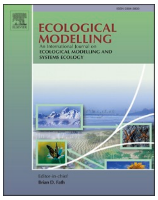
Ecological Modelling 493 (2024) 110739

Contents lists available at ScienceDirect

Ecological Modelling

journal homepage: www.elsevier.com/locate/ecolmodel



# On the trade-off between accuracy and spatial resolution when estimating species occupancy from geographically biased samples

Robin J. Boyd *, Diana E. Bowler, Nick J.B. Isaac, Oliver L. Pescott

*UK Centre for Ecology and Hydrology, Benson Ln., Wallingford, Oxfordshire, OX10 8BB, UK*

ARTICLE INFO

*Keywords:*
Sampling bias
Spatial grain
Representativeness
Biodiversity monitoring
Spatial pattern

ABSTRACT

Species occupancy is often defined as the proportion of areal units (sites) in a landscape that the focal species occupies, but it is usually estimated from the subset of sites that have been sampled. Assuming no measurement error, we show that three quantities–the degree of sampling bias (in terms of site selection), the proportion of sites that have been sampled and the variability of true occupancy across sites–determine the extent to which a sample-based estimate of occupancy differs from its true value across the wider landscape. That these are the only three quantities (measurement error notwithstanding) to affect the accuracy of estimates of species occupancy is the fundamental insight of the "Meng equation", an algebraic re-expression of statistical error. We use simulations to show how each of the three quantities vary with the spatial resolution of the analysis and that absolute estimation error is lower at coarser resolutions. Absolute error scales similarly with resolution regardless of the size and clustering of the virtual species' distribution. Finely resolved estimates of species occupancy have the potential to be more useful than coarse ones, but this potential is only realised if the estimates are at least reasonably accurate. Consequently, wherever there is the potential for sampling bias, there is a trade-off between spatial resolution and accuracy, and the Meng equation provides a theoretical framework in which analysts can consider the balance between the two. An obvious next step is to consider the implications of the Meng equation for estimating a time trend in species occupancy, where it is the confounding of error and true change that is of most interest.

## 1. Introduction

Species' range sizes are often measured in terms of occupancy, which is to say, the proportion of "sites" that they occupy within some landscape (Kéry and Royle, 2016; MacKenzie et al., 2002). Sites were originally conceived as discrete habitat patches or relatively small sampling units, but increasingly they represent contiguous larger-scale units defined by the analyst (e.g. squares on a map; Van Strien et al., 2013). This latter definition has often been used when estimating species occupancy at national and supranational scales (Boyd, August, et al., 2023; Coomber et al., 2021; Outhwaite et al., 2019; Powney et al., 2019).

In most circumstances—and particularly at fine scales across large areas—data are not available for all sites, so occupancy must be estimated from the subset of sites that have been sampled (Kéry and Royle, 2016). If the focal species is more or less likely to occupy sampled than non-sampled sites, then the sample is geographically biased (a formal definition is provided below), and the sample-based estimate will differ from its true value across the wider landscape (Boyd, Powney, et al., 2023; Meng, 2018). Geographic sampling biases are just one source of error when estimating species occupancy, the other major source being measurement error at sampled sites (MacKenzie et al., 2002).

A further complication when estimating species occupancy is that it varies with spatial resolution. Occupancy always increases as the resolution is coarsened, but the rate at which it increases depends on the size and clustering of the species' distribution at the finer scales (Azaele et al., 2012; Kunin, 1998; Wilson et al., 2004). All else being equal, fine scale estimates of species occupancy are preferable to coarse ones. For example, colonisations and local extinctions at small-scale sites are more probable than at larger scales, so working at a finer resolution means that occupancy is more sensitive to change (Dennis et al., 2019).

Although estimates of occupancy are nominally more useful at fine scales, there are reasons to work at coarser resolutions too. One reason is that, given finite resources, sampling at a fine scale might come at the expense of sampling over a large geographic area. Another is that the effects of sampling bias become more pronounced where there are more sites in the landscape (Boyd, Powney, et al., 2023; Meng, 2018a), which

---

* Corresponding author.
*E-mail address:* robboy@ceh.ac.uk (R.J. Boyd).

https://doi.org/10.1016/j.ecolmodel.2024.110739
Received 20 February 2024; Received in revised form 17 April 2024; Accepted 26 April 2024

is obviously the case at finer resolutions (i.e. where the sites are smaller). The fact that sampling biases are likely to be more pervasive at finer spatial resolutions raises questions about how the accuracy of estimates of species occupancy scales with resolution. Although working at coarser resolutions will clearly improve accuracy at the extremes—we can be surer a species occupies planet Earth than a set of small plots on its surface—how accuracy varies along the gradient from fine to coarse resolutions under sampling bias has not, to our knowledge, been investigated in ecology.

Here then, we investigate how the error of sample-based estimators of species occupancy vary with spatial resolution. Assuming no false absences (or that a model has adequately corrected them), we begin by demonstrating that three, and only three, quantities determine the magnitude of the error: the degree of sampling bias (in terms of site selection), the proportion of sites sampled and the variability of true occupancy across sites. That these are the only quantities affecting estimation error is a key implication of Meng's (2018) decomposition of survey error. We use simulations to show how each of the three quantities, and both relative and absolute error, vary with spatial resolution under sampling bias (at the finest resolution) and how varying the level of sampling bias affects the error. A trade-off emerges between finely resolved and accurate estimates, which we discuss in detail.

## 2. Methods

### 2.1. Quantifying estimation error

We consider a landscape comprising $N$ contiguous sites of equal area. The presence of at least one individual of the focal species is a binary variable $Y$ taking the value 1 at sites where it is present and 0 elsewhere. Occupancy $P(Y=1)$ is the proportion of sites at which the species is present, which is equivalent to the mean of $Y$ across sites $\overline{Y}$. Of the $N$ sites, a subset $n$ are sampled. Whether each site is one of the $n$ sampled sites is another binary variable $R$ ($R=1$ where the site is sampled and $R=0$ otherwise). It is not possible to calculate mean occupancy across all $N$ sites, $\overline{Y}_N$, because information on $Y$ is not available for sites with $R=0$. Instead, it is common to *estimate* $\overline{Y}_N$ as mean occupancy across sampled sites $\overline{Y}_n$.

Assuming no measurement errors, or that a model has corrected them, the absolute error of $\overline{Y}_n$ as an estimator of $\overline{Y}_N$ is (Meng, 2018)

$$\overline{Y}_n - \overline{Y}_N = \rho(R,Y)\sqrt{\frac{1-f}{f}}\,\sigma_Y. \tag{1}$$

The first quantity on the right, $\rho(R,Y)$, is the (population) correlation between $Y$ and $R$. It is a measure of both the sign and magnitude of *sampling bias*. In simple terms, $\rho(R,Y)$ is positive where $Y$ is generally larger in the sample than in the population (often the result of "preferential sampling"; Aubry et al., 2024) and vice versa. $f$ is the sampling rate $(n/N)$, and the second quantity on the right is a measure of *data quantity*. The final quantity $\sigma_Y$ is the population standard deviation of $Y$. It is 0 where $Y$ is constant, in which case a sample size of 1 is sufficient to estimate $\overline{Y}_N$ with no error, and it is largest where $Y$ is most variable. Hence, it can be considered a measure of "*problem difficulty*" (Meng, 2018), although we refer to it as occupancy variability given the context in which we are working.

Importantly, Eq. (1) gives the absolute error of $\overline{Y}_n$ as an estimator of $\overline{Y}_N$ for a given sample: that is, for one realisation of $R$. In what follows, we consider replicate realisations of $R$ from given $R$-generating (i.e. sampling) mechanisms and the average $\overline{Y}_n - \overline{Y}_N$ across those samples.

### 2.2. Effects of spatial resolution on error

Eq. (1) provides a basis for understanding the effects of resolution on absolute error when estimating species occupancy. Assuming perfect detection, it implies that there are three, and only three, ways to reduce error: decrease the sampling bias $\rho(R,Y)$, increase the sampling rate $f$ and/or decrease the occupancy variability $\sigma_Y$. Below we describe a set of simulations that demonstrate the effects of coarsening the spatial resolution on each of these quantities and on both absolute and relative error.

### 2.3. Simulation setup

#### 2.3.1. Virtual landscape, species and samples

The virtual landscape comprises a square grid of $N = 6400$ cells (80 × 80) at the finest resolution. Each cell might represent, say, a 1 × 1 km grid square, but the precise definition is not important for drawing general conclusions.

We simulated six species' geographic distributions of different sizes and with different levels of clustering in the virtual landscape. Our approach was a simplified version of the one used by Guélat and Kéry (2018). For each species, the first step was to populate every cell in the landscape with a continuous index $X$ sampled from a multivariate normal distribution

$$X \sim \mathcal{N}(\mu, \delta), \tag{2}$$

where $\mu$ is an $N$-vector of zeros (i.e. mean $X$ for each grid cell) and $\delta$ is an $N \times N$ covariance matrix. We used an exponential decay function to define the covariance matrix

$$\delta = e^{-\varphi\, D_{i,j}}, \tag{3}$$

where $\varphi$ is the decay constant and $D_{i,j}$ is the Euclidian distance between grid cells $i$ and $j$. Larger values of $\varphi$ result in patchier distributions, because the covariance between grid cells diminishes faster with the distance between them.

The next step was to convert the continuous index $X$ to a binary one (i.e. occupied vs unoccupied) with a specified proportion of cells being occupied. For each species, we set a threshold percentile of $X$ across grid cells $(1-\overline{Y}_N)$ above which the cell was designated occupied and below which it was designated unoccupied. Table 1 lists the parameters used to simulate each species' geographic distribution and the resulting properties of those distributions.

It was important that the simulated species' distributions spanned a range of plausible sizes and levels of clustering, because these properties determine how $\overline{Y}_N$ scales with resolution (Kunin, 1998). We tested whether the distributions covered sufficiently wide ranges of these parameters using their fractal (Kunin, 1998). The fractal dimension $D$ of a

**Table 1**
Properties of the six virtual species' distributions at the finest spatial resolution. The autocorrelation parameter is the exponential decay constant in Eq. (3), and higher values produce a more dispersed distribution. The theoretical limits for the fractal dimension are 0, representing a highly dispersed species, and 2, representing a very clustered one. The fractal dimension also varies with $\overline{Y}_N$ (Wilson et al., 2004).

| Distribution properties | Exponential decay parameter in autocorrelation function | Proportion of sites occupied (at the finest scale) | Fractal dimension |
|---|---|---|---|
| Rare and sparse | 0.6 | 0.01 | 0.19 |
| Rare and clustered | 0.1 | 0.01 | 0.88 |
| Medium and sparse | 0.6 | 0.25 | 1.19 |
| Medium and clustered | 0.1 | 0.25 | 1.42 |
| Common and sparse | 0.6 | 0.5 | 1.58 |
| Common and clustered | 0.1 | 0.5 | 1.71 |

species' distribution is given by $D = 2(1 - b)$, where $b$ is the slope of its scale-area curve or occupancy-area relationship (Hartley and Kunin, 2003). We calculated $b$ over the finest three resolutions, because, for the medium and common species, including the coarsest two resolutions resulted in nonlinear scale-area curves (i.e. their distributions are non-fractal at coarse scales). The theoretical limits of the fractal dimension are 0, representing a species whose distribution is very sparse, and 2, representing a species whose distribution is very clustered (Hartley and Kunin, 2003). Our virtual species' distributions spanned most of this range (0.19−1.71). Like Wilson et al. (2004), we found that $D$ is positively related to $\overline{Y}_N$, which reflects the facts that a small distribution can only be so clustered, and a large distribution can only be so dispersed. See Fig. S1 for maps of the virtual species' distributions.

For each species, we simulated 100 virtual samples at the finest resolution. Whilst it might seem more logical to simulate one set of samples for all species, this would not allow control over $\rho(R, Y)$, the sampling bias, which depends on the focal species' geographic distribution. For most simulations, we simulated the samples in such a way that $E_R[\rho(R, Y)] \sim 0.05$ and $f = 0.1$, where $E_R[\rho(R, Y)]$ is the expectation (average) of $\rho(R, Y)$ over the 100 simulated samples (i.e. with respect to $R$). See the supplementary Fig. S2 for the distributions of $\rho(R, Y)$ across samples for each species. We based the values of $\rho(R, Y)$ and $f$ on an empirical example: a citizen science dataset on vascular plant sampling and the species *Calluna vulgaris'* occupancy in Britain (Boyd, Powney, et al., 2023). Whilst we generally set $E_R[\rho(R, Y)] \sim 0.05$ and $= 0.1$, we also demonstrate the effects of varying both parameters (in the supplementary material for $f$). Switching the sign of $\rho(R, Y)$ (i.e. whether occupancy is larger or smaller in the sample than the population) would switch the sign of the error in the estimate of mean occupancy, but for simplicity we only present the positive case.

Our algorithm for generating samples with the prescribed $\rho(R, Y)$ uses "brute force". It starts by creating a random sample with the desired $f$ then iteratively seeks the target value of $\rho(R, Y)$ by flipping values of $R$ at selected sites (i.e. from *sampled* $= 1$ to *not sampled* $= 0$ or vice versa). It stops once the values of $\rho(R, Y)$ and $f$ are within small tolerance limits of the target values. We are not aware of an analytical approach to generating correlated binary vectors ($R$ and $Y$) with fixed proportions of 1 s.

### 2.4. Analysis of error at each resolution

The goal of our analysis was to determine how the absolute error of $\overline{Y}_n$ as an estimator of $\overline{Y}_N$ ($\overline{Y}_n - \overline{Y}_N$; assuming perfect detection) varies with spatial resolution. Starting at the finest resolution, we calculated the value of each quantity in Eq. (1) (including the absolute error; averaged across the 100 samples). We then coarsened the resolution by aggregating every square of four grid cells into one (i.e. doubling the length and width of the site). After coarsening the resolution, we recalculated each quantity in Eq. (1), coarsened the resolution again and repeated the process until each grid cell was 16× its original height and width (see Fig. S5 for the results of additional coarsening on a larger grid). Fig. 1 shows how a species' distribution (medium and clustered; Table 1) and a sample vary with resolution.

## 3. Results

### 3.1. Error

For all virtual species, estimates of occupancy are more accurate at coarser resolutions. This result is evident both in terms of the absolute actual error (Fig. 2A), which is on the left side of Eq. (1), and the relative actual error (Fig. 2B), which expresses the absolute error as a percentage of true occupancy. Relative error is larger for rare species. Absolute error is larger for the medium and common species, particularly at the finer resolutions. There is little difference in absolute or relative error between clustered and dispersed species.

### 3.2. True occupancy

Although well-documented (Azaele et al., 2012; Kunin, 1998), it is worth revisiting the scaling properties of $\overline{Y}_N$ (true occupancy) here, because they provide insight into the scaling properties of error. $\overline{Y}_N$ always increases with resolution, but the rate at which it increases depends on the properties of the species' distribution at the finest resolution (Fig. 2C). Species that are common and sparsely distributed at the finest resolution quickly reach $\overline{Y}_N = 1$ as the resolution is coarsened. By contrast, species that are rare and clustered at the finest resolution do not reach $\overline{Y}_N = 1$ at any of the resolutions we considered (Fig. 2A).

### 3.3. Sampling bias

In our simulations, the sampling bias $\rho(R, Y)$ tends towards 0 as the resolution is coarsened (Fig. 2D). There are plausible scenarios in which it will not (e.g. when the samples are highly clustered), however, a point that we expand on in the Discussion.

### 3.4. Sampling rate

Like $\overline{Y}_N$, the sampling rate always increases with resolution.

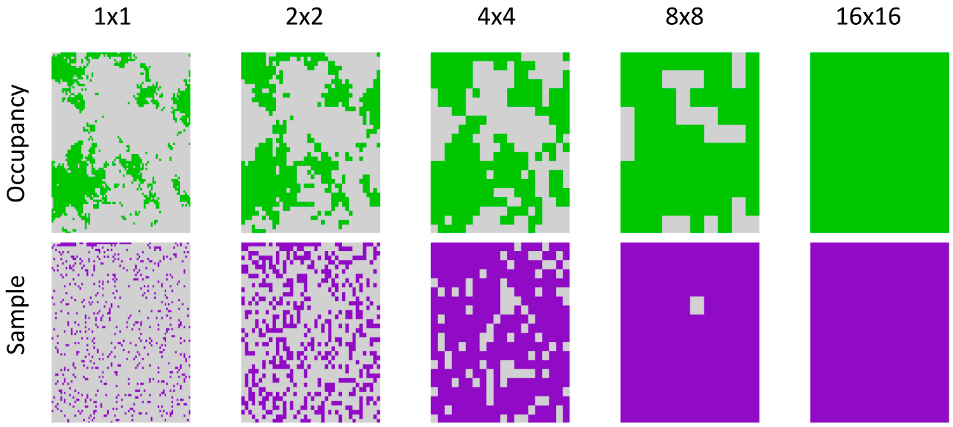


**Fig. 1.** Top row: a virtual species' ("medium and clustered"; Table 1) geographic distribution at each spatial resolution. Green cells are occupied, and grey cells are not. Bottom row: a virtual sample at each resolution. $\rho(R, Y) \sim 0.05$ and $f \sim 0.1$ at the finest resolution ($1 \times 1$). Purple cells are sampled, and grey cells are not. Sampled cells may be either occupied or not.

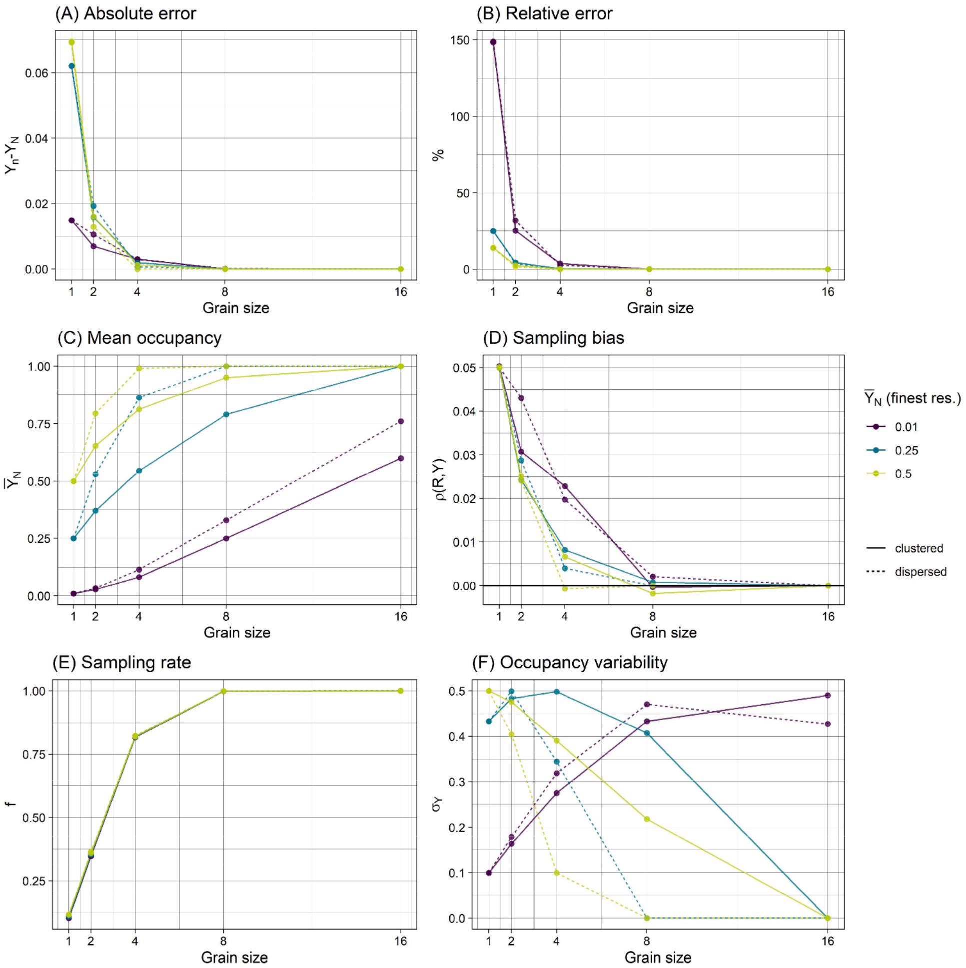


**Fig. 2.** (A) absolute error, (B) relative error (i.e. the absolute error expressed as a percentage of true occupancy), (C) mean occupancy (i.e. true occupancy), (D) sampling bias, (E) sampling rate and (F) occupancy variability $\sigma_Y$ at each resolution. The resolution is the height and width of the grid cells in arbitrary units. Points represent the average of each statistic over 100 simulated samples. At the finest resolution, $\rho(R,Y) \sim 0.05$ and $f \sim 0.1$, the target values for the simulations.

### 3.5. *Occupancy variability*

As occupancy is binary, its standard deviation $\sigma_Y$ is given by $\sqrt{\overline{Y}_N(1-\overline{Y}_N)}$. $\sigma_Y$ is largest where $\overline{Y}_N$ is near 0.5 and smallest where $\overline{Y}_N$ is near 0 or 1. Given that $\overline{Y}_N$ increases with resolution (Fig. 2C), coarsening the resolution for species with $\overline{Y}_N < 0.5$ increases $\sigma_Y$ until $\overline{Y}_N = 0.5$ (Fig. 2F). Further coarsening the resolution decreases $\sigma_Y$, because $\overline{Y}_N$ moves away from 0.5 and towards 1. For species with $\overline{Y}_N \geq 0.5$ at the finest resolution, coarsening the resolution always decreases $\sigma_Y$.

### 3.6. *Scaling of error with resolution at different levels of sampling bias*

In most simulations, we set $\rho(R,Y) \sim 0.05$ at the finest resolution, but it is instructive to see how actual error scales with resolution under different levels of sampling bias. Absolute error generally scales in the same way with resolution regardless of the level of sampling bias but is greater in magnitude under stronger sampling bias (Fig. 3; the same is true of relative error [Fig. S3]). Under a simple random sample at the finest resolution, where the expected sampling bias $E_R[\rho(R,Y)] \sim 0$, there is roughly no error at any resolution (recalling that we present the average error across samples, which essentially removes sampling error). Note that we were not able to simulate highly biased samples ($E_R[\rho(R,Y)] \sim 0.15$) for the common species (green lines in Fig. 3). For these species, $\overline{Y}_N$ is very different to $f$, which makes a large and positive $\rho(R,Y)$ highly unlikely, and our algorithm for generating the samples could not achieve it (two binary variables with very different proportions of 1 s can only be so positively correlated).

## 4. Discussion

Nobody would dispute the fact that estimates of species occupancy are more accurate at coarse scales asymptotically: we can be surer that a species occupies Britain than it does some 1 km grid square therein. Our contribution has been to show that accuracy varies somewhat predictably along the spectrum from fine to coarse resolutions. Indeed, Meng's (2018) three-part decomposition of statistical error provides a clear theoretical framework within which analysts can consider quantities like the potential sampling bias and the sampling rate when deciding on the appropriate resolution at which to estimate occupancy. Coarsening the resolution may be particularly beneficial where sampling biases are likely to be large (e.g. when using citizen science data; Pescott et al., 2019; Stroh et al., 2023a).

The Meng (2018) equation tells us that to increase the accuracy of estimates of species occupancy, we should work at the spatial resolution at which the sampling bias and the variability of occupancy in the landscape are smallest and at which the sampling rate is highest. Maximising the sampling rate is simplest in theory, because it always increases with resolution (practice of course introduces issues of resourcing and planning). The effect of resolution on the variability of occupancy in the landscape depends on the species' prevalence (i.e. $\overline{Y}_N$) at the finest resolution. If there is good reason to think that

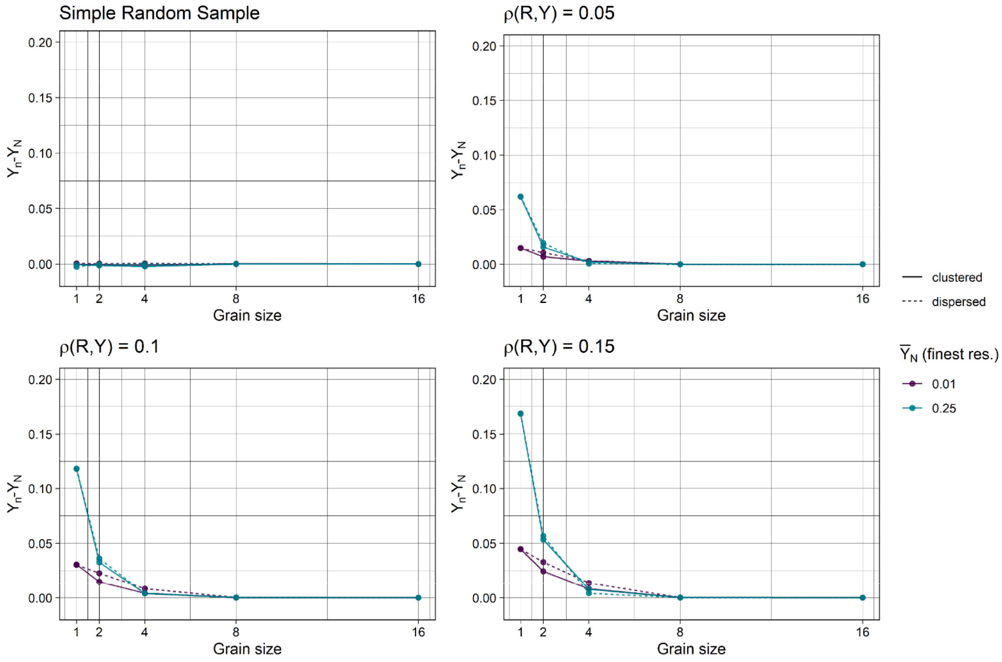


**Fig. 3.** Absolute error at each resolution under four levels of sampling bias $\rho(R, Y)$ (at the finest resolution). The resolution is the height and width of the grid cells in arbitrary units. The simple random sample has approximately no sampling bias at the finest resolution. Points represent the average of each statistic over 100 simulated samples. $f \sim 0.1$ at the finest resolution in all cases.

$\overline{Y}_N \geq 0.5$—say, from an expert drawn range map—then coarsening the resolution will always reduce $\sigma_Y$. On the other hand, if there is good reason to think that $\overline{Y}_N$ is truly low, then coarsening the resolution will increase $\sigma_Y$ until the $\overline{Y}_N$ reaches 0.5.

In our simulations, sampling bias was clearly lower at coarser resolutions (Fig. 2D), but this will not be universally true. One minor thing to note is that we presented the average $\rho(R, Y)$ across 100 samples: for some of the individual samples, $\rho(R, Y)$ occasionally increased from one resolution to the next. More importantly, our algorithm for creating samples starts by simulating random samples then adjusts them to reach the desired $\rho(R, Y)$. Starting with a random sample makes it unlikely that the final sample will be highly clustered (Fig. 1). One might expect real samples to be clustered in accessible or attractive areas: say, near where people live or in nature reserves (Tulloch et al., 2013). It is not clear whether simulating clustered samples would alter our finding that $\rho(R, Y)$ decreases with spatial resolution; more detailed simulations that account for drivers of species' occupancy and sampling would be needed to answer this question.

As it is often time trends in species occupancy, rather than one-off estimates, that are of interest, it is worth considering estimation error in this context. It is generally understood that time-varying sampling bias (and therefore error) can confound true change in occupancy (Bowler et al., 2022), but knowing how sampling bias changes over time is made difficult by the various sampling schemes and analytical approaches that might be employed by researchers. The simplest scenario is where the analyst estimates occupancy separately for multiple time-periods and calculates the differences between them. If the sampling bias changes over time, then the estimated differences will be erroneous. Another way to estimate time trends in occupancy is to restrict the analysis to the pool of sites that were sampled at some point within the relevant timeframe and to predict (or impute) missing values in each time-period (Boyd, August, et al., 2023; Isaac et al., 2014). Putting to one side the fact that there are almost certain to be prediction errors, one ends up in a situation where the distribution of $R$ across sites is effectively time-invariant. Crucially, however, this does not mean that the sampling bias will remain constant over time unless the distribution of $Y$ across sites is also time-invariant (i.e. the species' distribution does not change over time at the relevant scale). A similar scenario arises when occupancy is estimated using unrepresentative monitoring data whose geographic distribution does not change over time: for example, long-term monitoring of protected sites.

Understanding how the potential for confounding of error and true temporal change in occupancy varies with spatial resolution is difficult, but the Meng equation provides several insights here too. For example, working at coarser resolutions means less temporal variation in $\overline{Y}_N$ (as colonisations and local extinctions are less probable), which means less temporal variation in $\sigma_Y$. It is also likely to mean less variation in $\rho(R, Y)$—especially if occupancy is predicted across a fixed pool of sites in each year, in which case the distribution of $R$ is effectively constant over time (again, one must also consider the fact that the predictions could be wrong at unsampled site/time-period combinations). Reducing temporal variation in the quantities in Eq. (1) will reduce temporal variation in error, which should reduce the potential for confounding of error and true change in occupancy in many cases. An obvious exception is where the per-period errors cancel each other out over long timeframes, in which case they will not bias the estimated trend; however, it is not likely that biodiversity monitors will know that they are in this situation—if the per period direction of error was known, then it could be modelled. More elaborate simulations and theoretical work are needed to fully understand the effects of spatial scale on error when estimating time trends in species occupancy.

Although not the focus of this paper, the Meng equation could also shed light on how accuracy scales with spatial resolution when estimating mean *abundance*. Whether measuring occupancy or abundance, $f$ always increases as the resolution is coarsened. For abundance, which is numeric, $\sigma_Y = \sqrt{1/N\sum(y_i - \overline{Y}_N)^2}$, where $i$ indexes the site. Consequently, $\sigma_Y$ is likely to be smaller at coarser resolutions, because aggregating multiple cells into one should smooth over local variations in abundance. Smoothing over local variation in abundance by coarsening the spatial resolution might also reduce $\rho(R, Y)$ if it means that differences in $Y$ between sampled and non-sampled sites become smaller. One might speculate then, that estimates of mean abundance are likely to be more accurate at coarser resolutions, and it would be useful to test this assertion more thoroughly (noting that measurement error is likely to more prevalent when measuring species abundance than occupancy—especially at coarser resolutions).

The fact that error in estimates of species occupancy is likely to be lower at coarser spatial resolutions sets up a trade-off between accuracy and "usefulness". Estimates of species occupancy clearly have the potential to be more useful at fine scales. For example, working at a finer resolution, at which local extinctions and colonisations are more probable, means having a greater power to detect change. (Of course, this argument supposes that the estimates are accurate or at least consistently inaccurate over time. It also supposes that the power to detect change at some significance level is of primary interest, which is not always true.) Working at coarse resolutions also means that results are potentially (i) less relevant to policy (Spake et al., 2022) and (ii) less biologically meaningful (e.g. if the site is much larger than the species' home range size; Altwegg and Nichols, 2019). When deciding on the appropriate resolution at which to analyse their data, analysts must balance the need for accurate and useful estimates and remember that an estimate will not be useful if it is completely wrong.

A good example of the potential for bias being balanced against the desire for finely resolved estimates of species occupancy is found in the latest plant atlas of the Botanical Society of Britain and Ireland (Stroh et al., 2023). The data were analysed at a 10 × 10 km scale—much coarser than the 1 × 1 km resolution used by others in the area (Boyd, August, et al., 2023)—and some time-periods were omitted, due to serious concerns about sampling biases affecting species data at finer scales across the 20th century. For example, both rarer and more challenging to identify taxa were more likely to be reported at finer scales in the early part of the time series. Moreover, $f$ was known to be far smaller at smaller scales in these earlier periods (Pescott et al., 2019).

Like all simulations, ours are a simplification of reality, which might have implications for the wider applicability of our results. We did not account for the fact that additional data tend to be available at coarser resolutions; for example, digitised specimens may be resolved only to some vague locality, and historic distribution data from species' Atlases tend to be more coarsely resolved than contemporary data (Groom et al., 2018; Kunin et al., 2000; Pescott et al., 2019). These additional data would increase the sampling rate $f$ at coarse resolutions, which, as we have shown, would be likely to increase the accuracy of sample-based estimates of mean occupancy. [Note that it is possible to combine fine and coarse data using integrated distribution models and to draw inferences at the finer scale (Pacifici et al., 2019). Whether the fact that data might be available solely at coarse scales for historic time-periods, and at multiple scales for recent ones, will impact inference is an open question.] Our assumption of perfect detection (i.e. no false absences) is also unrealistic, so it is worth considering whether the prevalence of false absences is likely to be lower at fine or coarse resolutions. On the one hand, if a coarse resolution is chosen when planning data collection, false absences might be higher if the portions of the larger cells that are sampled are not suitable for the focal species (Altwegg and Nichols, 2019). On the other, if the resolution is chosen at the analysis stage, coarsening the spatial resolution increases the number of sampling events per site, so, all else being equal, it is more likely that the focal species will be detected if it is present.

Rather than accepting false absences, it is common practice to try to correct them using some sort of occupancy-detection model (MacKenzie et al., 2002; Royle, 2006). Coarsening the resolution of the analysis risks violating the closure assumption of occupancy-detection models (Altwegg and Nichols, 2019; Jönsson et al., 2021), but it also increases the amount of repeat visits to the same site, which are needed to estimate detectability and correct false absences. Interesting possibilities are that multi-scale occupancy models (Mordecai et al., 2011), which relax the closure assumption, could be used and that fine-scale sampling events could be used as spatial replicates to estimate detection probabilities and correct false absences at coarser scales (Srivathsa et al., 2018). While failing to correct false absences can make estimates of species occupancy worse, it is important to remember that successfully correcting them only reduces error to its baseline level determined by sampling biases (Meng, 2018).

Coarsening the resolution of an analysis is one approach to counter some of the error introduced by sampling biases, but there are alternatives. One is to estimate mean occupancy in the population using a *weighted* sample mean, where the weights are equal to the inverse of the (possibly estimated) sample inclusion probabilities (Boyd, Stewart, et al., 2023; Johnston et al., 2020). If successful, weighting of this type brings the distribution of occupancy in the sample closer to its distribution in the population and can be recast as a means to minimising $\rho(R, Y)$ (Meng, 2022). Several approaches to estimating sampling weights for unstructured (i.e. nonprobability) samples, the principal type of data used to estimate species occupancy, exist (Boyd, Stewart, et al., 2023; Elliott and Valliant, 2017). Weighting is often more successful where available covariates explain larger portions of the variance in sample inclusion (i.e. $R$) and the variable of interest (occupancy; Collins et al., 2001), and it would be useful to investigate how this scales with spatial resolution.

## 5. Conclusions

Analysts consider several factors when deciding on the appropriate resolution at which to estimate species occupancy. Examples include the focal species' home range sizes (Wilson and Schmidt, 2015), the scale at which they use the landscape more generally (Powney et al., 2019), the number of replicate visits to the same site within closure periods (Outhwaite et al., 2019) and the resolution at which the data were collected (Higa et al., 2015). We propose that analysts should also consider the fact that estimates are likely to be more accurate at coarse resolutions, because a highly erroneous finer-scale estimate is unlikely to be useful for most applications. The Meng (2018) equation provides a theoretical framework in which accuracy and the desire for finely resolved information can be balanced.

## CRediT authorship contribution statement

**Robin J. Boyd:** Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Diana E. Bowler:** Writing – review & editing, Investigation. **Nick J.B. Isaac:** Writing – review & editing, Supervision, Investigation. **Oliver L. Pescott:** Writing – review & editing, Supervision, Conceptualization.

## Declaration of competing interest

The authors have no interests to disclose.

## Data availability

All R code needed to fully reproduce our analysis, as well as the session information, is available on Zenodo (https://doi.org/10.5281/zenodo.11078096).

## 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". All authors were supported by the NERC award number NE/R016429/1 as part of the UK-SCAPE programme delivering National Capability.

## Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ecolmodel.2024.110739.

## References

Altwegg, R., Nichols, J.D., 2019. Occupancy models for citizen-science data. Methods Ecol. Evol. 10 (1), 8–21. https://doi.org/10.1111/2041-210X.13090.

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

Azaele, S., Cornell, S.J., Kunin, W.E., 2012. Downscaling species occupancy from coarse spatial scales. Ecol. Appl. 22 (3), 1004–1014. https://doi.org/10.1890/11-0536.1.

Bowler, D.E., Callaghan, C.T., Bhandari, N., Henle, K., Barth, M.B., Koppitz, C., Klenke, R., Winter, M., Jansen, F., Bruelheide, H., Bonn, A., 2022. Temporal trends in the spatial bias of species occurrence records. Ecography. https://doi.org/10.1111/ecog.06219.

Boyd, R.J., August, T., Cooke, R., Logie, M., Mancini, F., Powney, G., Roy, D., Turvey, K., Isaac, N., 2023a. An operational workflow for producing periodic estimates of species occupancy at large scales. Biol. Rev. 9 https://doi.org/10.32942/OSF.IO/2V7JP.

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

Boyd, R.J., Stewart, G.B., Pescott, O.L., 2023c. Descriptive inference using large, unrepresentative nonprobability samples: an introduction for ecologists. Ecology. https://doi.org/10.1002/ecy.4214.

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

Coomber, F.G., Smith, B.R., August, T.A., Harrower, C.A., Powney, G.D., Mathews, F., 2021. Using biological records to infer long-term occupancy trends of mammals in the UK. Biol. Conserv. 264 (February), 109362 https://doi.org/10.1016/j.biocon.2021.109362.

Dennis, E.B., Brereton, T.M., Morgan, B.J.T., Fox, R., Shortall, C.R., Prescott, T., Foster, S., 2019. Trends and indicators for quantifying moth abundance and occupancy in Scotland. J. Insect Conserv. 23 (2), 369–380. https://doi.org/10.1007/s10841-019-00135-z.

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

Groom, Q.J., Marsh, C.J., Gavish, Y., Kunin, W.E., 2018. How to predict fine resolution occupancy from coarse occupancy data. Methods Ecol. Evol. 9 (11), 2273–2284. https://doi.org/10.1111/2041-210X.13078.

Guélat, J., Kéry, M., 2018. Effects of spatial autocorrelation and imperfect detection on species distribution models. Methods Ecol. Evol. 9 (6), 1614–1625. https://doi.org/10.1111/2041-210X.12983.

Hartley, S., Kunin, W.E., 2003. Scale dependency of rarity, extinction risk, and conservation priority. Conserv. Biol. 17 (6), 1559–1570. https://doi.org/10.1111/j.1523-1739.2003.00015.x.

Higa, M., Yamaura, Y., Koizumi, I., Yabuhara, Y., Senzaki, M., Ono, S., 2015. Mapping large-scale bird distributions using occupancy models and citizen data with spatially biased sampling effort. Divers. Distrib. 21 (1), 46–54. https://doi.org/10.1111/ddi.12255.

Isaac, N.J.B., van Strien, A.J., August, T.A., de Zeeuw, M.P., Roy, D.B., 2014. Statistics for citizen science: extracting signals of change from noisy ecological data. Methods Ecol. Evol. 5 (10), 1052–1060. https://doi.org/10.1111/2041-210X.12254.

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

Jönsson, G.M., Broad, G.R., & Umner, S.S. (2021). *A century of social wasp occupancy trends from natural history collections : spatiotemporal resolutions have little effect on model performance*. 14(5), 543–555. https://doi.org/10.1111/icad.12494.

Kéry, M., Royle, J.A., 2016. Applied Hierarchical Modelling in ecology: Analysis of Species distribution, Abundance and Species Richness in R and BUGS. Academic press.

Kunin, W.E., 1998. Extrapolating species abundance across spatial scales. Science 281 (5382), 1513–1515. https://doi.org/10.1126/science.281.5382.1513.

Kunin, W.E., Hartley, S., Lennon, J.J., 2000. Scaling down: on the challenge of estimating abundance from occurrence patterns. Am. Natur. 156 (5), 560–566. https://doi.org/10.1086/303408.

MacKenzie, D.I., Nichols, J.D., Lachman, G.B., Droege, S., Royle, A.A., Langtimm, C.A., 2002. Estimating site occupancy rates when detection probabilities are less than one. Ecology 83 (8), 2248–2255. https://doi.org/10.1890/0012-9658(2002)083[2248:ESORWD]2.0.CO;2.

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 (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. Surv. Methodol. 48 (2), 1–22.

Mordecai, R.S., Mattsson, B.J., Tzilkowski, C.J., Cooper, R.J., 2011. Addressing challenges when studying mobile or episodic species: hierarchical Bayes estimation of occupancy and use. J. Appl. Ecol. 48 (1), 56–66. https://doi.org/10.1111/j.1365-2664.2010.01921.x.

Outhwaite, C., Powney, G., August, T., Chandler, R., Rorke, S., Pescott, O.L., Harvey, M., Roy, H.E., Fox, R., Roy, D.B., Alexander, K., Ball, S., Bantock, T., Barber, T., Beckmann, B.C., Cook, T., Flanagan, J., Fowles, A., Hammond, P., Isaac, N.J.B., 2019. Annual estimates of occupancy for bryophytes, lichens and invertebrates in the UK, 1970-2015. Sci. Data 6 (1), 259. https://doi.org/10.1038/s41597-019-0269-1.

Pacifici, K., Reich, B.J., Miller, D.A.W., Pease, B.S., 2019. Resolving misaligned spatial data with integrated species distribution models. Ecology 100 (6), 1–15. https://doi.org/10.1002/ecy.2709.

Pescott, O.L., Humphrey, T.A., Stroh, P.A., Walker, K.J., 2019. Temporal changes in distributions and the species atlas: how can British and Irish plant data shoulder the inferential burden? Br. Irish Botany 1 (4), 250–282. https://doi.org/10.33928/bib.2019.01.250.

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. Nat. Commun. 10 (2019), 1–6. https://doi.org/10.1038/s41467-019-08974-9.

Royle, J.A., 2006. Site occupancy models with heterogeneous detection probabilities. Biometrics 62 (1), 97–102. https://doi.org/10.1111/j.1541-0420.2005.00439.x.

Spake, R., Barajas-Barbosa, M.P., Blowes, S.A., Bowler, D.E., Callaghan, C.T., Garbowski, M., Jurburg, S.D., Van Klink, R., Korell, L., Ladouceur, E., Rozzi, R., Viana, D.S., Xu, W.B., Chase, J.M., 2022. Detecting thresholds of ecological change in the anthropocene. Annu. Rev. Environ. Resour. 47, 797–821. https://doi.org/10.1146/annurev-environ-112420-015910.

Srivathsa, A., Puri, M., Kumar, N.S., Jathanna, D., Karanth, K.U., 2018. Substituting space for time: empirical evaluation of spatial replication as a surrogate for temporal replication in occupancy modelling. J. Appl. Ecol. 55 (2), 754–765. https://doi.org/10.1111/1365-2664.13005.

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 Univ. Press.

Tulloch, A.I.T., Mustin, K., Possingham, H.P., Szabo, J.K., Wilson, K.A., 2013. To boldly go where no volunteer has gone before: predicting volunteer activity to prioritize surveys at the landscape scale. Divers. Distrib. 19 (4), 465–480. https://doi.org/10.1111/j.1472-4642.2012.00947.x.

Van Strien, A.J., Van Swaay, C.A.M., Termaat, T., 2013. Opportunistic citizen science data of animal species produce reliable estimates of distribution trends if analysed with occupancy models. J. Appl. Ecol. 50 (6), 1450–1458. https://doi.org/10.1111/1365-2664.12158.

Wilson, R.J., Thomas, C.D., Fox, R., Roy, D.B., Kunin, W.E., 2004. Spatial patterns in species distributions reveal biodiversity change. Nature 432 (7015), 393–396. https://doi.org/10.1038/nature03031.

Wilson, T., Schmidt, J.H., 2015. Scale dependence in occupancy models: implications for estimating bear den distribution and abundance. Ecosphere 6 (9). https://doi.org/10.1890/ES15-00250.1.