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

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

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

*corresponding author email: robboy@ceh.ac.uk

Orcid ID's: OLP- https://orcid.org/0000-0002-0685-8046

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 as the proportion of *sampled* sites in which the species has been observed. Assuming perfect detection (i.e. no false absences), 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 the truth. That these are the only three quantities 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 actual estimation error is lower at coarser resolutions. Although finely resolved estimates of species occupancy have the potential to be more useful than coarse ones, 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.

Key words: sampling bias; spatial grain; representativeness; biodiversity monitoring

Introduction

Species occupancy, which we define as the proportion of areal units (sites) in some defined landscape occupied by the focal species, is often of interest to ecologists (Kéry & Royle, 2016). It is used to quantify species' range dynamics (Dennis et al., 2019; Outhwaite et al., 2020; Powney et al., 2019; Stroh et al., 2023), identify correlates and drivers of those range dynamics (Cooke et al., 2023; Woodcock et al., 2016), track the spread of invasive species and their effects on native taxa (Roy et al., 2012) and monitor progress towards (inter-) national biodiversity targets (Boyd, August, et al., 2023). Clearly, information on species occupancy has the potential to be useful, but realising this potential is conditional on available data being an accurate reflection of reality.

A major source of inaccuracy when estimating species occupancy is geographic sampling bias. In most circumstances—and particularly at fine scales across large areas—it is not possible to sample all sites, so occupancy must be estimated from the subset of sites that have been sampled (Kéry & Royle, 2016). If occupancy differs between sampled and non-sampled sites, then the sample is not representative, and the sample-based estimate of species occupancy will differ from its true value in the wider landscape (Boyd, Powney, et al., 2023; Meng, 2018). Sampling biases are just one of many sources of error when estimating species occupancy (e.g. Isaac et al., 2014; MacKenzie et al., 2002).

Further complicating estimation of 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 fine-scale properties of the species' geographic distribution (Azaele et al., 2012; Kunin, 1998; Wilson et al., 2004). Occupancy is a better surrogate for abundance, which is often of primary interest, at fine resolutions (Kunin, 1998). Indeed, where the scale of analysis is roughly the size of an individual, occupancy and abundance are equivalent. A species' abundance is more variable than its occupancy (e.g. Dennis et al., 2019), since local occupancy does not decline until local abundance reaches zero and cannot increase once it is above zero. Consequently, working at finer scales, where occupancy is a better surrogate for abundance, means having a greater power to detect change.

Although estimates of occupancy are nominally more useful at fine scales, there are reasons to work at coarser resolutions too. One reason is that resourcing constraints might preclude the additional sampling effort required to estimate occupancy at fine resolutions. Another is that the effects of sampling biases become more pronounced where there are more sites in the landscape (Boyd, Powney, et al., 2023; Meng, 2018), which is obviously the case at finer resolutions (i.e. where the sites are smaller). The fact that sampling biases are 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 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. Analysts should consider our results when deciding on the most appropriate resolution at which to estimate species occupancy.

Methods

Quantifying estimation error

We consider a landscape comprising N sites. 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 $\bar{Y}$. 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, $\bar{Y}_N$, because information is not available on sites with $R = 0$. Instead, it is common to *estimate* $\bar{Y}_N$ as mean occupancy across sampled sites $\bar{Y}_n$.

Assuming no measurement error (e.g. false absences), the actual error of $\bar{Y}_n$ as an estimator of $\bar{Y}_N$ is (Meng, 2018)

$$\bar{Y}_n - \bar{Y}_N = \rho(R, Y)\sqrt{\frac{1-f}{f}}\,\sigma_Y. \qquad \text{equation 1}$$

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 negative where Y is generally smaller in the sample than in the population 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 σ_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 $\bar{Y}_N$ 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 actual error of $\bar{Y}_n$ as an estimator of $\bar{Y}_N$ 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 $\bar{Y}_n - \bar{Y}_N$ across those samples.

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 σ_Y. Below we describe a set of simulations that demonstrate the effects of coarsening the spatial resolution on each of these quantities and on error.

Simulation setup

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 & Kéry (2018). For each species, the first step was to populate every cell in the landscape with an index X sampled from a multivariate normal distribution

$$X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\delta}), \qquad \text{equation 2}$$

where $\boldsymbol{\mu}$ is an N-vector of 0's (i.e. mean X for each grid cell) and $\boldsymbol{\delta}$ is an $N \times N$ covariance matrix. We used an exponential decay function to define the covariance matrix

$$\boldsymbol{\delta} = e^{-\varphi\, \boldsymbol{D_{i,j}}}, \qquad \text{equation 3}$$

where φ is the decay constant and $\boldsymbol{D_{i,j}}$ is the Euclidian distance between grid cells i and j. Larger values of φ 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 $(1 - \bar{Y}_N)$ 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 $\bar{Y}_N$ scales with resolution (Kunin, 1998). We tested whether the distributions covered sufficiently wide ranges of these parameters using their fractal dimensions (Kunin, 1998). The fractal dimension D of a species' distribution is given by $D = 2(1 - b)$, where b is the slope of its scale-area curve (i.e. a plot of the logarithm of range size against the logarithm of the area of each grid cell; Hartley & 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 & Kunin, 2003). Our virtual species' distributions spanned the majority of this range (0.31−1.64). Note that D is positively related to $\bar{Y}_N$ (Wilson et al., 2004).

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 $\bar{Y}_N$ (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	0.31
Rare and clustered	0.1	0.01	0.87
Medium and sparse	0.6	0.25	1.20
Medium and clustered	0.1	0.25	1.39
Common and sparse	0.6	0.5	1.57
Common and clustered	0.1	0.5	1.64

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. S1 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 et al., 2023). Whilst we generally set $E_R[\rho(R, Y)] \sim 0.05$ and $f = 0.1$, 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.

Analysis of error at each resolution

The goal of our analysis was to determine how the actual error of $\bar{Y}_n$ as an estimator of $\bar{Y}_N$ ($\bar{Y}_n - \bar{Y}_N$; assuming perfect detection) varies with spatial resolution. Starting at the finest resolution, we calculated the value of each quantity in eq. 1 (including the actual 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. Fig. 1 shows how a species' distribution (medium and clustered; Table 1) and a sample vary with resolution.

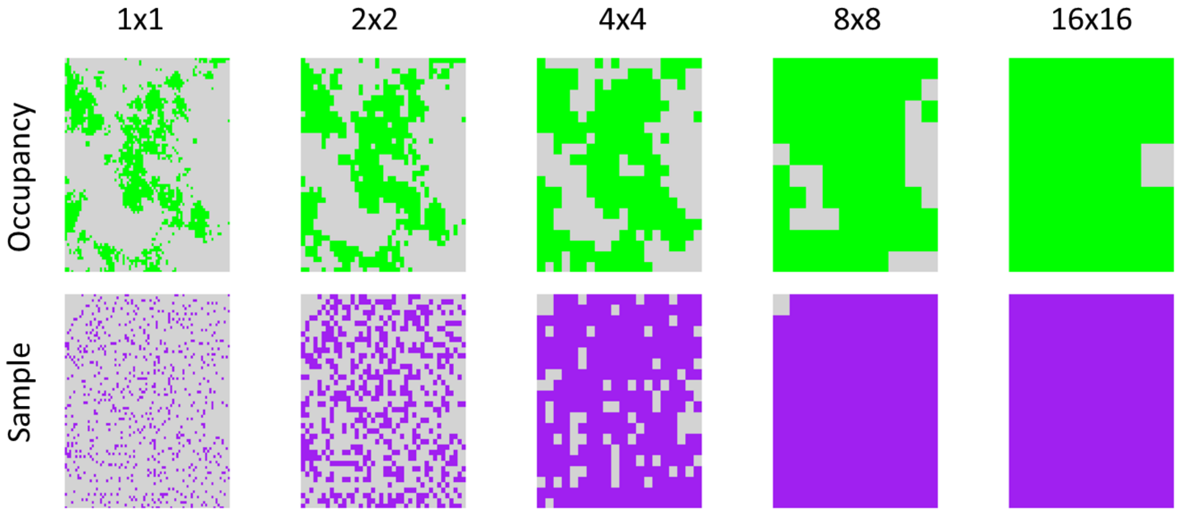

Figure 1. Top row: a virtual species' ("medium and clustered"; Table 1) geographic distribution at each spatial resolution. Green cells are occupied, and grey cells are not. Bottom row: a virtual sample at each resolution. $\rho(R,Y)\sim 0.05$ and $f\sim 0.1$ at the finest resolution (1×1). Purple cells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

Results

Error

For all virtual species, estimates of occupancy are more accurate at coarser resolutions. This 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.

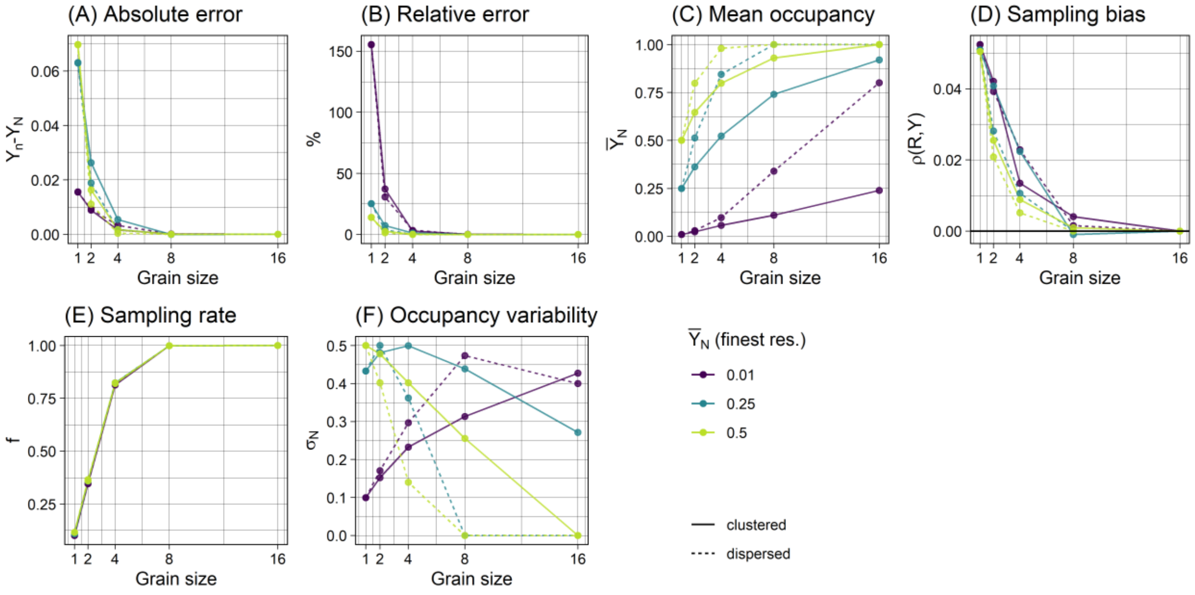

Figure 2. (A) absolute error, (B) relative error (i.e. the absolute error expressed as a percentage of true occupancy), (C) mean occupancy (i.e. true occupancy), (D) sampling bias, (E) sampling rate and (F) occupancy variability σ_Y at each resolution. The resolution is the height and width of the grid 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.

True occupancy

Although well-documented (Azaele et al., 2012; Kunin, 1998), it is worth revisiting the scaling properties of $\bar{Y}_N$ (i.e. a species' true occupancy) here, because they provide insight into the scaling properties of error. $\bar{Y}_N$ 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 $\bar{Y}_N = 1$ as the resolution is coarsened. By contrast, species that are rare and clustered at the finest resolution do not reach $\bar{Y}_N = 1$ at any of the resolutions we considered (Fig. 2A).

Sampling bias

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

Sampling rate

The sampling rate f scales in a similar way with resolution to $\bar{Y}_N$. It always increases with resolution, and sparser samples increase at a greater rate. In our simulations, sparser samples are slightly more likely for the sparsely distributed species, because we forced a correlation between the species' distributions and the samples (i.e. a sampling bias). Hence, f does not increase at exactly the same rate for all species.

Occupancy variability

As occupancy is binary, $\sigma_Y = \sqrt{\bar{Y}_N(1 - \bar{Y}_N)}$ (Bradley et al., 2021). It is largest where $\bar{Y}_N$ is near 0.5 and smallest where $\bar{Y}_N$ is near 0 or 1. Given that $\bar{Y}_N$ increases with resolution (Fig. 2C), coarsening the resolution for species with $\bar{Y}_N < 0.5$ increases σ_Y until $\bar{Y}_N = 0.5$ (Fig. 2F). Further coarsening the resolution decreases σ_Y, because $\bar{Y}_N$ moves away from 0.5 and towards 1. For species with $\bar{Y}_N \geq 0.5$ at the finest resolution, coarsening the resolution always decreases σ_Y.

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. 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). Under a simple random sample at the finest resolution, where the expected sampling bias $E_R[\rho(R, Y)] = 0$, 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 (blue lines in Fig. 3). For these species, $\bar{Y}_N$ 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.

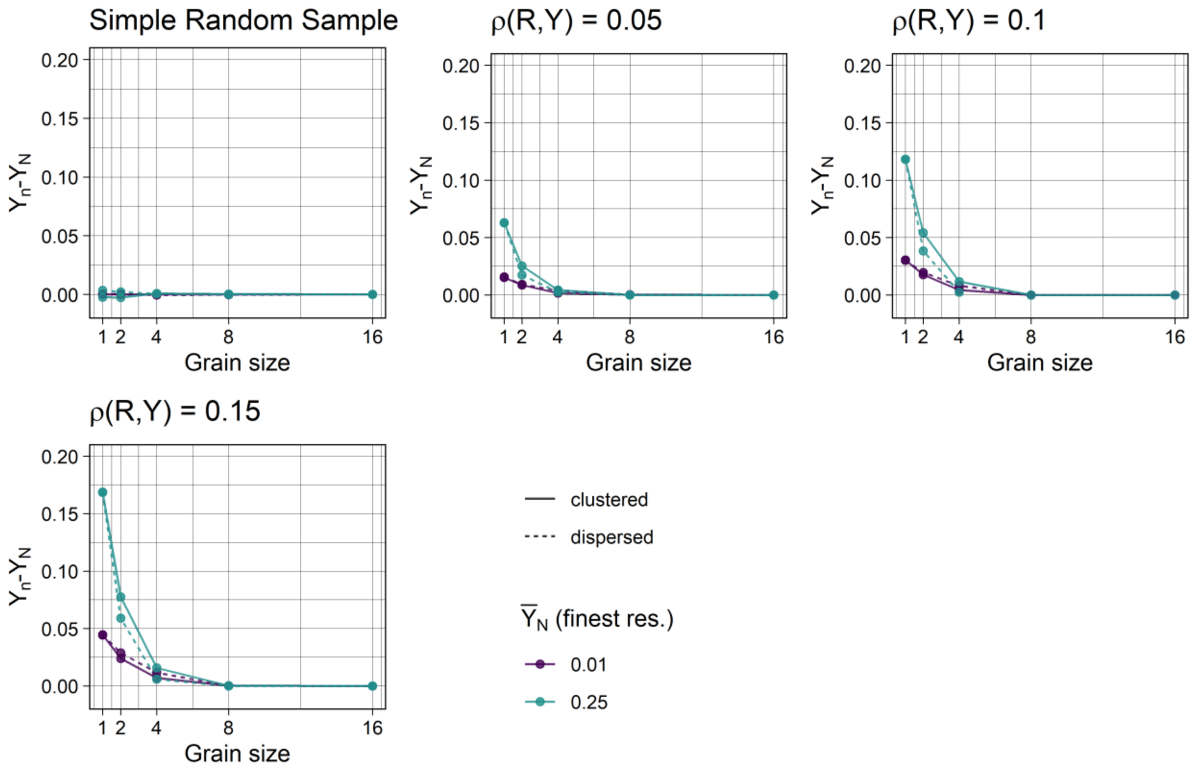

Figure 3. Absolute error at each resolution under four levels of sampling bias $\rho(R,Y)$ (at the finest resolution). The resolution is the height and width of the grid cells in arbitrary units. The simple random sample has approximately no sampling bias at the finest resolution. Each line represents one virtual species: red = rare, green = medium and blue = common. Solid lines depict species with clustered distributions at the finest resolution and dashed lines indicate species that are highly dispersed 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.

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 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 will 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., 2023).

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, because it always increases with resolution. The effect of resolution on the variability of occupancy in the landscape depends on the species' prevalence (i.e. $\bar{Y}_N$) at the finest resolution. If there is good reason to think that $\bar{Y}_N \geq 0.5$—say, from an expert drawn range map—then coarsening the resolution will always reduce σ_Y. On the other hand, if there is good reason to think that the species is rare, then coarsening the resolution will increase σ_Y until the $\bar{Y}_N$ reaches 0.5. The effect of spatial resolution on sampling bias $\rho(R,Y)$ is the most difficult to assess of the three quantities that determine error.

In our simulations, $\rho(R, Y)$ generally decreased as the spatial resolution was coarsened, but this will not be universally true. Recall 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 general insight into how $\rho(R, Y)$ might scale with resolution in other situations can be gleaned from the formula for Pearson's correlation coefficient. $\rho(R, Y)$ is the Pearson's correlation between R and occupancy Y, which is to say their covariance divided by the product of their standard deviations. We have already seen that coarsening the resolution of analysis increases the standard deviation of Y σ_Y until $\bar{Y}_N \geq 0.5$, at which point further coarsening the resolution reduces it. The same logic applies to the standard deviation of the R, which is also a binary variable. It follows that the denominator in the formula for $\rho(R, Y)$, the product of the standard deviations of Y and R, increases as the resolution is coarsened to the point where $\bar{Y}_N \geq 0.5$ and $P(R = 1) \geq 0.5$, at which point further coarsening the resolution reduces it. For a given covariance between occupancy and R then, coarsening the resolution of analysis will reduce $\rho(R, Y)$ where $\bar{Y}_N \geq 0.5$ and $P(R = 1) \geq 0.5$. Further work is needed to understand how the covariance between occupancy and R will vary with spatial resolution under different conditions.

Of course, error is not the sole criterion on which analysts should base their decision about the spatial resolution at which to work, because estimates of species occupancy become less useful at coarse resolutions (assuming a given level of accuracy). For one, the power to detect change is greater at fine scales, because trends at some fine scale might not be evident at a coarser one (Jönsson et al., 2021). Coarsening the resolution of estimation thus stands somewhat in opposition to the principle espoused by the Convention on Biological Diversity (CBD) that indicators should be sensitive to change (https://www.cbd.int/indicators/indicatorprinciples.shtml; although the CBD also ask for "scientific soundness" and "policy relevance", implying minimal error as a strongly desirable property). Other limitations of working at coarse resolutions are that occupancy is a better surrogate for abundance and often more relevant to policy at fine scales (Kunin, 1998; Spake et al., 2022), and that modelling the ecological or data generating processes becomes more difficult where the scale of analysis is much coarser than the scales at which they operate (but see Hill, 2012). Clearly, there is a trade-off between the usefulness and accuracy of estimates of species occupancy.

Importantly, however, the usefulness of an estimate is conditional on it being at least reasonably accurate. Imagine a species whose occupancy declines at some fine scale over time. It is sampled in two time-periods, and the sampling bias is strong in both periods. If the sampling bias switches direction from negative in the first period to positive in the next, then we may fail to detect the decline or even spuriously detect an increase (depending on the relative magnitudes of the sampling bias; Bowler et al., 2022; Pescott et al., 2019). Working at a coarser resolution might reduce the error in both time-periods to the point where the actual trend (at the coarser scale) is detectable and the chance of detecting a spurious trend is low. Of course, if the sampling bias has the same sign in both time-periods, then we may be able to detect the decline at the fine resolution despite under- or overestimating occupancy in both periods (Pocock et al., 2023). Ultimately intuition about the likelihood of such scenarios requires familiarity with the species' datasets being used for an analysis and clear assessments of the likely risk of bias (Boyd et al., 2022; Boyd, Powney, et al., 2023).

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 (e.g. Boyd, August, et al., 2023)—and particular time-periods were omitted, because of serious concerns about sampling biases affecting species data at finer scales across the 20th century. For example, rarer and more critical 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. Moreover, it is worth noting that the parameters of any such integrated model will also be subject to potential biases in estimation in the face of important unmodelled sampling variation.] 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 & 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 grid cell, 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 & Nichols, 2019; Jönsson et al., 2021), but 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 (cf. 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 (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 & 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.

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 & 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.

Code availability

All code needed to fully reproduce our analysis is available at https://github.com/robboyd/biasVsResolution.

References

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

Azaele, S., Cornell, S. J., & Kunin, W. E. (2012). Downscaling species occupancy from coarse spatial scales. *Ecological Applications*, *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. (2023). An operational workflow for producing periodic estimates of species occupancy at large scales. *Biological Reviews*, *9*. https://doi.org/10.32942/OSF.IO/2V7JP

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*, *xx*(xx), 1–11. 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. *Ecoevorxiv*, *April*. https://doi.org/10.32942/X2359P

Bradley, V. C., Kuriwaki, S., Isakov, M., Sejdinovic, D., Meng, X. L., & Flaxman, S. (2021). Unrepresentative big surveys significantly overestimated US vaccine uptake. *Nature*, *600*(7890), 695–700. https://doi.org/10.1038/s41586-021-04198-4

Collins, L. M., Schafer, J., & Kam, C. (2001). A Comparison of Restrictive Strategies in Modern Missing Data Procedures. *Psychological Methods*, *6*(June). https://doi.org/10.1037/1082-989X.6.4.330

Cooke, R., Mancini, F., Boyd, R., Evans, K. L., Shaw, A., Webb, T. J., & Isaac, N. J. B. (2023). Protected areas support more species than unprotected areas in Great Britain , but lose them equally rapidly. *Biological Conservation*, *278*(December 2022), 109884. https://doi.org/10.1016/j.biocon.2022.109884

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. *Journal of Insect Conservation*, *23*(2), 369–380. https://doi.org/10.1007/s10841-019-00135-z

Elliott, M. R., & Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, *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 in Ecology and Evolution*, *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 in Ecology and Evolution*, *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. *Conservation Biology*, *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. *Diversity and Distributions*, *21*(1), 46–54. https://doi.org/10.1111/ddi.12255

Hill, M. O. (2012). Local frequency as a key to interpreting species occurrence data when recording effort is not known. *Methods in E*, *3*(2012), 195–205. https://doi.org/10.1111/j.2041-210X.2011.00146.x

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 in Ecology and Evolution*, *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. *Ecological Modelling*, *422*(December 2019), 108927. https://doi.org/10.1016/j.ecolmodel.2019.108927

Jönsson, G. M., Broad, G. R., Sumner, S., & Isaac, N. J. B. (2021). A century of social wasp occupancy trends from natural history collections: spatiotemporal resolutions have little effect on model performance. *Insect Conservation and Diversity*, *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. *American Naturalist*, *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. *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.

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. *Journal of Applied Ecology*, *48*(1), 56–66. https://doi.org/10.1111/j.1365-2664.2010.01921.x

Outhwaite, C., Gregory, R. D., Chandler, R. E., Collen, B., & Isaac, N. J. B. (2020). Complex long-term biodiversity change among invertebrates, bryophytes and lichens. *Nature Ecology & Evolution*. https://doi.org/10.1038/s41559-020-1111-z

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. *Scientific 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? *British & Irish Botany*, *1*(4), 250–282. https://doi.org/10.33928/bib.2019.01.250

Pocock, M. J. O., Logie, M., Isaac, N. J. B., Fox, R., & August, T. (2023). The recording behaviour of field-based citizen scientists and its impact on biodiversity trend analysis. *Ecological Indicators*, *151*(April), 110276. https://doi.org/10.1016/j.ecolind.2023.110276

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

Roy, H. E., Adriaens, T., Isaac, N. J. B., Kenis, M., Onkelinx, T., Martin, G. S., Brown, P. M. J., Hautier, L., Poland, R., Roy, D. B., Comont, R., Eschen, R., Frost, R., Zindel, R., Van Vlaenderen, J., Nedvěd, O., Ravn, H. P., Grégoire, J. C., de Biseau, J. C., & Maes, D. (2012). Invasive alien predator causes rapid declines of native European ladybirds. *Diversity and Distributions*, *18*(7), 717–725. https://doi.org/10.1111/j.1472-4642.2012.00883.x

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. *Annual Review of Environment and Resources*, *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. *Journal of Applied Ecology*, *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.

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

Woodcock, B. A., Isaac, N. J. B., Bullock, J. M., Roy, D. B., Garthwaite, D. G., Crowe, A., & Pywell, R. F. (2016). Impacts of neonicotinoid use on long-term population changes in wild bees in England. *Nature Communications*, *7*. https://doi.org/10.1038/ncomms12459