of index value on year. The uncertainty in both panels was derived by bootstrapping across sites. The break in the *Y* axis on the left panel was introduced to show the full uncertainty in SM_2 whilst also enabling comparison of the three models. The vertical lines in the right-hand panel represent the median slopes.

4 | DISCUSSION

We have demonstrated how causal diagrams and superpopulation models might be used to estimate the landscape-wide mean of some variable of interest *Y* (e.g. a species' abundance or occupancy) from geographically biased monitoring data. Our simulations show that, if the variables that explain the geographic bias (i.e. *A*) are measured (without error), known and included in the superpopulation model, then that model is unbiased (Figure 3). Clearly, these ideal conditions are unlikely, a point we come back to below. By fitting models for multiple time-periods, we estimated time trends in the mean relative abundances of two species of butterfly using data from the UKBMS. The trends are different to those estimated using the UKBMS's existing method, which does not adjust for geographic bias (Figures 5 and 6).

The experts (DBR, IM, MB and RF) were not confident about whether the superpopulation models more accurately captured the UK-wide trends for the two species than the existing UKBMS method. The UKBMS conducts a regional breakdown of each species' trend in the UK (https://ukbms.org/official-statistics). For the small pearl-bordered fritillary, it reports a strong decline in England but a strong increase in Scotland. Scotland is relatively underrepresented in the UKBMS data, so a good adjustment should place more weight on the trend in this region. The superpopulation models clearly give more weight to the Scottish trend than the UKBMS method—one even suggests that the species is increasing at the UK level. Nevertheless, two experts (DBR and RF) are sceptical that the small pearl-bordered fritillary is not declining at the UK level (based on e.g. an observed decline in its distribution). All experts felt that the superpopulation model estimates for the meadow brown are plausible (although none were highly confident). Their reasoning is that the species is probably doing better or at least as well at sampled than at non-sampled locations.

Unfortunately, it is not possible to say for sure whether a superpopulation model has successfully corrected sampling biases, that is reduced the correlation between *R* and *Y*, because data on *Y* are not available where $R = 0$ (Lohr, 2022a). Instead, the analyst must use extra-statistical logic to assess the plausibility of the assumption that sampling bias is smaller having accounted for *A*. Insight into the plausibility of this assumption can be gleaned directly from the causal diagrams. For example, the diagram might reveal that a variable on which no data are available induces a dependence between *R* and *Y*, in which case, according to that diagram, the superpopulation model will not have eliminated the bias.

Since it is not possible to quantify sampling biases, or the extent to which any given adjustment has corrected them, qualitative "risk-of-bias" assessment tools might prove useful. One example, which was developed specifically for biodiversity monitoring, is ROBITT (Boyd et al., 2022). ROBITT comprises a series of questions about whether there is evidence of sampling bias (geographic and otherwise) and how the analyst intends to correct it. Causal diagrams depicting the assumed drivers of the sampling bias could be used to answer these questions.

From a theoretical perspective, there are some limited cases in which the UKBMS model and others like it could capture the true trend in the landscape-wide mean of a species' abundance *Y* despite not adjusting for *A*. If, for example, the bias is time-invariant in terms of both sign and magnitude, then no adjustment is needed. This scenario is highly unlikely, however. The distribution of sample inclusion *R* changes over time, because different sites are sampled in different years. So too does the distribution of *Y* unless the focal species' relative abundance remains the same at each site in each time-period, which is clearly implausible. As the distributions of *R* and *Y* change over time, it is highly unlikely that the correlation between the two, that is the bias, would remain constant. We also note that the existing UKBMS model could produce unbiased estimates of alternative estimands using the UKBMS data. For example, one might be interested in the trend in the mean abundance across occupied sites (i.e. where $Y \geq 1$), in which case sampling all populations would permit unbiased inference. Of course, this would require knowing the locations of all populations, and range shifts would complicate matters.

Whether a superpopulation model estimate of the landscape-wide mean of *Y* is less biased than a naïve (i.e. unadjusted) one depends on several factors, some of which were captured by our simulations. Where data on all *A* are available, but they were measured with error, the adjustment should reduce but not eliminate bias (Figure 3). Matters are more complex where some variables in *A* are omitted from the model and/or where some variables that are not in *A* are erroneously included (both of which may result from mis-specifying the causal diagram). Erroneously included/omitted variables that are more strongly related to *R* and *Y* have greater potential to cause a bias, because they can induce a larger correlation between the two (Collins et al., 2001; Thoemmes & Rose, 2014). However, the sign of the relationships is also relevant: if one erroneously omitted/included variable induces a negative geographic bias and another induces a positive one, then the two might cancel each other out and cause no bias (Thoemmes & Rose, 2014). Our simulations do not capture the effects of mis-specifying the causal diagrams, because any choice of a select few scenarios of strengths and directions of effects and causal diagram structures would have been necessarily arbitrary, but analysts should recognise that the success of the superpopulation modelling approach is contingent on accuracy of the causal diagram.

Indeed, constructing accurate causal diagrams is the major practical limitation of our approach. Analysts might have hundreds or even thousands of species in mind, and experts might not have the knowledge to construct causal diagrams for each one (noting that, for multispecies surveys such as the UKBMS, it is only the *Y* part of the causal diagram that must be constructed separately for every species, not the *R* part). Even if they did, it would be a time-consuming exercise (although less costly than full probability sampling, which would be needed to satisfy the assumptions of conventional methods). One option for reducing this burden might be to identify and adjust for a set of variables that have very large causal effects on *R* and *Y* for many species: for example, woodland in the case of woodland birds or grassland in the case of grassland butterflies. Such a strategy is likely to miss important variables in *A* for some species, however, so it should only be used as a last resort. Another option might be to crowd-source information on the causes and effects of *R* and *Y* from taxon and dataset experts. This strategy could yield feedback from more than two experts (as here) and potentially many possible adjustment sets *A*. A decision would need to be made about whether to fit a model for each unique *A* or whether to ask the experts to reach a consensus. We think that fitting models for each *A* is preferable, as it allows the analyst to assess the sensitivity of the results to the experts' feedback. Indeed, it is apparent in Figure 6 that the results can vary appreciably depending on the structure of the causal diagram and by extension which variables are included in the superpopulation model. If needed, a single measure of change could be derived by averaging across the models whilst propagating the inter-model uncertainty.

Another question for the analyst is how comprehensive the causal diagrams need to be. Variables that explain a negligible portion of the correlation between *R* and *Y* do not have the capacity to cause much bias (Thoemmes & Rose, 2014) and can be safely omitted from the causal diagram (assuming they are truly "unimportant"). Such variables might be distal in the sense that their effects are highly indirect (i.e. they are far upstream of *R* and *Y*), or they could have small but direct effects. In our examples, the experts did not report on variables whose causal effects on *R* and *Y* they considered to be very small (of course, this was based on intuition), and we did not ask them to report on distal variables. We do not claim that limiting the diagrams in this way will always be the best approach, but unnecessarily complex causal diagrams will result in a large set of variables in *A*, which could increase the variance in the superpopulation model estimates for very little benefit (especially if collinearity becomes a problem).

Having identified the variables in *A*, superpopulation models are just one of several methods that could be used to adjust for them. Alternatives include inverse probability weighting (Fink et al., 2023; Johnston et al., 2020), poststratification (Boyd, Stewart, & Pescott, 2023; Van Swaay et al., 2002) and more elaborate techniques such as Multilevel Regression and Poststratification (MRP; first conceived by Gelman & Little, 1997). If *Y* is uncorrelated with *R* for given values of *A*, then sufficient information is available for each of these methods to recover the landscape-wide mean of *Y* in principle (Lohr, 2022a; Meng, 2022). In practice, however, violation of modelling assumptions (e.g. about the form of the relationships between *R*, *Y* and *A*) can introduce bias, so it might be preferable to opt for a model that permits flexible functional forms etc. (e.g. Fink et al., 2023).

Alternative methods exist that do not require complete knowledge of or data on the variables in *A*, but these too come with strong assumptions. A practical option is to model geographic bias using spatial random fields, which assign a value of sampling intensity to every location in the landscape based on spatial autocorrelation in the sampling locations (e.g. Simmonds et al., 2020). Use of this approach essentially substitutes the assumption that *R* and *Y* are independent given *A* for the assumption that they are independent given the spatial field and other covariates in the model (Diggle et al., 2010). Selection models, which involve modelling *R* and *Y* separately and making assumptions about their joint distribution, are another alternative (Bailey, 2023b). For an interesting discussion of the relative merits of methods that do and do not assume that *R* and *Y* are uncorrelated given *A*, see Bailey (2023a) and commentaries thereon.

5 | CONCLUSIONS

Although alternatives are available, most model-based estimators of population parameters (here the landscape-wide mean of a species' abundance) assume that there is no sampling bias once a set of auxiliary variables has been adjusted for. In any given situation, there might exist a set of auxiliary variables that satisfy this assumption but identifying them is notoriously challenging. Historically, analysts have relied on empirical rules of thumb to identify the auxiliaries, but this strategy is unlikely to yield the correct set (Schuessler & Selb, 2023). Consequences of adjusting for the wrong set of variables range from failing to eliminate bias (best case) to inadvertently increasing it (Thoemmes & Rose, 2014). Expert-informed causal diagrams are a viable alternative to empirical criteria and can, in theory, identify the exact set of auxiliary variables needed to eliminate sampling bias (Thoemmes & Mohan, 2015). Readers considering the use of causal diagrams to identify auxiliary variables might find our code (Boyd, 2025) helpful and should work with domain experts who have knowledge of the chosen dataset and taxon in question.

AUTHOR CONTRIBUTIONS

Robin J. Boyd and Oliver L. Pescott conceived the idea and developed the methodology. Robin J. Boyd led the writing of the manuscript. All authors commented critically on the manuscript. Ian Middlebrook and David B. Roy provided the information needed to construct the causal diagrams. Colin Harrower and Emily Dennis provided the data and demonstrated how to use it. Marc Botham quality checked the data.

ACKNOWLEDGEMENTS

R.J.B. and O.L.P. were supported by the NERC Exploring the Frontiers award number NE/X010384/1 "Biodiversity indicators from nonprobability samples: Interdisciplinary learning for science and society". R.J.B. was also supported by the UKCEH National Capability for UK Challenges programme NE/Y006208/1. O.L.P. was also supported by the NERC award number NE/R016429/1 as part of the UK-SCAPE programme delivering National Capability. The UK Butterfly Monitoring Scheme is organised and funded by Butterfly Conservation, the UK Centre for Ecology & Hydrology, British Trust for Ornithology, and the Joint Nature Conservation Committee. The UKBMS is indebted to all volunteers who contribute data to the scheme.

CONFLICT OF INTEREST STATEMENT

The authors have no conflicts of interest to disclose.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/2041-210X.14492.

DATA AVAILABILITY STATEMENT

All code and data needed to reproduce the small pearl-bordered fritillary example and the simulations are available on Zenodo at https://doi.org/10.5281/zenodo.14605048 (Boyd, 2025).

ORCID

Robin J. Boyd https://orcid.org/0000-0002-7973-9865
Emily Dennis https://orcid.org/0000-0003-2692-4415
Richard Fox https://orcid.org/0000-0001-6992-3522
Oliver L. Pescott https://orcid.org/0000-0002-0685-8046

REFERENCES

Aubry, P., Francesiaz, C., & Guillemain, M. (2024). On the impact of preferential sampling on ecological status and trend assessment. *Ecological Modelling*, *492*, 110707. https://doi.org/10.1016/j.ecolmodel.2024.110707

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

Bailey, M. A. (2023b). *Polling at a crossroads: Rethinking modern urvey research*. Cambridge University Press (forthcoming). https://doi.org/10.2753/RSL1061-1975230163

Bowler, D. E., Bhandari, N., Repke, L., Beuthner, C., Callaghan, C. T., Eichenberg, D., Henle, K., Klenke, R., Richter, A., Jansen, F., Bruelheide, H., & Bonn, A. (2022). Decision-making of citizen scientists when recording species observations. *Scientific Reports*, *12*(1), 1–12. https://doi.org/10.1038/s41598-022-15218-2

Boyd, R. J. (2025). Code and data from: Using causal diagrams and superpopulation models to correct geographic biases in biodiversity monitoring data. *Zenodo*. https://doi.org/10.5281/zenodo.14605048

Boyd, R. J., Powney, G. D., Burns, F., Danet, A., Duchenne, F., Grainger, M. J., Jarvis, S. G., Martin, G., Nilsen, E. B., Porcher, E., Stewart, G. B., Wilson, O. J., & Pescott, O. L. (2022). ROBITT: A tool for assessing the risk-of-bias in studies of temporal trends in ecology. *Methods in Ecology and Evolution*, *13*(March), 1497–1507. https://doi.org/10.1111/2041-210X.13857

Boyd, R. J., Powney, G. D., & Pescott, O. L. (2023). We need to talk about nonprobability samples. *Trends in Ecology & Evolution*, *38*(6), 521–531. https://doi.org/10.1016/j.tree.2023.01.001

Boyd, R. J., Stewart, G. B., & Pescott, O. L. (2023). Descriptive inference using large, unrepresentative nonprobability samples: An introduction for ecologists. *Ecology*, *105*(2), e4214.

Brereton, T., Cruickshanks, K. L., Risely, K., Noble, D. G., & Roy, D. B. (2011). Developing and launching a wider countryside butterfly survey across the United Kingdom. *Journal of Insect Conservation*, *15*(1), 279–290. https://doi.org/10.1007/s10841-010-9345-8

Brereton, T., Roy, D. B., Middlebrook, I., Botham, M., & Warren, M. (2011). The development of butterfly indicators in the United Kingdom and assessments in 2010. *Journal of Insect Conservation*, *15*(1), 139–151. https://doi.org/10.1007/s10841-010-9333-z

Buckland, S. T., Goudie, I. B. J., & Borchers, D. L. (2000). Wildlife population assessment: Past developments and future directions. *Biometrics*, *56*(March), 1–12. https://academic.oup.com/biometrics/article/56/1/1/7263482

Cinelli, C., Forney, A., & Pearl, J. (2022). A crash course in good and bad controls. *Sociological Methods & Research*, *53*, 1–30. https://doi.org/10.1177/00491241221099552

Collins, L. M., Schafer, J., & Kam, C. (2001). A comparison of restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*(June), 330–351. https://doi.org/10.1037/1082-989X.6.4.330

Dennis, E. B., Freeman, S. N., Brereton, T., & Roy, D. B. (2013). Indexing butterfly abundance whilst accounting for missing counts and variability in seasonal pattern. *Methods in Ecology and Evolution*, *4*(7), 637–645. https://doi.org/10.1111/2041-210X.12053

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. https://doi.org/10.1111/biom.12506

Dennis, E. B., Morgan, B. J. T., Freeman, S. N., Brereton, T. M., Roy, D. B., Ecology, C., Lane, B., & Gifford, C. (2016). A generalized abundance index for seasonal invertebrates. *Biometrics*, *72*, 1305–1314. https://doi.org/10.1111/biom.12506

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

Elliott, M. R., & Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, *32*(2), 249–264. https://doi.org/10.1214/16-STS598

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

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*, e12951. https://doi.org/10.1111/conl.12951

Geldmann, J., Heilmann-Clausen, J., Holm, T. E., Levinsky, I., Markussen, B., Olsen, K., Rahbek, C., & Tøttrup, 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. https://doi.org/10.1111/ddi.12477

Gelman, A., & Little, T. (1997). poststratification into many categories using hierarchical regression. *Survey Methodology*, *23*(2), 127–335.

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

Greenland, S., Pearl, J., & Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology*, *10*(1), 37–48. https://doi.org/10.1097/00001648-199901000-00008

Hughes, A., Orr, M., Ma, K., Costello, M., Waller, J., Provoost, P., Zhu, C., & Qiao, H. (2020). Sampling biases shape our view of the natural world. *Ecography*, *44*, 1259–1269. https://doi.org/10.1111/ecog.05926

Johnston, A., Moran, N., Musgrove, A., Fink, D., & Baillie, S. R. (2020). Estimating species distributions from spatially biased citizen science data. *Ecological Modelling*, *422*(December 2019), 108927. https://doi.org/10.1016/j.ecolmodel.2019.108927

Lohr, S. (2022a). Comments on "Statistical inference with non-probability survey samples". *Survey Methodology*, *48*(12), 331–338.

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

Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 us presidential election. *Annals of Applied Statistics*, *12*(2), 685–726. https://doi.org/10.1214/18-AOAS1161SF

Meng, X.-L. (2022). Comments on the Wu (2022) paper by Xiao-Li Meng 1: Miniaturizing data defect correlation: A versatile strategy for handling non-probability samples. *Survey Methodology*, *48*(2), 1–22.

Mohan, K., Pearl, J., & Tian, J. (2013). Graphical models for inference with missing data. *Advances in Neural Information Processing Systems*, 1277–1285.

Morton, R., Marston, C., O'Neil, A., & Rowland, C. (2022). *Land cover map 2018 (1km summary rasters, GB and N. Ireland)*. NERC EDS Environmental Information Data Centre. https://doi.org/10.5285/9b68ee52-8a95-41eb-8ef1-8d29e2570b00

Moussy, C., Burfield, I. J., Stephenson, P. J., Newton, A. F. E., Butchart, S. H. M., Sutherland, W. J., Gregory, R. D., McRae, L., Bubb, P., Roesler, I., Ursino, C., Wu, Y., Retief, E. F., Udin, J. S., Urazaliyev, R., Sánchez-Clavijo, L. M., Lartey, E., & Donald, P. F. (2022). A quantitative global review of species population monitoring. *Conservation Biology*, *36*(1), 1–14. https://doi.org/10.1111/cobi.13721

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

Pescott, O. L., Walker, K. J., Harris, F., New, H., Cheffings, C. M., Newton, N., Jitlal, M., Redhead, J., Smart, S. M., & Roy, D. B. (2019). The design, launch and assessment of a new volunteer-based plant monitoring scheme for the United Kingdom. *PLoS One*, *14*(4), 1–30. https://doi.org/10.1371/journal.pone.0215891

Pescott, O. L., Walker, K. J., Pocock, M. J. O., Jitlal, M., Outhwaite, C. L., Cheffings, C. M., Harris, F., & Roy, D. B. (2015). Ecological monitoring with citizen science: The design and implementation of schemes for recording plants in Britain and Ireland. *Biological Journal of the Linnean Society*, *115*(3), 505–521. https://doi.org/10.1111/bij.12581

Pollard, E. (1977). A method for assessing changes in the abundance of butterflies. *Biological Conservation*, *12*(2), 115–134. https://doi.org/10.1016/0006-3207(77)90065-9

Pollard, E., & Yates, T. J. (1996). *Monitoring butterflies for ecology and conservation*. Chapman & Hall.

Powney, G. D., Carvell, C., Edwards, M., Morris, R. K. A., Roy, H. E., Woodcock, B. A., & Isaac, N. J. B. (2019). Widespread losses of pollinating insects in Britain. *Nature Communications*, *10*(2019), 1–6. https://doi.org/10.1038/s41467-019-08974-9

Robbins, C., Bystrak, D., & Geissler, P. (1986). *The breeding bird survey: Its first fifteen years, 1965-1979*. https://pubs.usgs.gov/publication/5230189

Schuessler, J., & Selb, P. (2023). Graphical causal models for survey inference. *Sociological Methods & Research*, 1–32. https://doi.org/10.1177/00491241231176851

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

Stroh, P. A., Walker, K., Humphrey, T. A., Pescott, O. L., & Burkmar, R. (2023). *Plant atlas 2020: Mapping changes in the distribution of the British and Irish Flora*. Princeton University Press.

Textor, J., van der Zander, B., Gilthorpe, M. S., Liśkiewicz, M., & Ellison, G. T. (2016). Robust causal inference using directed acyclic graphs: The R package "dagitty". *International Journal of Epidemiology*, *45*(6), 1887–1894. https://doi.org/10.1093/ije/dyw341

Thoemmes, F., & Mohan, K. (2015). Graphical representation of missing data problems. *Structural Equation Modeling*, *22*(4), 631–642. https://doi.org/10.1080/10705511.2014.937378

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. https://doi.org/10.1080/00273171.2014.931799

Van Swaay, C. A. M., Plate, C. L., & Van Strien, A. J. (2002). Monitoring butterflies in the Netherlands: How to get unbiased indices. *Proceedings of the Section Experimental and Applied Entomology of the Netherlands Entomological Society*, *13*, 21–27.

Yoccoz, N. G., Nichols, J. D., & Boulinier, T. (2001). Monitoring of biological diversity in space and time. *Trends in Ecology & Evolution*, *16*(8), 446–453. https://doi.org/10.1016/S0169-5347(01)02205-4

SUPPORTING INFORMATION

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

Supporting Information S1. Expert feedback form (expert 2)

Supporting Information S2. Expert feedback form (expert 1)

Supporting Information S3. Figure S1. Causal diagram depicting causes and effects of sample inclusion R and a species' abundance Y. **Figure S2.** Estimates of change in the landscape-wide mean of Y relative to its mean in period 1 from the superpopulation model (spm) and the model with site and year effects (lm). All estimates are averaged over 200 bootstrap samples. **Figure S3.** Land cover in Great Britain in 1990 and 2015. Categories are 1 = woodland, 2 = arable, 3 = grassland, 4 = freshwater, 5 = built-up and 6 = "other" (e.g. saltwater and inland rock). **Table S1.** Details of the auxiliary variables used as covariates in the superpopulation models. Some of the variables in this table were not used, but we include details of the full set for completeness.

How to cite this article: Boyd, R. J., Botham, M., Dennis, E., Fox, R., Harrower, C., Middlebrook, I., Roy, D. B., & Pescott, O. L. (2025). Using causal diagrams and superpopulation models to correct geographic biases in biodiversity monitoring data. *Methods in Ecology and Evolution*, *00*, 1–13. https://doi.org/10.1111/2041-210X.14492