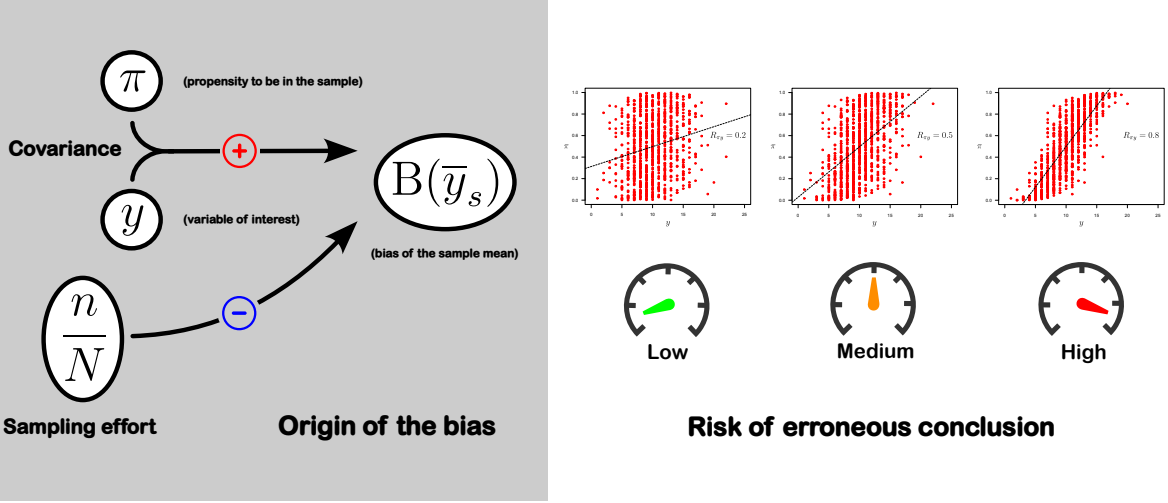
On the impact of preferential sampling on ecological status and trend assessment

Philippe Aubry, Charlotte Francesiaz, Matthieu Guillemain

Graphical Abstract

On the impact of preferential sampling on ecological status and trend assessment

Philippe Aubry, Charlotte Francesiaz, Matthieu Guillemain



Highlights

On the impact of preferential sampling on ecological status and trend assessment

Philippe Aubry, Charlotte Francesiaz, Matthieu Guillemain

- Preferential sampling can be modeled by Poisson or conditional Poisson sampling.
- We document the impact of preferential sampling on population mean estimation.
- Preferential sampling can lead to biased mean estimation, if not accounted for.
- The bias increases with the covariance between sample membership and the variable.
- The bias decreases with increasing sampling effort (expected sampling fraction).

On the impact of preferential sampling on ecological status and trend assessment

Philippe Aubry[a,*], Charlotte Francesiaz[b], Matthieu Guillemain[c]

[a] *OFB - Office français de la biodiversité - Direction surveillance, évaluation, données - Unité données et appui méthodologique, Saint Benoist, BP 20, 78612 Le Perray-en-Yvelines, France*
[b] *OFB - Office français de la biodiversité - Direction de la recherche et de l'appui scientifique - Service conservation et gestion durable des espèces exploitées, Les Portes du Soleil, 147 avenue de Lodève, 34990 Juvignac, France*
[c] *OFB - Office français de la biodiversité - Direction de la recherche et de l'appui scientifique - Service conservation et gestion durable des espèces exploitées, La Tour du Valat, Le Sambuc, 13200 Arles, France*

Abstract

Assessments of the status and trends of abiotic and biotic indicators are two central objectives in many ecological studies and monitoring programs. Given the impracticality of making measurements or observations at every point in geographic space, even within a limited domain, consideration of spatial sampling is crucial to ensure the reliability of statistical inference regarding such status or temporal trends.

The sampling units in geographic space (e.g., sites, plots, quadrats) for field observations are often selected with a preference for those expected to be species-rich or those with the highest abundances or occupancy probabilities. This sampling approach, called *preferential sampling*, can be based on probability sampling theory, but in practice, it is usually a form of nonprobability sampling.

Introducing a selection force that disproportionately includes units in the sample based on the expected values of the variables of interest can lead to (severely) biased inferences. This is because inclusion probabilities — referred to here as *propensities* for units to be part of the sample — cannot be accounted for in statistical estimators when they are unknown to the sampler.

In this article, we model sampling processes (considered without replacement) for a finite spatial population of sampling units using probability sampling designs. We consider four designs: Bernoulli sampling, Poisson sampling, simple random sampling, and conditional Poisson sampling. We document the bias introduced by preferential sampling in the estimation of a mean, whether for a status assessment (e.g., mean species richness) or a trend assessment (e.g., trend in mean abundance). For this purpose, we use Monte Carlo simulations and an analytical expression for the bias of the sample mean.

This analytical expression shows that the bias of the sample mean (1) increases with increasing covariance between the propensities and the values of the variable of interest and (2) decreases with increasing sampling effort (sampling fraction or expected sampling fraction). This fundamental statistical result is neither widely known nor appreciated by most ecologists, even though it has the potential to ruin status or trend assessments and to lead to erroneous conclusions.

The findings on preferential sampling in ecology presented in this article are reviewed from a methodological perspective, mainly for an audience of quantitative ecologists, wildlife statisticians, and biometricians involved in the design or implementation of ecological studies and monitoring programs. To facilitate future exchange among researchers on this topic by clarifying the concepts, in the discussion we also examine the terminology found in the literature for the notions related to preferential sampling.

Keywords:
biased site selection, spatial sampling processes, propensities, range shift, abundance gradient, conditional Poisson sampling

*Corresponding author
Email addresses: philippe.aubry@ofb.gouv.fr (Philippe Aubry), charlotte.francesiaz@ofb.gouv.fr (Charlotte Francesiaz), matthieu.guillemain@ofb.gouv.fr (Matthieu Guillemain)

1. Introduction

Although it has been abandoned in modern physics, the separation of space and time proves convenient and operational at macroscopic scales relevant to ecology. This allows us to consider all ecological phenomena as occurring in space and time, whether these dimensions are considered jointly or separately in their description, analysis and modeling.

In practice, the measurements or observations made, whether on abiotic or biotic variables, concern a given spatial domain (which we denote $\mathcal{D}$) and which take place over a given period (which we denote $\mathcal{T}$). Even if $\mathcal{D}$ and $\mathcal{T}$ are not large, it is generally impossible to make measurements or observations at every point in $\mathcal{D}$ or every time in $\mathcal{T}$. Thus, a fundamental sampling problem immediately arises.

In the following, we assume that $\mathcal{D}$ is a two-dimensional spatial domain lying in Euclidean space. Although not mandatory, $\mathcal{D}$ is often partitioned into a finite population of areal units (see, e.g., Aubry and Francesiaz, 2022, Fig. 1). The same is true for the time domain, which can be discretized into a finite population of time units (e.g., potential daily counting sessions for bird counting).

Regardless of whether the population of spatiotemporal units is finite or considered as infinite, data collection is performed in two steps: (i) by selecting a subset of units from $\mathcal{D} \times \mathcal{T}$ to form a sample s and (ii) by making measurements or observations on the units of s.

In the technical mathematical sense, an *error* is the deviation between a state (or a summary of a state) defined on $\mathcal{D} \times \mathcal{T}$ and its estimate or prediction computed from the sample s at hand. The first step mentioned above introduces a sampling error because only a portion of the sampled spatiotemporal population is considered when collecting the data. The second step generates an error that can be treated as a measurement or observation error, for instance, when counting individuals (e.g., Aubry et al., 2012) or when assessing the size of vast groups of wild animals (e.g., Vallecillo et al., 2021). When individuals are counted, this error can also be treated as a sampling error — where the set of individuals counted is a sample of all individuals present — due to *imperfect detection* (e.g., White, 2005; Kellner and Swihart, 2014; Perret et al., 2023; Johnston et al., 2023, Sec. 4.2). In this article, we will focus only on the first source of error, concentrating on the sampling of $\mathcal{D}$. Since we are interested only in spatial sampling, we do not discuss methods for counting individuals or problems of imperfect detection or of imperfect ability to identify and to count them when they are detected (see, e.g., the references cited by Vallecillo et al., 2021 and Aubry et al., 2023).

In a general sense, we call the *sampling process* the way by which a sample s is selected from a statistical population, regardless of the method used. We reduce the set of sampling processes by assuming here that they are without replacement. We assume that the sampling process under consideration can be replicated — at least in principle — so that distinct samples (see Hedayat and Sinha, 1991, p. 2) can be obtained on the basis of selection probabilities $0 < p(s) < 1$. The individuals or groups of individuals who perform the sampling may be collectively referred to as the *sampler*. The sampling processes considered in this article are therefore human-based and, as such, are not necessarily easy to model in detail (but see ter Steege et al., 2011; Fernández and Nakamura, 2015).

Statistical inference from a sample can be used to estimate parameters of an actual finite statistical population or parameters of a hypothetical superpopulation describing the stochastic process under study (see, e.g., Aubry and Francesiaz, 2022, Sec. 2.3). Sampling is said to be *preferential* when the sampling process is related to a superpopulation (e.g., Diggle and Ribeiro, 2007, Sec. 4.4.2; Diggle et al., 2010; Watson et al., 2019; Gray and Evangelou, 2023) or to the finite population at hand in such a way that units are more likely to be included in the sample if they have the highest values for the indicator of interest. For example, in ecology, preferential sampling may refer to preferential inclusion in the sample of the most species-rich units, the units with the highest abundances or occupancy probabilities. Conversely, we use the expression *antipreferential sampling* to denote situations where units with the lowest values are more likely to be included in the sample. In all the other cases where the probability of inclusion in the sample is independent of the characteristics of the unit with respect to the variables of interest, sampling is said to be nonpreferential.

With the previous definition in mind, preferential sampling should be distinguished from (i) *convenience sampling* (sometimes called *accessibility sampling*; see, e.g., Young and Young, 1998, p. 93 or Barnett, 2002, p. 17) and (ii) *purposive sampling* (also known as *judgment sampling*). In the first case, units are selected for inclusion in the sample for convenience (e.g., ease of access and/or safety for the observers). In the second case, units judged to be typical or appropriate for the survey are selected from the statistical population under study, for example, when "*ecologists actively seek the event of interest, such as an active breeding nest of a bird, or a specific plant species*" (Edwards et al., 2006). The previous two sampling methods correspond to two forms of nonprobability sampling. Preferential sampling can be either a nonprobability sampling method or be based on a random selection mechanism. We distinguish between the two cases (i.e., nonprobability vs. probability preferential sampling) because the statistical implications are quite different. Indeed, in the first case, statistical inference

must be based on a model — which can be implicit — and its key assumptions, while in the second case, the properties of the estimator used can be based solely on the random selection mechanism (see, e.g., Aubry and Francesiaz, 2022).

As a concrete situation, we assume in this article that sampling is performed in a nonprobability setting. We consider two objectives that are central to many ecological studies or monitoring programs: assessment of (i) status (i.e., the state observed at a given point in time) or (ii) trend (i.e., a smooth pattern of state variation over time) (see, e.g., Gitzen et al., 2012). We do not discuss the scientific relevance of either of these depending on the context in which they occur (see, e.g., Vos et al., 2000; Yoccoz et al., 2001, 2003; Nichols and Williams, 2006). We merely consider status and trend as objects of statistical inference, and our aim is to examine the impact of (nonprobability) preferential (spatial) sampling on their assessment, mainly in terms of estimation bias. To this end (i) we formalize preferential sampling in the context of finite population sampling; (ii) we briefly consider the one-stage probability sampling designs that can be used to model basic without-replacement sampling processes; (iii) using appropriate sampling process models, we examine the case of status and trend estimation under preferential sampling through Monte Carlo simulation studies; (iv) we introduce to ecology a fundamental formula that enables us to understand the very nature of the mean estimation bias that can be caused by preferential sampling. Finally, we draw lessons from the results presented, particularly with respect to trend assessment. Fig. 1 is intended to provide the reader with an overview of the article organization and of the topics covered before the discussion and perspectives sections.

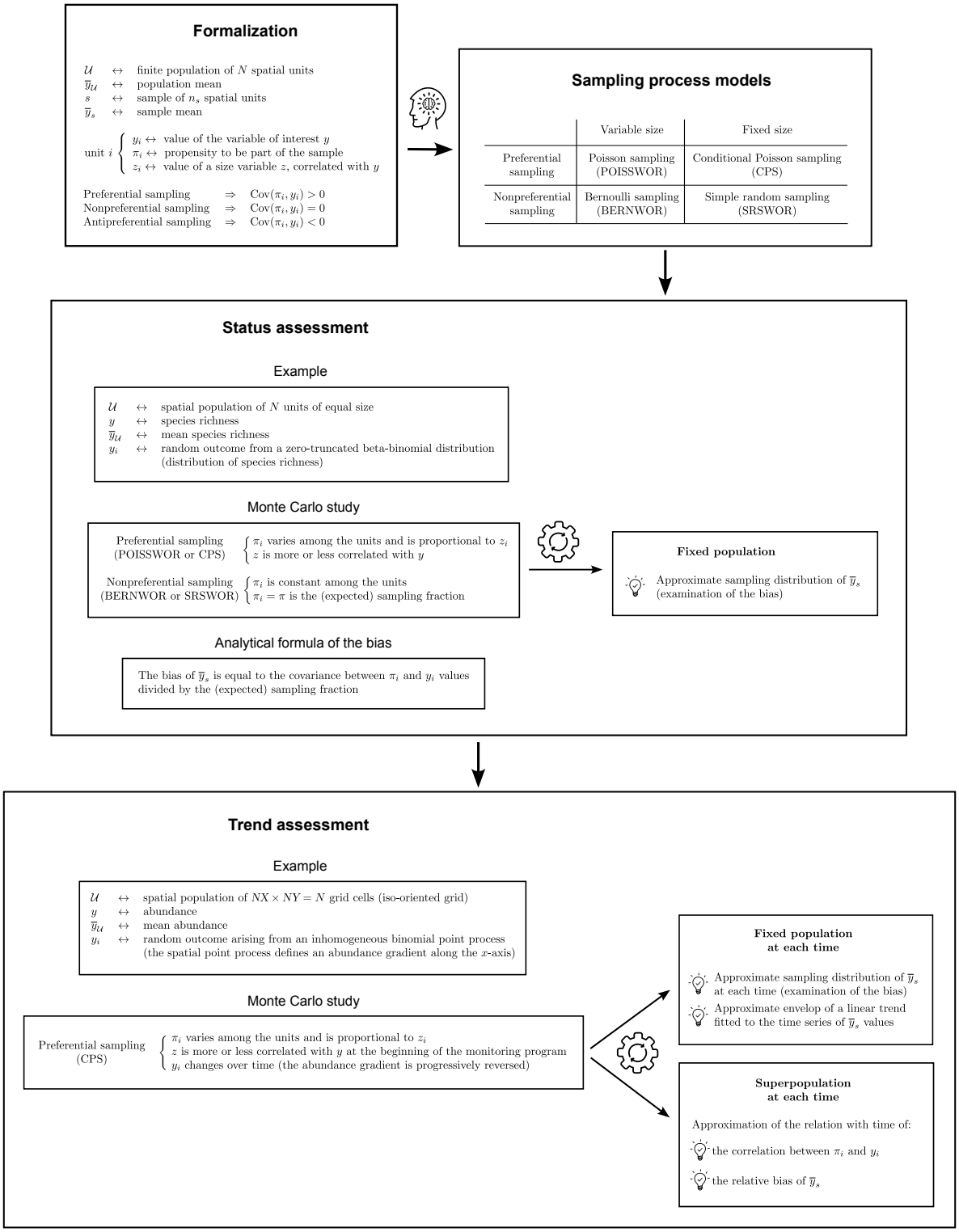


Figure 1: Overview of the article organization before the discussion and perspectives sections. We first formalize preferential sampling in the context of finite population sampling. Next, we consider the probability sampling designs that can be used to model basic nonpreferential/preferential sampling processes. With these models, we then use Monte Carlo simulations to study the impact of preferential sampling, first for an example of status assessment, second for an example of trend assessment. The type of results to be examined is indicated to the right, either in a fixed population context or in a superpopulation context.

2. Formalization

From a population of spatial units (e.g., sites, plots, quadrats) $\mathcal{U}$ of size N, spatial sampling consists of selecting a subset $s \subseteq \mathcal{U}$ of size $n_s \leq N$, called a sample. The sampled spatial population should not be confused with a biological population, i.e., here a set of organisms present in the domain $\mathcal{D}$.

The size of the spatial population N is not necessarily known to the sampler. In addition, the sample size n_s may be variable or fixed. It is variable, for example, when the sample is drawn by a group of people who do not necessarily consult each other and who each add one or more units to the sample. The sample size can be fixed if the sample selection is more centralized, for example, to predetermine the sampling effort to be allocated.

In the following, we consider a univariate situation, i.e., there is only one variable of interest y, which takes fixed values for the units $i \in \mathcal{U}$. We denote $\mathbf{y} = (y_1, y_2, \ldots, y_N)$ as the vector of y-values for $i \in \mathcal{U}$. By extension, this vector can be called the population (implicitly of y-values) associated with the population of spatial sampling units $\mathcal{U}$. The $\mathbf{y}$-vector results from a bioecological stochastic process, which may be modeled if appropriate. The observed $\mathbf{y}$-vector is then assumed to be a realization of a random vector (i.e., a superpopulation).

2.1. Inclusion probabilities

In any sampling process, the probability $0 < \pi_i \leq 1$ that a unit $i \in \mathcal{U}$ is part of s: (i) is a probability of inclusion known to the sampler, as is the case in finite population sampling theory (e.g., Särndal et al., 1992; Tillé, 2020); or (ii) expresses the greater or lesser propensity of a unit to be part of the sample, which is unknown to the sampler. In the latter case, the probabilities π_i ($i \in \mathcal{U}$) are *propensity scores*, here simply called *propensities*. We denote $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_N)$ as the vector of probabilities for $i \in \mathcal{U}$. Let us note that the case $\pi_i = 0$ corresponds to the exclusion of unit i from the sampled population, a trivial case not of interest here.

Whether we are dealing with inclusion probabilities in the strict sense or with propensities, the $\boldsymbol{\pi}$-vector results from the sampler's view of the variation of the variable of interest across geographic space, without actually knowing it precisely. In particular, the probabilities considered in this article are not modified when y-values are observed; therefore, the sampling process can be considered nonadaptive. Thus, in this article, we assume that the samples are generated by a sampling process governed by probabilities that are fixed in advance, whether consciously or not.

Where appropriate, the probabilities π_i can be defined as being proportional to a variable z ($z_i > 0$ for all $i \in \mathcal{U}$), which assigns more or less importance to the units — this type of variable is known as a *size variable* — and is assumed to be correlated to some level with the variable of interest y.

2.2. Preferential sampling

The sampling process is said to be preferential when there is a monotonically increasing relation between the y-values and the inclusion probabilities, whether the latter are known or unknown. If the relation decreases, we have antipreferential sampling.

We do not assume that the relation is linear, as this assumption is too restrictive to be realistic. Indeed, a linear relation assumes that the propensity increases at a constant rate as a function of the y-variable, whereas in reality, it may increase faster at higher y-values, for example. Moreover, the relation in question is not necessarily analytical; it can be only statistical — corresponding to the conditional expectation — if the propensity varies for each y-value.

A necessary condition for sampling to be preferential (or antipreferential) is that the probabilities π_i ($i \in \mathcal{U}$) differ significantly between units or at least between groups of units. Sampling is inherently nonpreferential if all units have, at least approximately, the same probability of being included in the sample.

2.3. Notation

In any sampling problem, we can distinguish at least two sources of stochasticity (see, e.g., Aubry and Francesiaz, 2022): (i) one that relates only to the sampling process $p(s)$ of the finite population of units, resulting in the $\boldsymbol{\pi}$-vector and (ii) the other related to the stochastic process ξ at the origin of the $\mathbf{y}$-vector (superpopulation). To avoid any ambiguity, operators (expectation, variance, etc.) are denoted by using p as a subscript for the first source and ξ for the second. For a parameter ω, the sampling distribution of an estimator $\hat{\omega}$, i.e., the distribution of all the values that $\hat{\omega}$ can take for all the samples that can be generated, is called the p-distribution. The mean of the p-distribution is called the p-expectation of $\hat{\omega}$, its variance is the p-variance (i.e., the sampling variance), and so on. We define the p-bias as $\mathrm{B}_p(\hat{\omega}) = \mathrm{E}_p(\hat{\omega}) - \omega$ and the relative p-bias as $\mathrm{B}_p(\hat{\omega})/\omega$. For the stochastic process ξ, we similarly denote the ξ-expectation, ξ-variance, and so on. When we consider both

sources of stochasticity at the same time, we associate the two subscripts, and we write, for example, ξp-expectation. Let us recall that in this article, in the context of double stochasticity, $\boldsymbol{\pi}$ is fixed, while $\mathbf{y}$ is random. Other notations used in this article are detailed in Appendix A.

2.4. Parameters and estimators

Regardless of whether the spatial population size N is known to the sampler, for a variable of interest y (discrete, including binary, or continuous), we assume that the parameter to be estimated is the population mean $\overline{y}_{\mathcal{U}}$.

In the following, we consider a concrete situation in which (a) the size of the spatial population is unknown to the sampler (there is no sampling frame) and (b) the inclusion probabilities are also unknown to the sampler (the sample is not the result of applying a probability sampling design). In the absence of auxiliary variables that would be (strongly) correlated with y and exhaustively known on $\mathcal{D}$, the only data available are the y_i values for $i \in s$ and the sample size n_s. Thus, by default, the population mean can be estimated by the sample mean $\overline{y}_s$, and the p-variance of $\overline{y}_s$ can be estimated using:

$$\widehat{\mathrm{V}}_p\left(\overline{y}_s\right) = \frac{s_y^2}{n_s} \tag{1}$$

3. Sampling process models

The fact that only a part (i.e., the sample) of the population of spatial sampling units is considered for data collection represents a source of uncertainty that may be modeled. This can be referred to as a *sampling model* (Hobbs and Hooten, 2015, Sec. 1.1.3) or an *inclusion model* (Gelman et al., 2014, Sec. 8.2), although both terms are polysemous which can lead to confusion.

As noted by Hájek (1964, p. 1492), any probability distribution $p(s)$ can be used as a mathematical model for any sampling procedure, experiment, or method. Thus, in this article, we model a sampling process using an appropriate probability sampling design (Aubry, 2023, Sec. 1.2.4). Basic sampling processes (without replacement) can be modeled by one of four one-stage sampling designs, described below, depending on whether the propensities are equal or unequal and on whether the sample size is fixed or variable (Fig. 2).

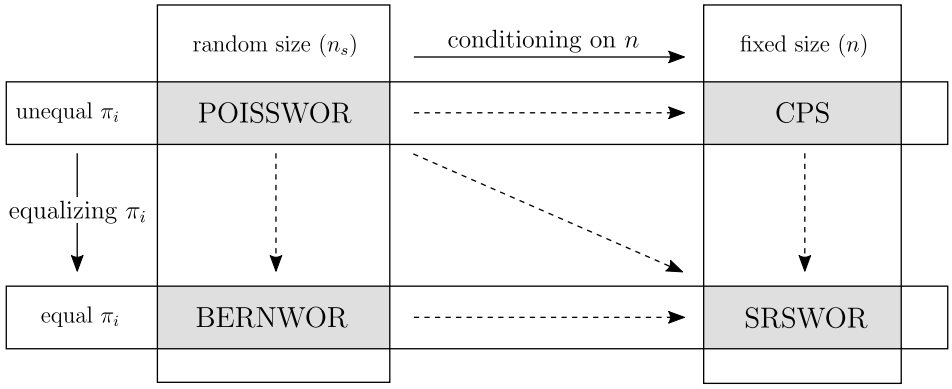


Figure 2: The four one-stage probability sampling designs used as without-replacement sampling process models. POISSWOR: Poisson sampling. BERNWOR: Bernoulli sampling. CPS: Conditional Poisson sampling. SRSWOR: Simple random sampling. The sampling designs are derived from each other by asymmetric relations of equalization of inclusion probabilities π_i ($i \in \mathcal{U}$) or conditioning on a fixed sample size n.

3.1. Preferential sampling — variable size (POISSWOR)

When the sample size is variable and the propensities vary across sampling units, the sampling process can be modeled by Poisson sampling, which we denote as POISSWOR.

In Poisson sampling, (i) the inclusion of a unit $i \in \mathcal{U}$ in sample s is governed by a Bernoulli trial with probability π_i and (ii) the resulting random sample size n_s follows a Poisson binomial distribution ($n_s \sim \mathrm{PoiBin}(N, \boldsymbol{\pi})$) (Hájek, 1981, Ch. 6; Särndal et al., 1992, Sec. 3.5; Tillé, 2006, Sec. 5.5; Tillé, 2020, Sec. 5.8). Let us note that the Poisson binomial distribution is here the generalization of the binomial distribution for Bernoulli trials with unequal probabilities (see, e.g., Wang, 1993) and should not be confused with the compound distribution described by Johnson et al. (2005, Sec. 9.5).

A necessary condition for using POISSWOR as a model for variable-size sampling processes without replacement is that unit inclusions must be assumed to result from independent events. This assumption is broadly consistent with the situation in which the sample results from the actions of many people who do not coordinate with each other.

3.2. Nonpreferential sampling — variable size (BERNWOR)

When the sample size is variable and the propensities are sufficiently similar to each other such that they can be considered equal to a constant probability π, the sampling process can be modeled by Bernoulli sampling (also known as *binomial sampling*), which we denote as BERNWOR.

In Bernoulli sampling, (i) the inclusion of any unit in the sample is governed by a Bernoulli trial with probability π and (ii) the resulting variable sample size n_s follows a binomial distribution ($n_s \sim \text{Bin}(N, \pi)$) (Särndal et al., 1992, Sec. 3.2; Tillé, 2006, Sec. 4.3; Tillé, 2020, Sec. 3.2).

BERNWOR is a special case of POISSWOR when the probabilities π_i ($i \in \mathcal{U}$) are equal (Fig. 2). When modeling a sampling process, BERNWOR has the same independence condition as POISSWOR.

3.3. Preferential sampling — fixed size (CPS)

Conditional on a fixed sample size ($n_s = n$ for all s), POISSWOR becomes conditional Poisson sampling, abbreviated as CPS (Hájek, 1981, Ch. 14; Tillé, 2006, Sec. 5.6; Tillé, 2020, Sec. 5.9) (Fig. 2). This is a general model suitable for a fixed-size preferential sampling process.

3.4. Nonpreferential sampling — fixed size (SRSWOR)

Conditional on a fixed sample size ($n_s = n$ for all s), BERNWOR becomes simple random sampling without replacement, abbreviated as SRSWOR, for which $\pi = n/N$ (Särndal et al., 1992, Sec. 3.3.1; Tillé, 2006, Sec. 4.4; Tillé, 2020, Sec. 3.3) (Fig. 2). SRSWOR is also a special case of CPS when the probabilities π_i ($i \in \mathcal{U}$) are equal (Fig. 2).

Let us recall that in this article, we assume that N is unknown to the sampler, so the propensity π is also unknown. SRSWOR is an appropriate model for a fixed-size sampling process with approximately constant propensities.

3.5. Estimators

Unlike the concrete situation we considered in Section 2.4, in Monte Carlo simulations, the $\boldsymbol{\pi}$-vector is known, as is the population size N. In this case, the population mean can be estimated without p-bias using the expansion estimator for the population total (Horvitz-Thompson estimator; see, e.g., Hedayat and Sinha, 1991, Ch. 3) to form the following weighted estimator:

$$\overline{y}_\pi = \frac{1}{N} \sum_{i \in s} \frac{y_i}{\pi_i} \tag{2}$$

For BERNWOR and SRSWOR, the estimator (2) simplifies to:

$$\overline{y}_\pi = \frac{1}{N} \frac{1}{\pi} \sum_{i \in s} y_i \tag{3}$$

which for SRSWOR gives the identity $\overline{y}_\pi = \overline{y}_s$ since, in this case, we have $\pi = n/N$.

For sampling processes modeled by BERNWOR or POISSWOR (variable sample sizes), the units are included in the sample independently. This gives a fairly simple expression for the p-unbiased estimator of the sampling variance (p-variance) of the weighted estimator (2) (see Särndal et al., 1992, p. 289). The expression for POISSWOR is written as:

$$\widehat{\mathrm{V}}_p\left(\overline{y}_\pi\right) = \frac{1}{N^2} \left[\sum_{i \in s} \frac{1 - \pi_i}{\pi_i^2} y_i^2 \right] \tag{4}$$

which simplifies for BERNWOR to:

$$\widehat{\mathrm{V}}_p\left(\overline{y}_\pi\right) = \frac{1}{N^2} \left[\frac{1 - \pi}{\pi^2} \sum_{i \in s} y_i^2 \right] \tag{5}$$

In the case of fixed-size designs, in general, to estimate without bias the p-variance, it is necessary to know the joint probabilities $\pi_{ij} > 0$ that two units $i \neq j \in \mathcal{U}$ are simultaneously included in the sample. The general expression for the p-unbiased estimator of the p-variance of the weighted estimator (2)

for an unequal-probability, fixed-size, without-replacement sampling design such as the CPS is (see Sen-Yates-Grundy estimator, Hedayat and Sinha, 1991, p. 52, Eq. 3.16):

$$\widehat{\mathrm{V}}_p\left(\overline{y}_\pi\right) = \frac{1}{N^2}\left[\sum\sum_{(i<j)\in s}\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2\right] \tag{6}$$

which simplifies for SRSWOR since then we have $\pi_i = n/N$ for all $i \in \mathcal{U}$ and $\pi_{ij} = n(n-1)/[N(N-1)]$ for all $i \neq j \in \mathcal{U}$. After simplification, we obtain (e.g., Tillé, 2020, p. 29, Result 3.1):

$$\widehat{\mathrm{V}}_p\left(\overline{y}_\pi\right) = (1-\pi)\,\frac{s_y^2}{n} \tag{7}$$

In the class of fixed-size without-replacement sampling designs with inclusion probabilities $\boldsymbol{\pi}$, the CPS is the sampling design of maximum entropy (Hájek, 1981; Tillé, 2006; Tillé, 2020, Sec. 5.9). It follows that an approximation to the p-variance estimator (6) may be obtained without involving the joint inclusion probabilities π_{ij} $(i \neq j \in \mathcal{U})$. We refer the reader to Tillé (2020, Sec. 5.14) and to the references cited by Aubry (2023, Sec. 1.3.2).

4. Status assessment

From an operational perspective, for a given spatial domain $\mathcal{D}$, status at a given point in time corresponds to either (i) the state of a variable of interest or (ii) the value of an indicator summarizing the state of the variable of interest. In practice, the point in time is actually a time interval — ideally, the shortest possible — which we have denoted $\mathcal{T}$. In the case of abundance, for example, at a certain infra-annual time interval $\mathcal{T}$ over which a biological population may be assumed to be (approximately) closed both geographically and demographically, status in sense (i) may correspond to the spatial distribution of abundance, while in sense (ii), it may refer to average abundance. In the following, we refer to status in sense (ii) (indicator of interest).

4.1. Status example

As an example of a status to be estimated over a $\mathcal{D} \times \mathcal{T}$ spatiotemporal domain, we consider here the average species richness for a group of L species and a finite statistical population $\mathcal{U}$ of N equal-size spatial sampling units, discretizing all or a part of the spatial domain $\mathcal{D}$. The variable of interest y is therefore species richness, and the indicator is the finite population mean $\overline{y}_\mathcal{U}$ (the time interval $\mathcal{T}$ is implicit and omitted from the notation).

In this example, spatial units are considered habitat patches that are disjunct within the spatial domain of interest (Wiens, 1976; Hall et al., 1997; Girvetz and Greco, 2007). A habitat patch may be included in the sample from a previously established list of all patches (sampling frame) that has been established by habitat mapping, in which case N is known. In the absence of a sampling frame, habitat patches are included in the sample based on the sampler's prior knowledge or on what the sampler encounters in the field, in which case N is generally unknown. This second case corresponds to the concrete situation of preferential sampling considered in this article (Section 2.4).

In the case of (inherently) nonpreferential sampling, units are selected with the same propensity π. With preferential sampling, units are selected based on a subjective assessment of their expected species richness, with propensities more or less positively correlated with y-values.

4.2. Population model example

As an illustrative example of a variable of interest y, we consider species richness with values ranging from 1 to $L = 25$, distributed according to a zero-truncated beta-binomial distribution $\mathrm{BetaBinZT}(L, \alpha, \beta)$ with shape parameters $\alpha = 1$ and $\beta = 5$. These values are realistic since they are of the same order of magnitude as those obtained for the species richness of a set of 25 waterbird species (ducks, geese, swans, coots, waders and grebes) breeding in European France (see the LIMAT scheme described in Aubry et al., 2023, Fig. 1). According to the typology proposed by Aubry and Francesiaz (2022, Section 2.4, Table 1), this statistical distribution is a type IV superpopulation model; that is, it does not include information on spatial structure or covariates (auxiliary variables). This model specifies an infinite set of values, but we are interested here in sampling a finite population containing $N = 5\,000$ spatial sampling units.

Among the infinite set of populations containing $N = 5\,000$ y-values following a $\mathrm{BetaBinZT}(25, 1, 5)$ distribution, we choose one such that $|\hat{\alpha} - \alpha| < 10^{-3}$ and $|\hat{\beta} - \beta| < 10^{-3}$. The estimator used for the shape parameters α and β of the zero-truncated beta-binomial distribution is the second estimator

proposed by Tripathi et al. (1994, Sec. 3.1). We can visually check that the histogram of the variable y associated with the finite population $\mathcal{U}$ we used is very close to that of the superpopulation model (Fig. 3). Thus, we have $\overline{y}_{\mathcal{U}} = 5.0062$ for comparison with its expected value in the model $\mathrm{E}_\xi(\overline{y}_{\mathcal{U}}) = 5$.

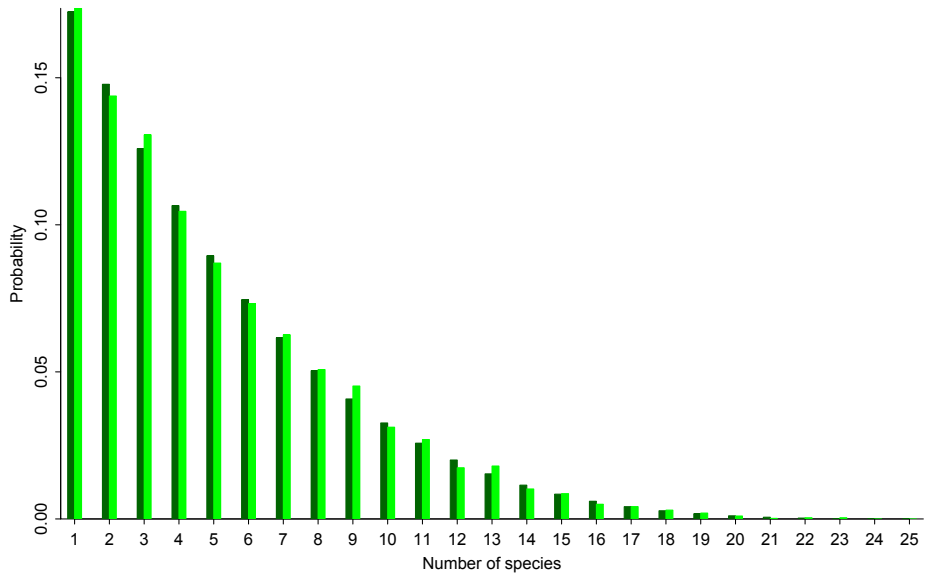


Figure 3: Distribution of the superpopulation model (in dark green) and histogram of the simulated population for $N = 5\,000$ (in light green).

4.3. Monte Carlo study

In this section, we examine the estimation of $\overline{y}_{\mathcal{U}}$ by $\overline{y}_s$ in four situations, combining the preferential vs. nonpreferential nature of the sampling process and a variable vs. fixed sample size.

In the case of preferential sampling, the propensities are defined as being proportional to a size variable z ($z_i > 0$ for all $i \in \mathcal{U}$):

$$\pi_i = P\,\frac{z_i}{Z} \qquad \text{with} \quad P = \sum_{i\in\mathcal{U}} \pi_i \qquad \text{and} \quad Z = \sum_{i\in\mathcal{U}} z_i \tag{8}$$

In the case of POISSWOR or BERNWOR, we have $P = \mathrm{E}_p(n_s)$ (variable sample size); in the case of CPS or SRSWOR, we have $P = n$ (fixed sample size).

Since we must respect the constraint $\pi_i \leq 1$, it follows that if we have $z_i > Z/P$ for at least one unit, we must perform the preprocessing described, for example, by Aubry (2023, Sec. 3.1) (see also Tillé, 2020, Sec. 5.2). From the above, we deduce that $R_{\pi y} = R_{zy}$ if and only if $z_i \leq Z/P$ for all $i \in \mathcal{U}$. In the following, we consider cases where this condition holds (for a counterexample, see Aubry et al., 2020, Sec. 4.4).

We define a model for the size variable z in which the correlation with y is ρ_{zy}. We vary ρ_{zy} between 0 and 1 by attenuating perfect positive correlation — for correlation attenuation; see, e.g., Charles (2005) — by adding a Gaussian error term $\epsilon \sim N(0, \sigma_\epsilon^2)$ to y. Thus, for each $i \in \mathcal{U}$:

$$\begin{array}{lll} z_i = \epsilon_i & \text{with } \sigma_\epsilon^2 > 0 & \text{for } \rho_{zy} = 0 \\ z_i = y_i + \epsilon_i & \text{with } \sigma_\epsilon^2 = S_y^2(1-\rho_{zy}^2)/\rho_{zy}^2 & \text{for } 0 < \rho_{zy} < 1 \\ z_i = y_i & \text{with } \sigma_\epsilon^2 = 0 & \text{for } \rho_{zy} = 1 \end{array} \tag{9}$$

Ultimately, to guarantee $z_i > 0$ (and hence, $\pi_i > 0$) for all $i \in \mathcal{U}$, we apply the linear transformation $z_i - \min(\mathbf{z}) + 0.1$.

In the same spirit as obtaining $\mathbf{y}$, we want the value of the finite population correlation R_{zy} to approximately match the superpopulation correlation ρ_{zy}. To achieve this, we generate random outcomes of the $\mathbf{z}$-vector until we obtain $|R_{zy} - \rho_{zy}| < 10^{-6}$. Having obtained the $\mathbf{z}$-vector, we then set a value for P and compute the $\boldsymbol{\pi}$-vector according to expression (8).

In the following, we replicate a given sampling process model with $P = 500$, i.e., with 10% as the sampling fraction (fixed sample size) or expectation of the sampling fraction (variable sample size). For each replication, we compute $\overline{y}_s$ and $\widehat{\mathrm{V}}_p(\overline{y}_s)$. We use 10^6 replications to accurately approximate the p-distributions of these two estimators. Therefore, we can assess the possible estimation bias of

the population mean $\overline{y}_{\mathcal{U}}$ by the sample mean $\overline{y}_s$ and also that of the p-variance $\mathrm{V}_p(\overline{y}_s)$ by the default estimator $\widehat{\mathrm{V}}_p(\overline{y}_s)$ (Eq. 1).

To evaluate the estimation bias of a parameter ω by using an estimator $\hat{\omega}$, we define a *bias index* $\mathrm{BI} = \mathrm{E}(\hat{\omega})/\omega$. The bias $\mathrm{E}(\hat{\omega}) - \omega$ is positive if $\mathrm{BI} > 1$, zero for $\mathrm{BI} = 1$, and negative if $\mathrm{BI} < 1$. Here, to assess the estimation bias for the p-variance of the sample mean, we form:

$$\mathrm{BI} = \frac{\mathrm{E}_{\mathrm{MC}}\left(\widehat{\mathrm{V}}_p(\overline{y}_s)\right)}{\mathrm{V}_{\mathrm{MC}}(\overline{y}_s)} \tag{10}$$

where $\mathrm{E}_{\mathrm{MC}}(\cdot)$ and $\mathrm{V}_{\mathrm{MC}}(\cdot)$ are the mean and variance calculated over the 10^6 replications of the Monte Carlo study.

4.3.1. Nonpreferential sampling — variable size (BERNWOR)

The sampling process model corresponding to (inherently) nonpreferential sampling of variable size is BERNWOR. The p-distribution of $\overline{y}_s$ for $\pi = 0.1$ ($\mathrm{E}_p(n_s) = 500$) is shown in Fig. 4.a.

In the BERNWOR case, the p-unbiased estimator of the population mean is the weighted estimator $\overline{y}_\pi$, which differs from the sample mean $\overline{y}_s$. This difference implies that the estimation of the population mean using the sample mean is p-biased. However, the bias is generally negligible and of no consequence, as shown in our example (Fig. 4.a). Here, although mathematically biased, the sample mean is actually a preferable estimator to the weighted estimator (see Särndal et al., 1992, p. 64, Eq. 3.2.6 and Tillé, 2020, p. 35).

The default estimator for the p-variance of $\overline{y}_s$ (Eq. 1) is slightly positively biased since we obtain $\mathrm{BI} \approx 1.1136$. The p-variance of $\overline{y}_s$ under the BERNWOR model is approximately the same as that under the SRSWOR model (Särndal et al., 1992, p. 65) (see Section 4.3.2 and Fig. 4).

4.3.2. Nonpreferential sampling — fixed size (SRSWOR)

The sampling process model corresponding to the (inherently) nonpreferential fixed-size sampling is SRSWOR. The p-distribution of $\overline{y}_s$ for $\pi = 0.1$ ($n = 500$) is shown in Fig. 4.b.

As recalled in Section 3, in the SRSWOR case, we have the identity $\overline{y}_\pi = \overline{y}_s$. This identity implies that the estimation of the population mean using the sample mean is p-unbiased (Fig. 4.b).

Conversely, the default estimator of the p-variance (Eq. 1) is positively biased since we obtain $\mathrm{BI} \approx 1.1111$. This is because the unbiased estimator in the SRSWOR model (Eq. 7) introduces a *finite population correction* (see Cochran, 1977, Sec. 2. 6) $\mathrm{fpc} = 1 - \pi$, which is closer to 0 the closer π is to 1. At the extreme, the p-variance becomes zero when $n = N$ (*finite population consistency*; see Cochran, 1977, Sec. 2.4, p. 21 or Hankin et al., 2019, p. 325). The magnitude of the p-variance overestimation depends on the value of the sampling fraction $\pi = n/N$, which remains unknown when N is not known. Let us note that the value obtained by Monte Carlo simulation for BI agrees with the theoretical value $(1 - \pi)^{-1}$ for $\pi = 0.1$.

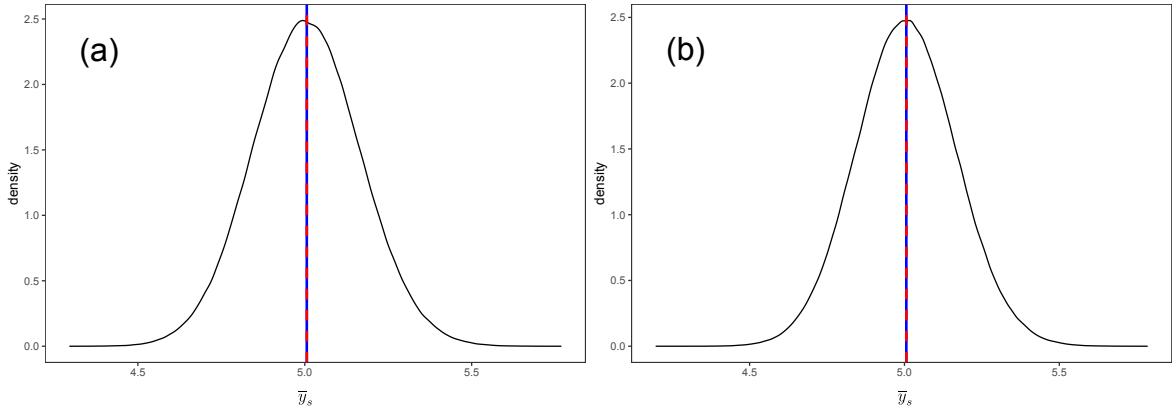


Figure 4: Approximate sampling distributions using 10^6 replications of each nonpreferential sampling model for $\pi = 0.1$. (a) BERNWOR. (b) SRSWOR. The population mean is shown by the red dashed line. The average of the values taken by $\overline{y}_s$ is shown by the blue line.

4.3.3. Preferential sampling — variable size (POISSWOR)

The sampling process model corresponding to variable-size preferential sampling is POISSWOR. The vector of probabilities $\boldsymbol{\pi}$ is computed as explained in Section 4.3 for $\mathrm{E}_p(n_s) = 500$ (variable sample size). The p-distributions of $\overline{y}_s$ for $R_{\pi y} = 0.5$ and $R_{\pi y} = 0.9$ are shown in Fig. 5.

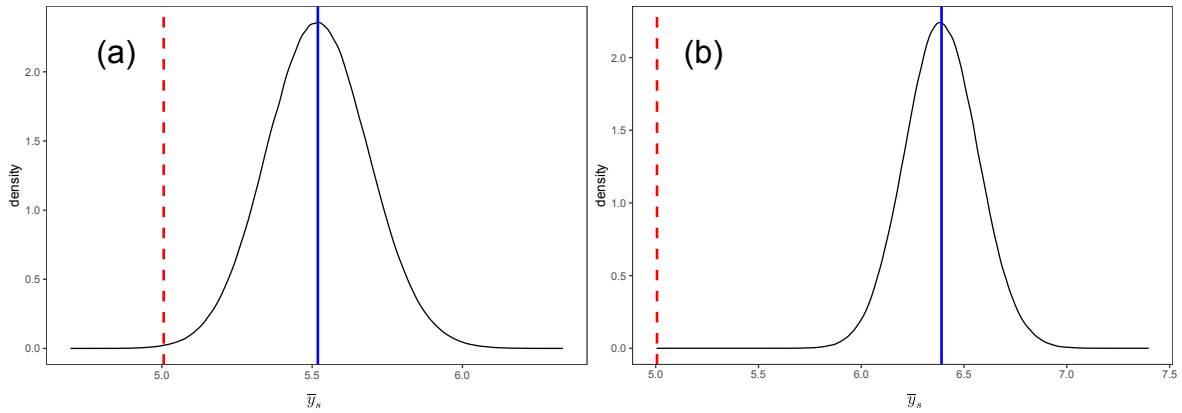


Figure 5: Approximate sampling distributions using 10^6 replications of the POISSWOR model with $P = 500$. (a) $R_{\pi y} = 0.5$. (b) $R_{\pi y} = 0.9$. The population mean is shown by the red dashed line. The average of the values taken by $\overline{y}_s$ is shown by the blue line.

In the case of POISSWOR, the p-unbiased estimator of the population mean is the weighted estimator $\overline{y}_\pi$, which is different from the sample mean $\overline{y}_s$. Therefore, like with BERNWOR, the estimation of the population mean using the sample mean is p-biased. However, unlike the BERNWOR case, this should not be neglected since the relative p-bias is approximately 10.2% for $R_{\pi y} = 0.5$ and 27.6% for $R_{\pi y} = 0.9$ (Fig. 5). The default estimator for the p-variance of $\overline{y}_s$ (Eq. 1) is positively biased since we obtain BI ≈ 1.1411 for $R_{\pi y} = 0.5$ and BI ≈ 1.1788 for $R_{\pi y} = 0.9$.

4.3.4. Preferential sampling — fixed size (CPS)

The sampling process model corresponding to fixed-size preferential sampling is CPS. The $\boldsymbol{\pi}$-vector is the same as that for POISSWOR. The p-distributions of $\overline{y}_s$ for $R_{\pi y} = 0.5$ and $R_{\pi y} = 0.9$ are shown in Fig. 6.

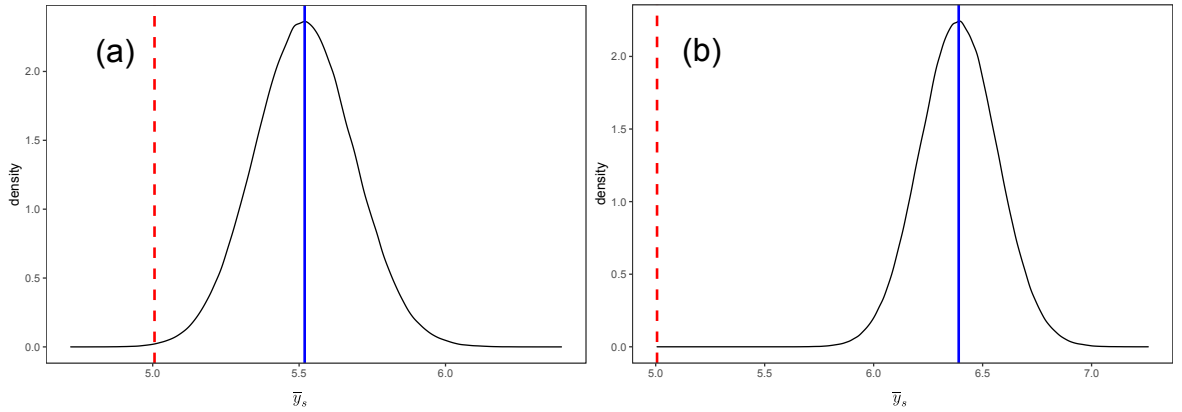


Figure 6: Approximate sampling distributions using 10^6 replications of the CPS model for $P = 500$. (a) $R_{\pi y} = 0.5$. (b) $R_{\pi y} = 0.9$. The population mean is shown by the red dashed line. The average of the values taken by $\overline{y}_s$ is shown by the blue line.

In our example, the case of CPS is very similar to that of POISSWOR since the relative p-bias of the sample mean is approximately 10.2% for $R_{\pi y} = 0.5$ and 27.6% for $R_{\pi y} = 0.9$ (Fig. 6). The default estimator for the p-variance of $\overline{y}_s$ (Eq. 1) is also positively biased since we obtain BI ≈ 1.1387 for $R_{\pi y} = 0.5$ and BI ≈ 1.1793 for $R_{\pi y} = 0.9$, values that are close to those obtained in the POISSWOR case.

4.3.5. Zero-correlation between the propensity and the variable of interest

Thus far, we have simulated preferential sampling for $R_{\pi y} = 0.5$ and $R_{\pi y} = 0.9$. In this section, we examine the degenerate situation where $R_{\pi y} = 0$.

When the correlation between the propensities and y-values is zero, the relative p-bias of the sample mean is essentially zero for both POISSWOR and CPS (Fig. 7). The default estimator for the p-variance of $\overline{y}_s$ (Eq. 1) is positively biased, with BI ≈ 1.1223 for the POISSWOR and BI ≈ 1.1201 for

the CPS. Although the propensities are variable, sampling is nonpreferential (Fig. 7). The situation is broadly equivalent to that encountered with the BERNWOR and SRSWOR models (Fig. 4).

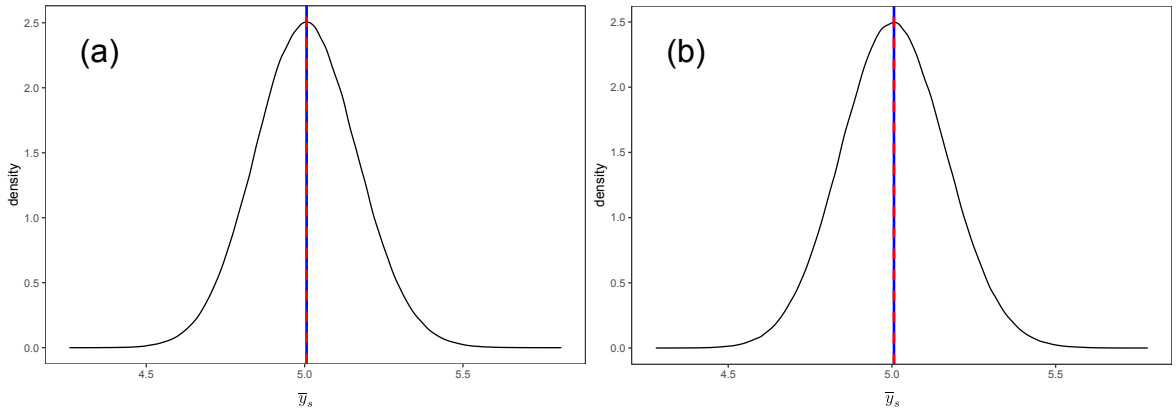


Figure 7: Approximate sampling distributions using 10^6 replications of each unequal propensity sampling model for $P = 500$ and $R_{\pi y} = 0$. (a) POISWOR. (b) CPS. The population mean is shown by the red dashed line. The average of the values taken by $\overline{y}_s$ is shown by the blue line.

4.4. An analytical formula for the bias of the sample mean

In this section, we focus on the formal expression of the p-bias of the sample mean $\overline{y}_s$ for estimating the population mean $\overline{y}_{\mathcal{U}}$. For algebraic details, the reader is referred to Appendix B.

For a sampling design of variable size, the p-bias of the sample mean can be written as (Appendix B.1):

$$\mathrm{B}_p\left(\overline{y}_s\right) = \mathrm{E}_p\left(\overline{y}_s\right) - \overline{y}_{\mathcal{U}} \approx \frac{R_{\pi y}\, S_\pi S_y}{\overline{\pi}_{\mathcal{U}}} \tag{11}$$

where $R_{\pi y} S_\pi S_y = S_{\pi y}$ is the (adjusted) covariance between $\boldsymbol{\pi}$ and $\mathbf{y}$ (see Appendix A).

In our example, we have $\overline{y}_{\mathcal{U}} \approx 5$, $S_y \approx 3.7787$, $N = 5\,000$ and $\overline{\pi}_{\mathcal{U}} = 0.1$. For $R_{\pi y} = 0.5$, we have $S_\pi \approx 0.0272$, from which we obtain $\mathrm{B}_p\left(\overline{y}_s\right) \approx 0.51$ and a relative p-bias of approximately 10.2%, in agreement with what we obtained via Monte Carlo simulation in the POISSWOR case (Section 4.3.3). For $R_{\pi y} = 0.9$, we have $S_\pi \approx 0.0407$, from which we obtain $\mathrm{B}_p\left(\overline{y}_s\right) \approx 1.38$ and a relative p-bias of approximately 27.6%, which again matches the value found earlier by Monte Carlo simulation.

In the case of a fixed-size sampling design, the p-bias of the sample mean admits an exact expression (Eq. B.11, Appendix B.2), but we can also use expression (11) as a close approximation when N is not too small. Consequently, we obtain the same results for CPS as for POISSWOR.

As expected, the observations in the cases studied by Monte Carlo simulation match the results computed by using formula (11). In the context of Monte Carlo simulation, when we are interested in the p-bias of the sample mean, this formula allows us to avoid replicating a sampling process model. We use this computational shortcut when appropriate throughout the remainder of this article.

Examination of formula (11) shows that the sample mean is a p-unbiased estimator of the population mean in the following two situations: (i) when the propensities are variable ($S_\pi > 0$), if there is no correlation between the propensities and the y-values ($R_{\pi y} = 0$) or (ii) when propensities do not vary ($S_\pi = 0$), since then the covariance and correlation are zero ($S_\pi = 0 \Rightarrow S_{\pi y} = 0 \Rightarrow R_{\pi y} = 0$). If the covariance between the propensities and the y-values remains unchanged (fixed $S_{\pi y}$), the lower the sampling fraction (fixed sample size) or its expectation (variable sample size) $\overline{\pi}_{\mathcal{U}}$ is, the greater the bias. As expected, the p-bias is zero in the case of a census since the propensities do not vary (and are all equal to 1), which is the same as in case (ii) above.

5. Trend assessment

For operational purposes, a trend can be defined as a pattern of change in the indicator of interest over time. In the following, we consider a monitoring program designed to study the trend of a population parameter over a period Δ. The monitoring program starts at time t_0 and ends at time $t_0 + \Delta$.

5.1. Trend example

As an example of a trend to be estimated, we consider here the trend in the abundance of a species over a bounded domain $\mathcal{D}$ of the Euclidean plane, partitioned by a population $\mathcal{U}$ of subdomains d_i $(i = 1, 2, \ldots, N)$, formally:

$$\mathcal{D} = \bigcup_{i \in \mathcal{U}} d_i \quad \text{with } d_i \cap d_j = \varnothing \text{ for } i \neq j \in \mathcal{U} \quad \text{and} \quad |\mathcal{D}| = \sum_{i \in \mathcal{U}} |d_i| \tag{12}$$

In a general sense, subdomains can have different shapes and sizes. To simplify the presentation, in what follows, we consider the case where $\mathcal{D}$ is an iso-oriented rectangular domain (i.e., its sides are parallel to the abscissa and ordinate axes), lying between $x_{\min}$ and $x_{\max}$ on the abscissa and between $y_{\min}$ and $y_{\max}$ on the ordinate. With such a domain, subdomains can simply be square cells of equal size. Thus, $\mathcal{D}$ is discretized into a spatial population of $NX \times NY = N$ grid cells of equal size, where NX and NY are the numbers of columns and rows in the grid, respectively.

For the species of interest, at time t (omitted from the notation for simplicity), an individual has a certain probability density $f(u) > 0$ of being present at a point $u \in \mathcal{D}$ with Cartesian coordinates (x_u, y_u). The value of the variable of interest y_i is the number of individuals present in grid cell d_i $(i \in \mathcal{U})$, i.e., the number of points u where the species is present such that $u \in d_i$ (here, a point represents only one individual).

When the size N of the sampled spatial population $\mathcal{U}$ is known, it is the same, to within a factor — i.e., N for the point estimator and N^2 for its p-variance — to estimate the mean or the total abundance. For the concrete situation of preferential sampling that we address in this article (Section 2.4), we consider that N is unknown to the sampler. Thus, in what follows, the parameter of interest is the population mean $\overline{y}_{\mathcal{U}}$, estimated by the sample mean $\overline{y}_s$, as in Section 4. We assume that the spatial variability of the variable of interest is greater on the abscissa than on the ordinate and that at time t_0, there is a spatial gradient in abundance from $x_{\min}$ to $x_{\max}$.

5.2. Population model example

5.2.1. Simulating an inhomogeneous binomial point process

The spatial distribution of a fixed number of individuals M over $\mathcal{D}$ can be modeled by an inhomogeneous binomial point process (abbreviated as IBPP) with a spatial intensity function $\lambda(u)$. Since a binomial point process is a Poisson point process conditional on M, simulation of an IBPP can be performed as for an inhomogeneous Poisson point process using the Lewis-Shedler method (see, e.g., Illian et al., 2008, p. 119; Baddeley et al., 2016, Sec. 5.4.2). An IBPP is simulated as follows:

1. Generate a point $u \in \mathcal{D}$ following a Bernoulli point process.
2. The probability of keeping this point is computed as $p(u) = \lambda(u)/\lambda^*$, where λ^* is the maximum value of the intensity on $\mathcal{D}$:
$$\lambda^* = \max_{u \in \mathcal{D}} \lambda(u)$$
3. A Bernoulli trial is performed with probability $p(u)$. If the trial is successful, then the point u is kept.
4. Steps 1 to 3 are repeated until M points are drawn.

From a realization of the IBPP, the $\mathbf{y}$-vector is obtained by counting the number of points within each grid cell d_i $(i \in \mathcal{U})$. Simulating the realizations of the variable of interest directly by allocating the M individuals among the N grid cells is equivalent to doing so. First, for $i = 1, 2, \ldots, N$, the probability p_i of drawing grid cell d_i is computed as follows (we recall that the grid cells form a partition of $\mathcal{D}$):

$$p_i = \frac{\displaystyle\int_{d_i} \lambda(u)du}{\displaystyle\sum_{i \in \mathcal{U}} \int_{d_i} \lambda(u)du} = \frac{\displaystyle\int_{d_i} \lambda(u)du}{\displaystyle\int_{\mathcal{D}} \lambda(u)du} \tag{13}$$

In our case, for all $i \in \mathcal{U}$, we have an area $|d_i|$ (i) that is very small compared to area $|\mathcal{D}|$ and (ii) that is a constant (the square grid cells have the same area). Therefore, in two steps, we obtain the following approximation:

$$p_i \approx \frac{\lambda(u_i)\,|d_i|}{\sum_{i \in \mathcal{U}} \lambda(u_i)\,|d_i|} = \frac{\lambda(u_i)}{\sum_{i \in \mathcal{U}} \lambda(u_i)} \tag{14}$$

where u_i is the barycenter of grid cell d_i.
The M individuals are allocated as follows:

1. Initialize $y_i \leftarrow 0$ for $i = 1, 2, \ldots, N$.
2. Select a grid cell of index j by unequal probability sampling with replacement (also known as *multinomial sampling*) (e.g., Johnson et al., 1997, Sec. 8; Aubry, 2023, Remark 6) with drawing probabilities p_i $(i = 1, 2, \ldots, N)$.
3. Increment $y_j \leftarrow y_j + 1$.
4. Steps 2 and 3 are repeated until M grid cells are drawn.

5.2.2. Spatial intensity function

As mentioned above, we assume that the spatial distribution of individuals over $\mathcal{D}$ is governed by a spatial intensity function $\lambda(x)$ describing a one-dimensional gradient along the x-axis (we now use x instead of x_u to simplify the notation). The shape of this spatial gradient is determined by the shape parameter η (Fig. 8) according to the expression:

$$\lambda(x) = \frac{\lambda^* \exp(\eta\, x')}{\delta} \qquad \text{with} \quad x \in [x_{\min}, x_{\max}] \tag{15}$$

where $x' \in [0, 1]$ is defined as follows:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{16}$$

and

$$\delta = \begin{cases} 1 & \text{if } \eta < 0 \\ \exp(\eta) & \text{otherwise} \end{cases} \tag{17}$$

The degenerate case $\lambda(x) = \lambda^*$ is obtained for $\eta = 0$ (Fig. 8) and corresponds to a homogeneous binomial point process and equiprobable allocation.

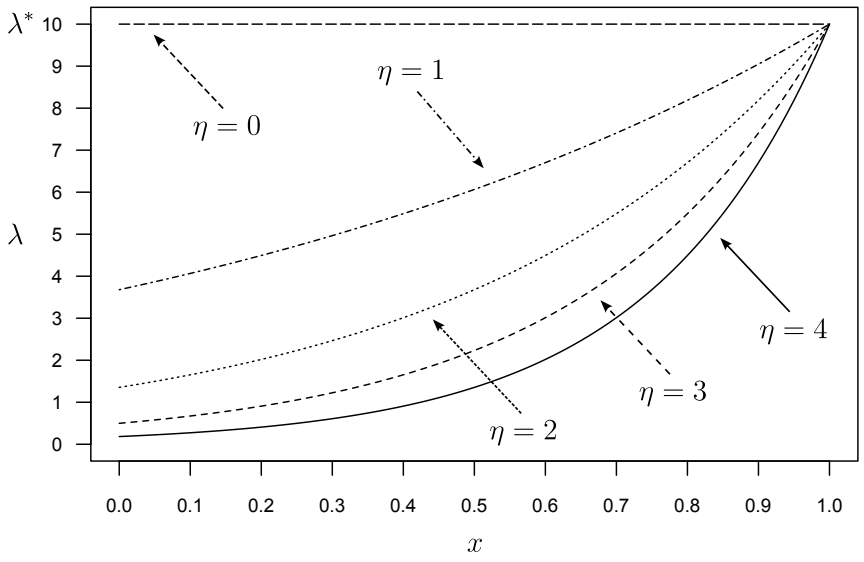


Figure 8: Plot of the intensity function $\lambda(x)$ on the interval $x \in [0, 1]$ with $\lambda^* = 10$, for $\eta = 0, 1, 2, 3, 4$.

We assume that the shape of the spatial gradient changes over time in such a way that its direction is completely reversed at time $t_0+\Delta$, i.e., it is ultimately from $x_{\max}$ to $x_{\min}$. This can be achieved with the shape parameter η (Eq. 15) by gradually decreasing its initial value η_0 at time t_0 until $\eta_\Delta = -\eta_0$ at time $t_0 + \Delta$ (example in Fig. 9).

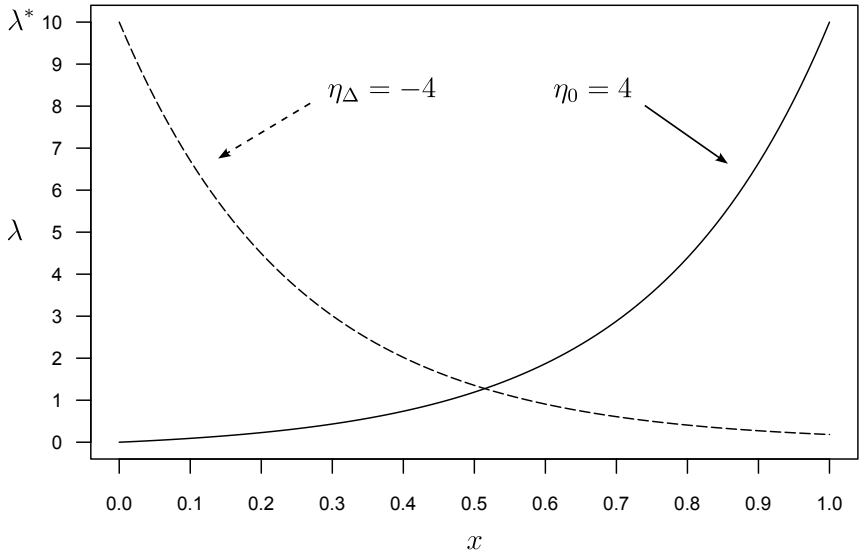


Figure 9: Plot of the intensity function $\lambda(x)$ on the interval $x \in [0, 1]$ with $\lambda^* = 10$, at time t_0 ($\eta_0 = 4$, solid line) and at time $t_0 + \Delta$ ($\eta_\Delta = -4$, dashed line).

5.3. Monte Carlo study

We assume that the number of individuals M in $\mathcal{D}$ remains constant over the monitoring period (no demographic change).

Preferential sampling of fixed size n is performed according to the sampler's view of the variation in abundance over $\mathcal{D}$ at time t_0, favoring the sampling units expected to be the richest in individuals, i.e., those with the highest y-values. As mentioned in Section 3, this type of preferential sampling process can be modeled by conditional Poisson sampling (CPS). Under preferential sampling, we have shown that using the sample mean biases the mean abundance estimation over $\mathcal{D}$ (Section 4). This is not a problem for assessing trends or changes between two points in time, as long as the bias remains (approximately) constant over the monitoring period, which here implies that the units that were expected to be the richest in individuals at time t_0 remain so between t_0 and $t_0 + \Delta$. Conversely, let us consider a directional change in the spatial distribution of individuals in $\mathcal{D}$ over the monitoring period. In this situation, using the sample mean to estimate the population mean results in estimating a spurious trend.

In our example, at the start of the monitoring program, the sampler knows that there is a spatial gradient in abundance from $x_{\min}$ to $x_{\max}$ but does not know its exact shape. By default, we simulate this situation by using a size variable expressed as $z_i = a \times x_i + b$ with (a, b) such that $z_i > 0$ for all $i \in \mathcal{U}$ to guarantee a probability $\pi_i > 0$ of being part of the sample.

We thus have (i) a fixed vector of propensities ($\boldsymbol{\pi}$) computed from the size variable ($\mathbf{z}$) (Eq. 8, Fig. 10, top panel) and (ii) a random vector of abundances ($\mathbf{y}$) obtained according to the intensity function $\lambda(x)$ (Eq. 15, Fig. 10, bottom panel). For a given realization of the spatial variation model defined by $\lambda(x)$, the correlation between the propensities and the abundances takes a certain fixed value $R_{\pi y}$. The model can generate an infinite number of $\mathbf{y}$-vectors, hence an infinite number of $R_{\pi y}$-values (within the range of variation of $R_{\pi y}$ for this model). If the population size is sufficiently large, for convenience, the correlation between the propensities and the abundances in the model ($\rho_{\pi y}$) can be assimilated to the ξ-expectation of the population correlation $R_{\pi y}$ (Appendix A, Fig. 10):

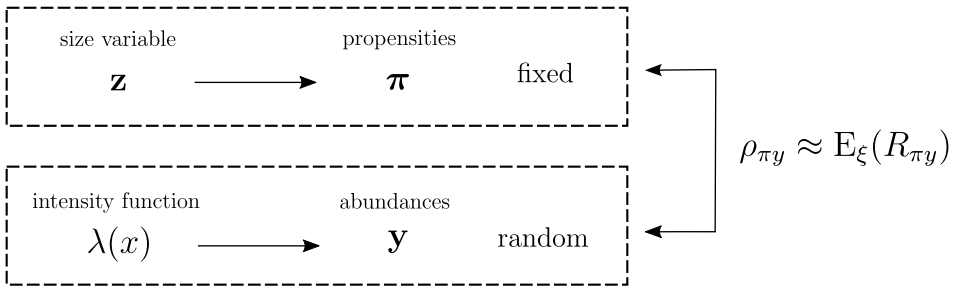


Figure 10: Schematic representation of the relationships between the simulation components. Top panel: The size variable ($\mathbf{z}$) determines the propensities ($\boldsymbol{\pi}$), which remain fixed. Bottom panel: The intensity function of the inhomogeneous binomial point process ($\lambda(x)$) determines the vector of abundances ($\mathbf{y}$), which is random. If the population size is sufficiently large, for convenience, the correlation in the model ($\rho_{\pi y}$) can be assimilated to the ξ-expectation of the finite population correlation $R_{\pi y}$.

In the following, we first consider the case of a fixed population (i.e., a given realization $\mathbf{y}$) and then the case of the superpopulation model itself. We decrease the shape parameter η to simulate the reversal of the spatial gradient in abundance over $\mathcal{D}$ between t_0 and $t_0 + \Delta$, where we assume that η decreases linearly with time. For simplicity, we decrease η by one unit per time unit (e.g., per year or decade). By varying the shape parameter from $\eta_0 = 4$ at time t_0 to $\eta_\Delta = -4$ at time $t_0 + \Delta$ (Fig. 9), we cover a monitoring period of $\Delta = 8$ time units and have a (short) time series of 9 mean abundance estimates to fit a (parametric) temporal trend model.

We simulate a simple situation where an iso-oriented rectangular domain $\mathcal{D}$ is discretized by $N = 2\,500$ square cells organized according to a grid of $NX = 100$ columns and $NY = 25$ rows. For simplicity, we set $x_{\min} = y_{\min} = 0$, $x_{\max} = 100$ and $y_{\max} = 25$ so that each grid cell has a unit area. The maximum intensity is $\lambda^* = 10$ (the number of points per unit area, i.e., also per grid cell).

Regardless of the value of η, $M = 10\,000$ individuals are randomly allocated among the $N = 2\,500$ grid cells, as explained in Section 5.2. Therefore, we have $\overline{y}_{\mathcal{U}} = M/N = 4$. To define the size variable z, we choose $a = b = 0.1$, that is, $z_i = 0.1 \times x_i + 0.1$ ($i \in \mathcal{U}$). We again use a sampling fraction of 10%, i.e., $n = 250$. The variable propensities that form the $\boldsymbol{\pi}$-vector result from the previous choices regarding z and n; they remain constant throughout the simulations (Fig. 10, top panel).

5.3.1. Fixed population

Like in Section 4, for each value of η, we keep a $\mathbf{y}$-vector of abundances for which the correlation $R_{\pi y}$ is approximately equal to a value of $\rho_{\pi y}$ compatible with the value of η. For each case, we generate realizations of the $\mathbf{y}$-vector until we obtain one for which we have $|R_{\pi y} - \rho_{\pi y}| < 10^{-4}$. We take $\eta = 4, 1, 0, -1, -4$. Compatible correlation values are $\rho_{\pi y} = 0.8, 0.5, 0.0, -0.5, -0.8$.

In the case of fixed populations, we are interested in the p-distribution of the sample mean ($\overline{y}_s$) and in the p-bias in particular. We replicate 10^6 times the conditional Poisson sampling (CPS) based on the $\boldsymbol{\pi}$-vector to accurately approximate the p-distribution of $\overline{y}_s$. Several samples obtained by CPS are shown in Fig. 11. As expected for a (putative) linear gradient in abundance from $x_{\min}$ to $x_{\max}$ and preferential sampling, we see a greater spatial concentration of selected units toward the higher abscissas and a unit deficit toward the lower abscissas, with a continuum of unit densities in between.

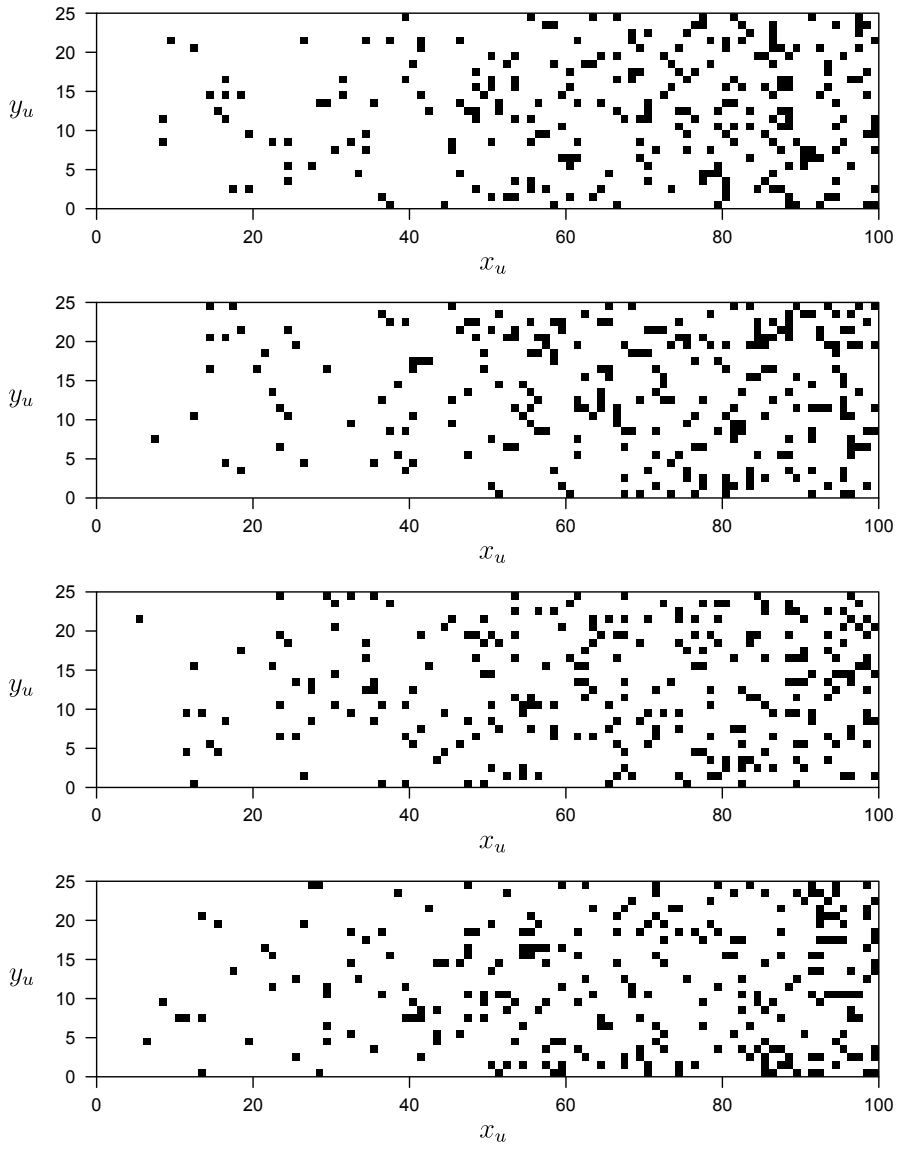


Figure 11: Several samples selected by CPS with inclusion probabilities proportional to a size variable z that varies linearly according to the equation $z_i = 0.1 \times x_i + 0.1$ ($i \in \mathcal{U}$) and for $n = 250$. See the text for details.

In agreement with the analytical formula for the bias of the sample mean (Eq. 11), it appears that the bias is initially positive for $R_{\pi y} > 0$, vanishes for $R_{\pi y} = 0$ and then becomes negative for $R_{\pi y} < 0$ (Fig. 12). Sampling that planned to be preferential at the beginning of the monitoring program becomes nonpreferential and eventually antipreferential due to spatial gradient reversal.

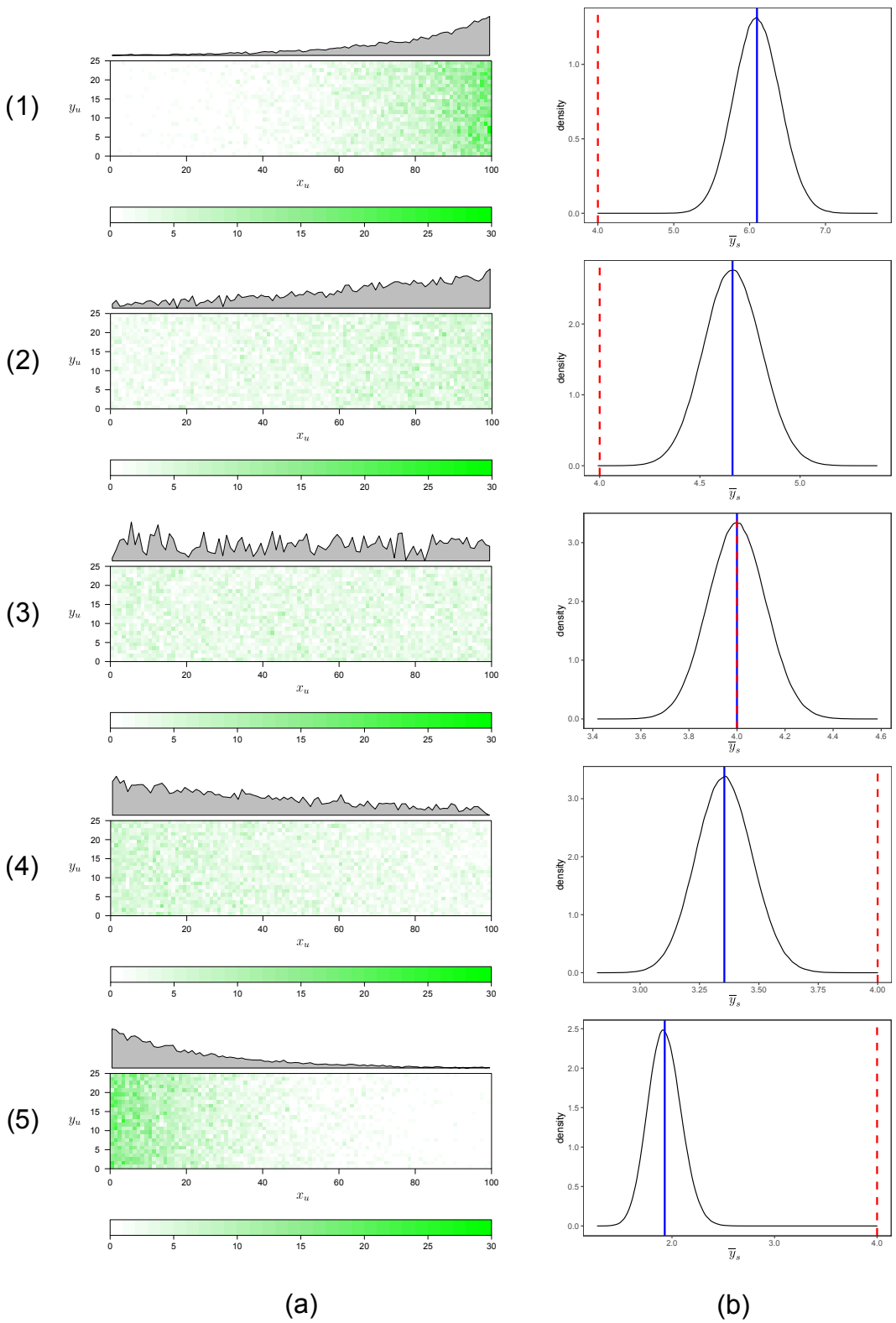


Figure 12: Results of the Monte Carlo study for the five fixed populations. (1) $\eta = 4$, $R_{\pi y} = 0.8$. (2) $\eta = 1$, $R_{\pi y} = 0.5$. (3) $\eta = 0$, $R_{\pi y} = 0$. (4) $\eta = -1$, $R_{\pi y} = -0.5$. (5) $\eta = -4$, $R_{\pi y} = -0.8$. (a) Finite populations simulated on a 100×25 grid as a function of the spatial variation model for the set of η-values. The marginal graph (in gray) represents the mean abundance in the grid cells along the x-axis. This reflects the realization of the spatial variation model. (b) Approximations of the p-distributions of the sample mean ($\overline{y}_s$); 10^6 samples are generated by conditional Poisson sampling. The finite population mean is represented by the red dashed line. The average of the values taken by $\overline{y}_s$ is shown by the blue line. The figures are adjusted to maximize visibility, so the scales may differ from one figure to another. See the text for details.

In practice, a sample of spatial units s obtained at time t_0 may remain unchanged over the monitoring period (permanent plots, fixed monitoring network). In the following, we assume that this is indeed the case. In our example of a reversal spatial gradient, with simulated preferential sampling and by using the sample mean $\overline{y}_s$ as an estimator of the population mean $\overline{y}_{\mathcal{U}}$, there is a serious risk of systematically highlighting a downward trend, while in fact, the abundance in $\mathcal{D}$ remains constant over the monitoring period (no demographic change).

To investigate this hypothesis, at time t_0, we simulate 10^4 samples using CPS. We compute $\overline{y}_s$ for each sample and for each of the abundance vectors at the times corresponding to the different values of η. This gives 10^4 time series of mean abundance estimates. We add four more η values than those used above, with $\eta = 3, 2, -2, -3$ for which compatible correlation values are $\rho_{\pi y} = 0.79, 0.72, -0.72, -0.79$. With a time series of 9 mean abundance estimates at hand (for $\eta = 4, 3, 2, 1, 0, -1, -2, -3, -4$), we fit a linear regression model as a function of time. We use weighted least squares regression (see, e.g., Press et al., 1989, Sec. 14.2) to account for the p-variance estimates. For each sample, we thus obtain a straight line with slope m, and for all the 10^4 samples generated by CPS, we obtain an envelope of straight lines reflecting the sampling variability of the trend highlighted by the monitoring program. As expected, in our example, a downward trend is obtained in all the cases, with an average slope of $\mathrm{E}_p(m) \approx -0.56$ (Fig. 13).

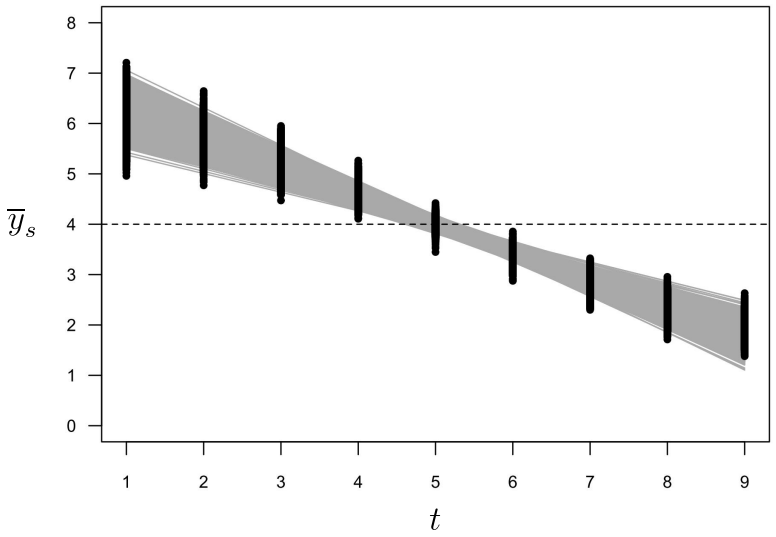


Figure 13: Weighted least squares adjusted linear trend model as a function of time (gray line) for the 9 estimates of the population mean by the sample mean ($\overline{y}_s$) (black dots). The results for 10^4 samples. See the text for details.

5.3.2. Superpopulation

In this section, we document the bias of the sample mean $\overline{y}_s$ in the context of both the preferential sampling process modeled by CPS and the process of spatial variation in abundance. The ξp-expectation of $\overline{y}_s$ is written as follows (we recall that the notation indicates the two sources of stochasticity involved, with subscript ξ for the superpopulation model and subscript p for the sampling process):

$$\mathrm{E}_{\xi p}(\overline{y}_s) = \mathrm{E}_{\xi}(\mathrm{E}_p(\overline{y}_s)) \tag{18}$$

In our model, the population mean $\overline{y}_{\mathcal{U}}$ is constant because we always allocate a fixed number of M individuals among a fixed number of N units (no demographic change). The ξp-bias is therefore:

$$\mathrm{B}_{\xi p}(\overline{y}_s) = \mathrm{E}_{\xi p}(\overline{y}_s) - \overline{y}_{\mathcal{U}} \tag{19}$$

and the relative ξp-bias is $\mathrm{B}_{\xi p}(\overline{y}_s)/\overline{y}_{\mathcal{U}}$. The p-expectation of $\overline{y}_s$ can be expressed by using the p-bias expression, which gives:

$$\mathrm{B}_{\xi p}(\overline{y}_s) = \mathrm{E}_{\xi}[\mathrm{B}_p(\overline{y}_s) + \overline{y}_{\mathcal{U}}] - \overline{y}_{\mathcal{U}} \tag{20}$$

If we are interested only in the value of the ξp-bias (or the relative ξp-bias), as mentioned in Section 4.4, we do not need to replicate the sampling process in our simulation since we know how to compute the p-bias analytically (Eq. 11 or Eq. B.11). However, it is necessary to simulate $\mathbf{y}$-vector realizations to approximate the ξ-expectation that appears in the right term of expression (20).

In the following, we simulate 10^4 $\mathbf{y}$-vectors for each η-value varying between $\eta = 4$ (spatial gradient in abundance from $x_{\min}$ to $x_{\max}$) and $\eta = -4$ (gradient from $x_{\max}$ to $x_{\min}$) with a step of 0.25. We can also compute the correlation in the model — assimilated here for simplicity to the ξ-expectation of the population correlation — for each simulated situation. On average, under the model, the population correlation varies between approximately 0.8 for $\eta = 4$ at time $t = 1$ and -0.8 for $\eta = -4$ at time $t = 9$, with $\rho_{\pi y} = 0$ for $\eta = 0$ at time $t = 5$ (Fig. 14.a). The relative ξp-bias ranges from approximately -52.7% for $\eta = -4$ at time $t = 9$ to 52.7% for $\eta = 4$ at time $t = 1$ (Fig. 14.b). As a result, a preferential sample set up at the beginning of the monitoring program — as explained above — results in a downward trend (approximately linear) on average under the model, while the number of individuals did not actually change over the Δ period, but the spatial gradient was reversed.

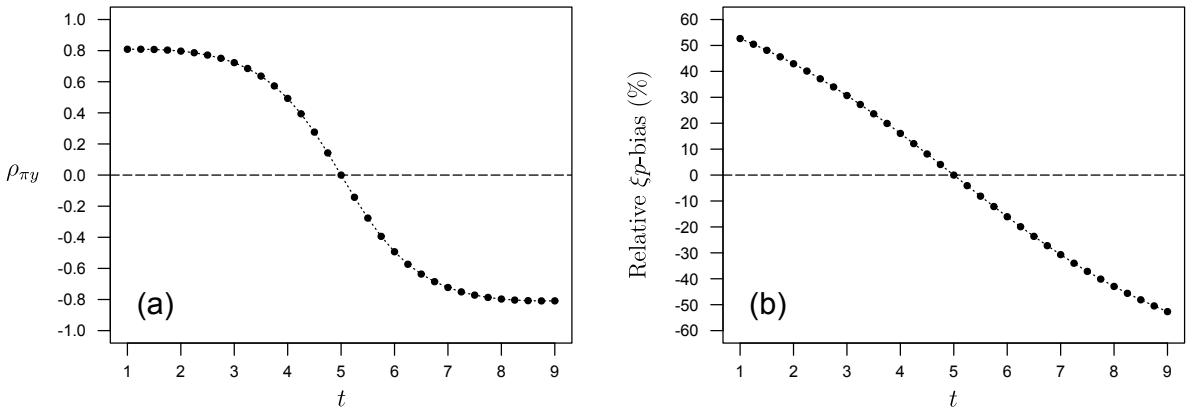


Figure 14: Approximation by Monte Carlo simulation (10^4 realizations of the model) of the relation with time (t) of (a) the ξ-expectation of the population correlation between the propensity and the variable of interest and (b) the relative ξp-bias of the sample mean.

6. Discussion

To date, the issue of nonprobability preferential sampling (as defined in Section 2.2) has not received sufficient attention in ecology. This may mean that most ecologists do not have a clear idea of the implications of this type of nonprobability sampling for their studies and monitoring programs. This observation led us to examine the issue in detail from a methodological perspective, through the example of a program aimed at estimating the status or trend of a finite population parameter.

First, we formalized the basic instances of without-replacement sampling processes by modeling them in terms of probability sampling designs. While we have applied this approach to spatial sampling, its versatility extends to temporal sampling, where the issue of preferential sampling may also arise (for example, migratory bird surveys are typically conducted at specific times of the year when bird migration is expected to peak).

Next, we used Monte Carlo simulations to study the bias of the sample mean as an estimator of the population mean, first in the case of status assessment and second in the case of temporal trend assessment. In the case of estimating the status of a population parameter (e.g., mean species richness), we illustrated the risk of preferential sampling in terms of bias in the estimation of the population mean since the relative bias may be unacceptable (e.g., greater than 10%), thereby leading to erroneous conclusions. In addition, we showed that the estimation of the sampling variance was also biased when simply using the formula for the variance of the mean s_y^2/n. We also noticed that the fact that the sampling process results in fixed or variable sample sizes did not appreciably alter the results obtained and is therefore not in itself a cause for concern. Thus, treating the sample size as fixed and estimating parameters conditional on its value is a legitimate practice, at least for the question with which we are concerned in this article, for a sufficiently large population size and moderate sampling fraction. In the case of estimating the temporal trend of a population parameter such as mean abundance, the simulated example shows that using preferential sampling at the beginning of the program to establish permanent plots or sites can lead to the identification of a downward demographic trend when there is a directional change in the spatial distribution of individuals, even though abundance is actually constant over the monitoring period.

Hence, we clearly demonstrated that, without knowledge of the inclusion probabilities, the sampler lacks awareness of the magnitude of the potential biases in either the sample mean or its sampling variance. The simulated case for the trend assessment is a clear example of *time-varying spatial bias* (Johnston et al., 2023, Sec. 3.1). The type of erroneous conclusions reached in this case has already been clearly stated by Fournier et al. (2019) for abundance and Palmer (1993) for community heterogeneity, but these authors referred to the phenomenon of *regression to the mean*, which differs from what we have considered in this article. In a different ecological scenario than the one we considered, McClure and Rolek (2023) noted that preferential sampling may delay the detection of a downward trend. For biological conservation purposes, the common conclusion is that preferential sampling can have a profound impact on trend estimation, either by highlighting a downward trend where none exists (Type I statistical error) or by failing to detect (in time) a downward trend that actually exists (Type II statistical error, insufficient statistical power). These two possibilities should be considered when evaluating the statistical power of a nonprobability preferential sampling monitoring program designed to detect downward trends.

In this article, a new perspective on biased site selection was introduced to quantitative ecology. We have modeled preferential sampling processes using probability sampling designs for the purpose of Monte Carlo simulation and also analytical formulation of the estimation bias. For status and trend assessment, we caution against the use of preferential sampling in a nonprobability framework. We showed that the bias in estimating the mean increases with the covariance between the propensity and the variable of interest (species richness, abundance, etc.) and decreases with increasing sampling effort. This is a simple but fundamental result that deserves greater recognition. In the following sections, we take a closer look at the discussion points that emerge from our article.

6.1. The scope of the simulated examples

The simulated examples in our article are intended to make our findings as clear as possible. Some objections can be raised: (1) real ecological studies and programs are not based on preferential sampling, so the problem documented in this article does not occur in practice; (2) the simulated situations bear no relation to reality.

Regarding point (1), it is clear that many programs actually use preferential sampling, at least implicitly. Observers have a general tendency to select sites that are species rich (with rare species) and/or high in abundance rather than opting for sites that may initially seem less interesting for naturalistic observations. This behavior applies to both flora and fauna data collection (see, e.g., Chytrý, 2001; Lepš and Šmilauer, 2007; Diekmann et al., 2007; Conn et al., 2017; Bowler et al., 2022). Many examples can be provided from the ornithological world, given birds are among the best studied of all animal groups. In the case of Galliforms, for example, a comprehensive review of data sources on a global scale shows that, with the exception of atlases, data tend to be collected preferentially from sites visited by tourists and bird specialists (biodiversity hotspots), while areas with few rare species or protected areas are neglected (Boakes et al., 2010, p. 5). This results in a spatial coverage problem typical of preferential sampling but which also involves convenience sampling because of access difficulties (for logistical or political reasons). Another example is the *Dutch Breeding Bird Monitoring Program* (BMP), which again relies on both convenience and preferential sampling since the observers are free to choose their study areas, and in each habitat, they may prefer the most attractive sites, i.e., those that are species-rich and have high bird densities (van Turnhout et al., 2008). An example of a program that relies on both purposive and preferential sampling is the *International Waterbird Count* (IWC) — a site-based counting scheme for monitoring waterbird numbers organized at the supranational level by *Wetlands International* — where sites are defined by the judgment of national coordinators and local observers and where decisions about which sites to count are based on their relative importance (Delany, 2010, Sec. 4). These few examples illustrate that programs based on nonprobability sampling of count sites do not fall under a single sampling type but generally involve some degree of preferential sampling.

With respect to point (2), when estimating population trends in the context of global warming, *range shifts* have already occurred in many places for different taxa and can be expressed in terms of latitude, longitude, elevation or depth (see, e.g., McCarty, 2001; Parmesan, 2006, 2019; Lenoir and Svenning, 2013, 2015; Dahms and Killen, 2023 and references cited therein). While real-world situations may not be as extreme as the complete reversal of a spatial gradient in abundance as simulated in this paper, the range shift phenomenon is undeniably real; moreover, its frequency is likely to increase with global warming, as temperature and drought (for terrestrial ecosystems) are the most limiting abiotic factors for many species. For birds, for example, range shifts in wintering areas have been well documented (e.g., Maclean et al., 2008, Lehikoinen and Sparks, 2010; Lehikoinen et al., 2013). For some species, this phenomenon may also act in synergy with a change in predation

pressure (Pokallus and Pauli, 2015). Changes in trophic interactions may contribute to a change in the geographic distribution of individuals. For example, this may be the case for the Eider Duck (*Somateria mollissima*), for which there is a possible shift from open islands to forested islands. Since the islands monitored are mostly open islands, the result can be the observation of a spurious downward trend (Ekroos et al., 2012, Sec. 4.2). Due to the gregarious behavior of migratory birds (for example) and communication between individuals, it is also possible that such shifts occur quickly.

In addition to the examples simulated in this article, as we have previously explained in detail in another context (Aubry et al., 2020, Sec. 5.1), the estimation by the sample mean is biased if the propensities vary significantly and are correlated with the variable of interest (e.g., species richness, abundance). This aligns well with the situation of preferential sampling described in this article. We note that a similar result occurs when estimating the variance of the sampled population, which may be a goal in itself (see Courbois and Urquhart, 2004), a topic not covered in this article.

6.2. The bias of the sample mean

Beyond the results obtained by Monte Carlo simulation, we formally showed that the bias of the sample mean as an estimator of the population mean can be written essentially as the population covariance between the propensities and y-values divided by the mean propensity, i.e., the sampling fraction in the case of a fixed-size sampling process or its expectation in the case of a variable-size sampling process (Appendix B). Although Aubry et al. (2020) and Boyd et al. (2023) have recently mentioned the key role played by the correlation between sample membership and the variable of interest, to our knowledge, the analytical expression for the bias of the sample mean has so far remained unknown to the ecological audience. In this respect, our paper fills a statistical gap in the field of quantitative ecology.

The analytical expression for the bias (or similarly, the relative bias) of the sample mean can also be found in the statistical literature specializing in the treatment of nonresponse (see Kalton and Maligalig, 1991, Eq. 1.3; Särndal et al., 1992, p. 577, Eq. 15.6.3, 15.6. 4; Bethlehem, 1988, Eq. 3.5; Bethlehem, 1999, p. 129; Bethlehem, 2002, p. 276; Särndal and Lundström, 2005, p. 92; Brick and Montaquila, 2009, Eq. 4; Bethlehem, 2009, p. 222; Bethlehem et al., 2011, p. 44; Haziza and Beaumont, 2017, Eq. 3.4). An algebraically equivalent formula, but of less pedagogical interest, is given by Groves et al. (2004b, Appendix) (see also Groves et al., 2004a, p. 182). We would like to draw the reader's attention to the fact that this specialized literature is a methodological resource of utmost interest that should be taken into account in quantitative ecology, as was also recently noted by Chadwick et al. (2024, Box 1).

To summarize, when propensities vary among the sampling units, for a given spatial population and variable of interest, there are three possible cases where using the sample mean may or may not result in a biased estimation of the population mean. We illustrate these three cases with the help of Fig. 15, where the variances of the propensities and y-values do not change between the three situations examined, with the fixed population being that of Fig. 12.1a. If there is a positive correlation between the propensities and y-values (Fig. 15.1a), then the sampling is preferential, and the sample mean bias is positive (overestimation) (Fig. 15.1b). If the correlation is zero, then, although the propensities are variable (Fig. 15.2a), the sampling is nonpreferential, and there is no bias (Fig. 15.2b). If the correlation is negative (Fig. 15.3a), then the sampling is antipreferential, and the bias is negative (underestimation) (Fig. 15.3b).

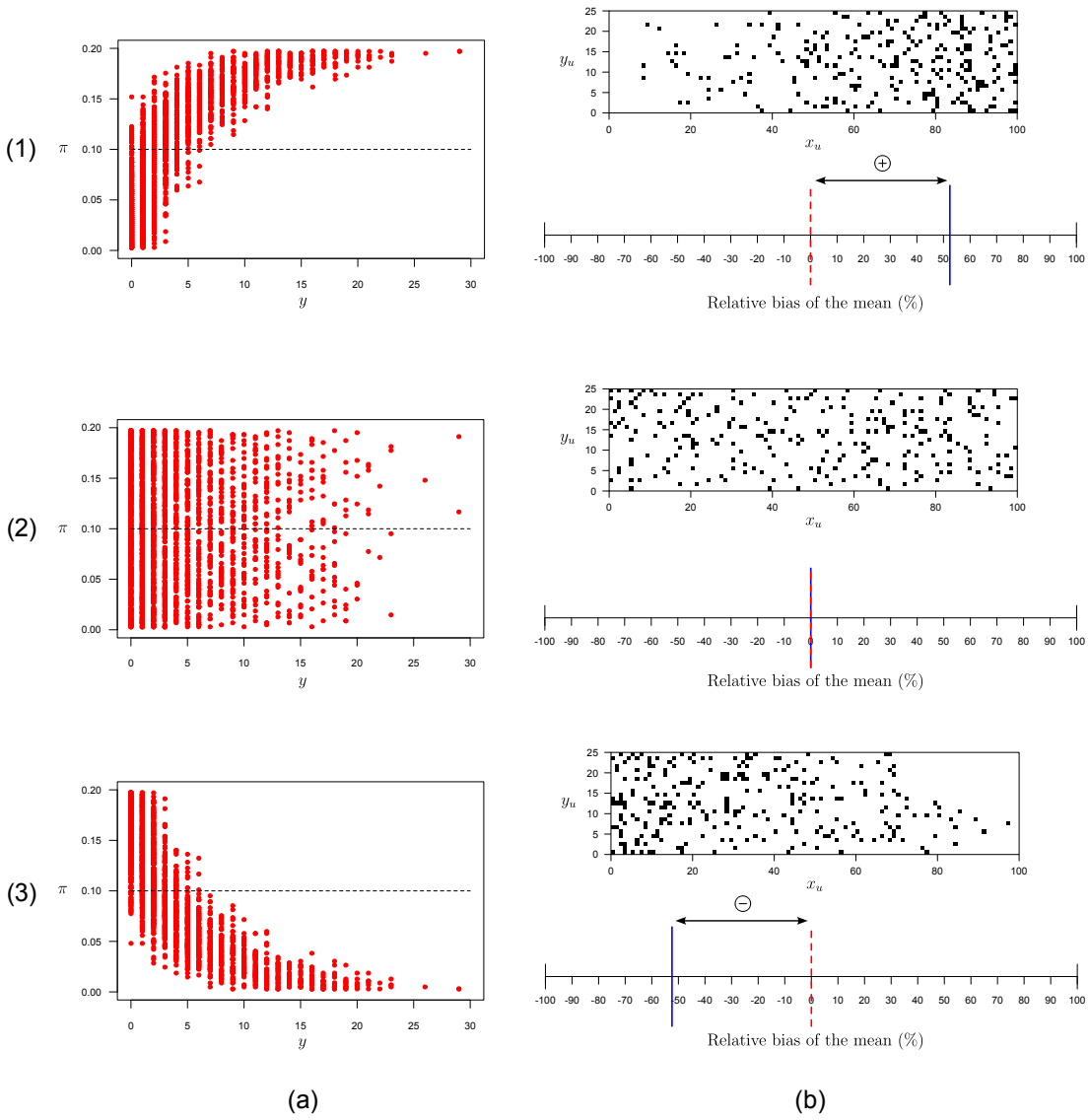


Figure 15: For the fixed population shown in Figure 12.1a, the bias of the sample mean according to the correlation between the propensities and y-values ($R_{\pi y}$). (1) $R_{\pi y} = 0.8$. (2) $R_{\pi y} = 0$. (3) $R_{\pi y} = -0.8$. (a) Scatter plot of the $N = 2\,500$ sampling units according to their propensity π and y-value. (b) Top panel: Example of a sample drawn by conditional Poisson sampling with the corresponding $\boldsymbol{\pi}$-vector. Bottom panel: Relative bias of the sample mean (%) for the corresponding sampling process.

6.3. Reasons for using preferential sampling

Given the risk that preferential sampling poses to the validity of the conclusions drawn from the estimates, we need to ask why it is used. To answer this question, we must first distinguish between two situations, depending on whether the sampler (i) has a sampling frame or (ii) does not.

A sampling frame implies that the size of the spatial population is known. Therefore, it is possible, at least in principle, to use a probability sampling design that allows for the knowledge of inclusion probabilities and their incorporation into the estimators. The sampler can then estimate population parameters within a *design-based* framework that is as objective as possible since there are no assumptions about either the statistical distribution of the variable of interest or its possible spatial structure (spatial trend and autocorrelation), unlike a *model-based* approach (see, e.g., de Gruijter and ter Braak, 1990; Brus and de Gruijter, 1993, 1997; Gregoire, 1998; Dumelle et al., 2022; Aubry and Francesiaz, 2022). In doing so, the sampler is aware of the statistical properties of the chosen estimators. This is a fundamental difference from the concrete situation discussed in this article (see Section 2.4).

There may be several reasons for using preferential sampling within the design-based framework. Usually, the goal is to optimize the efficiency of the sampling strategy — in the sense of a pair formed by a sampling design and an estimator (e.g., Hedayat and Sinha, 1991, p. 24) — by minimizing the p-variance (sampling variance). For a fixed-size sampling design, an examination of the Sen-Yates-Grundy variance formula (e.g., Hedayat and Sinha, 1991, p. 48) — or what leads to the same

conclusion, that of its estimator, used in Eq. (6) — shows that the p-variance of the weighted estimator is zero in the case of exact proportionality ($\pi_i \propto y_i$ for $i \in \mathcal{U}$). In practice, we do not know the y-values, but we may know a size variable z which is positively correlated with y. Thus, we approach the ideal proportionality relation by computing the inclusion probabilities as in Eq. (8). For a variable-size sampling design (i.e., Poisson sampling), a similar result holds (see Särndal et al., 1992, pp. 86-87). In a multispecies (i.e., multivariable) context, another reason may be to maximize the probability of covering the spatial distributions of several species, as recently illustrated by Aubry et al. (2023). The goal is to optimize the sampling effort to collect data for a maximum number of species in a limited number of site visits while remaining within the framework of design-based estimation.

If the sampler lacks a sampling frame, the selection of the sample of spatial units is typically not randomized — at least as we consider the sampling randomization in this article. Even leaving aside the question of randomization, specifying in advance the spatial units to be visited in such programs can be challenging. In these cases, the preferential nature of the sampling process is then no longer dictated by statistical considerations but rather by concerns about the adherence of the observers — often volunteers — to the program. The preferential nature of the sampling process often results from the collective action of observers who tend to select units in a similar manner (see, e.g., ter Steege et al., 2011).

In contrast to preferential sampling, antipreferential sampling may refer to preferential inclusion in the sample of species-poorest units, the units with the lowest abundances or occupancy probabilities. The objectives may be to overrepresent the distribution margins of a species in the sample, to study the effect of buffer zones in the establishment of protected areas or to study the natural recovery of degraded ecosystems.

6.4. Reasons for avoiding preferential sampling in trend assessment

In general, the use of (i) a permanent sample selected by preferential sampling and (ii) the sample mean as an estimator of the population mean is a sampling strategy that should not be followed for trend assessment. We have illustrated the problem of such a strategy that leads to the conclusion that there is a downward trend even when there is no trend at all. This spurious trend is caused by the concomitance of preferential sampling at the beginning of the program and a change in the geographic distribution of individuals during the monitoring period. In this case, if the propensities (inclusion probabilities) had been known — as would have been the case if the fixed-size sample had been obtained by applying an unequal-probability sampling design such as *conditional Poisson sampling* (for example) — then it would have been possible to estimate the mean abundance without bias, as well as the variance of the estimator (see Section 3.5). This would have attenuated the problem of spurious downward trend detection. However, in the simulated case, the sampling variance increases over time as the correlation between the propensities and the y-values decreases, vanishes (null correlation) and then changes sign (negative correlation) (Fig. 14.a). Thus, even if the inclusion probabilities were known and incorporated into the estimators, the strategy used was inappropriate and risky.

When assessing a temporal trend in abundance (for example), using a permanent sample of units selected by preferential sampling at the beginning of the monitoring program can potentially lead to three types of erroneous conclusions: (i) highlighting a downward trend when there is no demographic change (the case illustrated in this article); (ii) attenuating an upward trend (e.g., when there is a colonization of new habitat patches that are almost unrepresented in the sample, without changes in habitat patches previously occupied, which constitute the bulk of the sample); and (iii) exaggerating a downward trend (e.g., in the case of a density-dependent decline, more pronounced in units where there were many individuals initially, which represents the bulk of the sample). In the case of antipreferential sampling, the results are reversed. The only case for which this sampling strategy does not lead to a time-varying spatial bias is when the change in abundance is uniform across the spatial domain under consideration, i.e., when it occurs at the same rate for all sampling units. It is understandable that this case is the exception rather than the rule in ecology.

6.5. Consideration of the sampling process in ecology

Like others (e.g., Elzinga et al., 2001, p. 116; Albert et al., 2010; Smith et al., 2017; Aubry et al., 2020; Boyd et al., 2023), we recommend the use of probability sampling whenever possible to control for inclusion probabilities. This recommendation applies whether the planned inference is design-based or model-based. It is often thought that the use of a probability sampling design is not applicable in ecology, especially on a large scale. While we cannot deny that using probability sampling may indeed be difficult or even unfeasible (see, e.g., Roleček et al., 2007), there are counterexamples that

illustrate that this approach may be possible, at least for certain taxonomic groups and environments (e.g., Pavlacky et al., 2017; van Wilgenburg et al., 2020; Aubry et al., 2023).

When it is not possible to control for inclusion probabilities, it is important to be aware of the source of bias that preferential sampling may represent and to discuss it when communicating the results of a study or program. We agree with Boyd et al. (2023) on this point while recognizing that this is the least one can do. A further step forward is the recognition that statistically sound use of the data obtained by preferential sampling necessarily requires consideration of the sampling process, whether the chosen inferential framework is frequentist or Bayesian. This topic is an active area of research in survey sampling and spatial or ecological statistics, which goes beyond status and trend assessment (see Journel, 1994; Diggle et al., 2010; Pati et al., 2011; Zidek et al., 2014; Pan et al., 2015; Grisotto et al., 2016; Cecconi et al., 2016; Conn et al., 2017; Irvine et al., 2018; Watson et al., 2019; Dinsdale and Salibian-Barrera, 2019; Pennino et al., 2019; da Silva Ferreira, 2020; Olea, 2021; Gray and Evangelou, 2023; Vedensky et al., 2023). Such research is critical for existing ecological programs that were not built on the basis of probability sampling, but addressing the statistical approaches that can be considered *a posteriori* to deal with the problem of bias induced by preferential sampling is beyond the scope of this article. Nevertheless, for a recent introduction of this topic in ecology, we refer the reader to Boyd et al. (2024).

Regarding sampling processes in ecology, the main message of this article can be summarized as follows: "If we do not know the sampling process, we have no idea what we are doing with the data (at least from a statistical perspective)". While this statement may seem self-evident to statisticians, an examination of the ecological literature reveals a disturbing reality: the somewhat widespread neglect of this basic premise. In practice, statistical analysis and modeling are often performed without due consideration of the sampling process. This is not a problem if one is interested only in the data from the sample at hand, but it is a guarantee of poor statistical inference when the results must be extended beyond the sample. We believe that the oversight about the sampling process at work is a major contributor to the misuse of statistical methods in ecology (*analytical crisis*, Chadwick et al., 2024).

6.6. Focusing on the terminology

Scientific discourse requires accurate, monosemous and, ideally, stable and shared terminology (see, e.g., Schuster, 2020, Sec. 2). Unfortunately, there is no consensus on the terminology used to describe preferential sampling or related concepts. Since the lack of common understanding and naming of concepts is a major barrier to communication among researchers, we felt it important to devote part of the discussion in this article to this issue, to ensure proper and operational transfer in quantitative ecology. We fully agree with Hall et al. (1997) that "$[\cdots]$ *if we want to advance* $[\cdots]$ *ecology, we must be sure that the fundamental concepts with which we work are well defined, and hence, well understood*".

The term *preferential sampling* has been used in the literature to denote spatial sampling when it is neither regularly nor randomly distributed across the study area (e.g., Goovaerts, 1997, Sec. 4.1.1), as a synonym for *purposive sampling* of typical units (e.g., Orlóci, 1975, pp. 10, 12; Podani, 1984; Roleček et al., 2007; Swacha et al., 2017), for *convenience sampling*, mainly for ease of access (e.g., Clifford et al., 2011; Mentges et al., 2021). Other distinct meanings can also be found in the literature. For example, Merckx et al. (2011) use the term *preferential sampling* to refer to visiting some sites more frequently than others. In numerous articles, the term is not even defined by the authors who use it, which is problematic because it can mean different things to different people. Moreover, it covers different aspects that need to be distinguished because they may have different statistical consequences, requiring different statistical approaches to be properly handled. In this article, we use the term *preferential sampling* in a precise statistical sense to denote the existence of a nonnegligible correlation between the propensities of the units to be sampled and the values of the variable of interest. It is used in a similar way as McClure and Rolek (2023), except that our definition is more deeply rooted in statistics.

Preferential (or antipreferential) sampling is a special case of what is more generally known in the literature as *biased selection* in the sense that the sampling process produces samples in which some parts of the population are underrepresented while others are overrepresented (Zarkovich, 1966, p. 75). The term *selection bias* may be used by some authors to refer to the same idea (e.g., Eklund, 1959, Sec. 3.2), in the sense of samples or data biased by selection (*sample selection bias* or *selection-biased data*). This terminology refers to one of the usual meanings given to the term *bias*, i.e., that associated with a distortion or deformation (i.e., of a study, a result, a conclusion, etc.). However, the expression *selection bias* can also refer to the technical statistical meaning of the term *bias* when it concerns an estimator (e.g., Kotz et al., 2006, p. 483). We note that biased selection does not necessarily lead to biased estimation (e.g., Overton and Stehman, 1995, Example 6; Aubry et al., 2020, Sec. 5.1), for example,

when using a weighted estimator based on inclusion probabilities (e.g., using the Horvitz-Thompson estimator). Conversely, the absence of biased selection does not guarantee unbiased estimation (e.g., Stuart, 1984, Sec. 6). This terminology can therefore be confusing; the most important thing is to know what one is talking about and to use the terms consistently.

The term *site-selection bias* (e.g., Irvine et al., 2018; Fournier et al., 2019; Mentges et al., 2021) is another synonym for *preferential sampling* in the meaning used in this article, or includes it as a subcase (McClure and Rolek, 2023). The two synonymous terms *response biased sampling* or *response selective sampling* can also be found in the literature to refer to (or include) preferential sampling (e.g., Lawless, 1997; Lawless et al., 1999; Scott and Wild, 2011). However, these two terms seem to be used almost exclusively by (some) statisticians.

By using statistical terminology from the field of missing data (see, e.g., Allison, 2002; Enders, 2010; Molenberghs et al., 2015; Little and Rubin, 2019), whenever sampling is preferential (Fig. 15.1) or antipreferential (Fig. 15.3), the missingness mechanism is said to be *nonignorable*, and the data are said to be *missing not at random* (MNAR; also referred to as *not missing at random* or NMAR, see Little and Rubin, 2019, p. 28, Note 1). If the propensities vary minimally or vary significantly but are not correlated with the variable of interest (Fig. 15.2), then the mechanism is said to be *ignorable*, and the data are said to be *missing completely at random* (MCAR).

Boyd et al. (2023, 2024) use the term *representative* to refer to nonpreferential sampling, i.e., when sample membership is uncorrelated with the variable of interest. Given the already widely polysemous nature of the terms *representative sampling* and *representative sample* (see, e.g., Kruskal and Mosteller, 2006; Bethlehem, 2009, Sec. 2.4.1), this new definition is likely to create more confusion than clarification. Since sampling is a mechanism that produces missing data, it is more appropriate to use the terminology used in that field and simply refer to sampling as *ignorable* or *nonignorable*, as the case may be. When dealing with statistical issues in ecology, it is appropriate for clarity and consistency to refer to the vocabulary used in statistics, as Irvine et al. (2018) do, for example.

7. Perspectives

The points discussed in this article relate to spatial sampling in ecology — that is, sampling of spatial units to study or monitor ecological phenomena — but they are quite general from a statistical perspective and concern broader topics than those covered here (e.g., Aubry et al., 2020). We have used the (seemingly simple) example of status and trend assessment, but the issue of spatial preferential sampling has also been highlighted in other fields (e.g., spatial prediction, Gelfand et al., 2012). We have considered preferential spatial sampling — which is undoubtedly the most obvious instance of preferential sampling in ecology — but temporal sampling should also be accounted for in practice. For example, if the sampling period targets an annual peak in abundance but the phenology of the species of interest is gradually changing with global warming, then there is a serious risk of introducing a time-varying temporal bias, which is the counterpart of the time-varying spatial bias illustrated in this article.

We believe that the analytical formula governing the bias of the sample mean (Eq. 11) is fundamental and should be familiar to ecologists and wildlife biologists. Despite the recognized importance of sampling, it is paradoxical that it receives so little attention overall in quantitative ecology, a field largely dominated by modeling. Even in the case of a model-based approach, sampling issues remain central, as they largely determine the ability of the model fitted to the sample data to reliably estimate or predict the quantities of interest. Further attention and work are needed in this area, as we believe this topic is critical to the credibility of the results published in the ecological literature.

Demonstrating the risk of the sampling strategy documented in this article for trend assessment is undoubtedly useful, but it is even more useful to suggest strategies that are as robust as possible for detecting a trend that is not spurious, not attenuated or not exaggerated to guide the readers in their choices when designing a monitoring program. It is not simply a matter of drawing inspiration from existing programs but of justifying strategies in light of ecological, statistical and operational perspectives. Quoting Greenwood (2003): *"In designing surveys, however, we strike the balance between theoretical robustness and practicality: just as there is no point in running a survey so biased that the data are uninterpretable, there is no point in designing one that is so theoretically perfect that it is impossible to conduct."*. This may be the subject of future articles.

Another perspective concerns the situation in which probability sampling is ruled out (for various reasons). A statistical approach for dealing with nonprobability sampling data is to use a model, be it frequentist or Bayesian. This approach is known as the *model-based* approach. Strictly speaking, the *design-based* approach cannot be used, precisely because the sampling was not conducted by using a probability sampling design. However, a remaining question arises: Just as we have used probability

sampling designs to model sampling processes for simulation purposes, to what extent can they also be used for statistical inference purposes? This approach differs from design-based inference in the strictest sense of the term because, in this case, the inference is actually based on modeling the sampling process using a probability sampling design. This approach may be called *quasi-randomization* in the sense given by Oh and Scheuren (1983) or *pseudo design-based* (Baker et al., 2013). This rather unusual topic in quantitative ecology (see Boyd et al., 2024 for a primer) should be explored in detail in future articles and echoes the question posed by Boyd et al. (2023) " *What other methods for making inferences from nonprobability samples exist, and how reliable are they with real data?*". Such a study can be undertaken concurrently with an examination of the robustness of predictions based on a superpopulation model, a topic recently illustrated by Aubry and Francesiaz (2022).

8. Acknowledgments

We are grateful to Pr. Nigel G. Yoccoz and another anonymous reviewer for their comments, which provided us with the opportunity to improve the article. We thank American Journal Experts (AJE) for the final English language editing.

Appendix A. Notational conventions

We denote the mean of a variable x on a finite set A:

$$\overline{x}_A = \frac{1}{|A|} \sum_{i \in A} x_i$$

where $|A|$ is the size (cardinality) of A.
We denote the *adjusted variance* of x on the finite population $\mathcal{U}$:

$$S_x^2 = \frac{1}{N-1} \sum_{i \in \mathcal{U}} (x_i - \overline{x}_\mathcal{U})^2$$

With this notation, the population variance has $N-1$ as the denominator, following the convention used in survey sampling theory (see Cochran, 1977, p. 23). When the finite population is viewed as randomly drawn from a superpopulation, using $N-1$ as the denominator also makes sense since the population variance is then an unbiased estimator of the superpopulation variance (e.g., O'Neill, 2014, p. 283). Similarly, the (adjusted) sample variance has the denominator $n_s - 1$:

$$s_x^2 = \frac{1}{n_s - 1} \sum_{i \in s} (x_i - \overline{x}_s)^2$$

where n_s is the sample size, which may or may not depend on the realized sample s (variable or fixed sample size).
We denote the (adjusted) covariance between two variables x and y on the finite population $\mathcal{U}$:

$$S_{xy} = \frac{1}{N-1} \sum_{i \in \mathcal{U}} (x_i - \overline{x}_\mathcal{U})(y_i - \overline{y}_\mathcal{U})$$

The correlation between two variables x and y on the finite population $\mathcal{U}$ is defined as $R_{xy} = S_{xy} / [S_x S_y]$.

In a superpopulation model denoted ξ, the ξ-variance of a variable x is noted σ_x^2, and the ξ-covariance between two variables x and y is denoted σ_{xy}. The ξ-correlation between x and y is then defined as $\rho_{xy} = \sigma_{xy} / [\sigma_x \sigma_y]$. This is the definition used in Section 4. Mathematically, R_{xy} is not an unbiased estimator of ρ_{xy}; that is, $\rho_{xy} \neq \mathrm{E}_\xi(R_{xy})$. However, the bias is negligible in the case of a sufficiently large population. In Section 5, for convenience, we use the approximation $\rho_{xy} \approx \mathrm{E}_\xi(R_{xy})$.

Appendix B. Bias of the sample mean

Let a finite population $\mathcal{U}$ of size N be sampled by a without-replacement sampling design with inclusion probabilities $0 < \pi_k \leq 1$ $(k \in \mathcal{U})$. Let s be a sample of size n_s drawn from $\mathcal{U}$. Introducing the indicator variable $I_k = 1$ when unit $k \in \mathcal{U}$ is included in sample s and $I_k = 0$ otherwise, the sample mean is written:

$$\overline{y}_s = \frac{1}{n_s}\sum_{k\in s} y_k = \frac{1}{n_s}\sum_{k\in\mathcal{U}} I_k y_k = \frac{\sum_{k\in\mathcal{U}} I_k y_k}{\sum_{k\in\mathcal{U}} I_k} \tag{B.1}$$

Appendix B.1. Variable-size sampling design

In a variable-size sampling design, the sample size n_s is a random variable of expectation:

$$\mathrm{E}_p\left(n_s\right) = \mathrm{E}_p\left(\sum_{k\in\mathcal{U}} I_k\right) = \sum_{k\in\mathcal{U}} \mathrm{E}_p(I_k) = \sum_{k\in\mathcal{U}} \pi_k \tag{B.2}$$

By using the exact expression for the expectation of a ratio of two random variables (see Midzuno, 1950; Koop, 1951, 1972), the expectation of the sample mean (B.1) can be written as follows:

$$\mathrm{E}_p\left(\overline{y}_s\right) = \mathrm{E}_p\left(\frac{\sum_{k\in\mathcal{U}} I_k y_k}{\sum_{k\in\mathcal{U}} I_k}\right) = \frac{\mathrm{E}_p\left(\sum_{k\in\mathcal{U}} I_k y_k\right)}{\mathrm{E}_p\left(\sum_{k\in\mathcal{U}} I_k\right)} + \epsilon \tag{B.3}$$

with

$$\epsilon = \mathrm{E}_p\left[\left(\sum_{k\in\mathcal{U}} I_k y_k\right)\left\{\left(\sum_{k\in\mathcal{U}} I_k\right)^{-1} - \left(\sum_{k\in\mathcal{U}} \pi_k\right)^{-1}\right\}\right] \tag{B.4}$$

The ϵ-term is usually negligible, but this is not the case for very small populations. Therefore, from Eq. (B.3), we can use the following approximation:

$$\mathrm{E}_p\left(\overline{y}_s\right) \approx \frac{\mathrm{E}_p\left(\sum_{k\in\mathcal{U}} I_k y_k\right)}{\mathrm{E}_p\left(\sum_{k\in\mathcal{U}} I_k\right)} = \frac{\sum_{k\in\mathcal{U}} \pi_k y_k}{\sum_{k\in\mathcal{U}} \pi_k} = \frac{1}{N\overline{\pi}_\mathcal{U}}\sum_{k\in\mathcal{U}} \pi_k y_k \tag{B.5}$$

with an average inclusion probability of $\overline{\pi}_\mathcal{U} = N^{-1}\sum_{k\in\mathcal{U}} \pi_k$. Therefore, the bias is approximated as follows:

$$\mathrm{B}_p\left(\overline{y}_s\right) = \mathrm{E}_p\left(\overline{y}_s\right) - \overline{y}_\mathcal{U} \approx \frac{1}{N\overline{\pi}_\mathcal{U}}\sum_{k\in\mathcal{U}} \pi_k y_k - \overline{y}_\mathcal{U} = \frac{1}{\overline{\pi}_\mathcal{U}}\left[\frac{1}{N}\sum_{k\in\mathcal{U}} \pi_k y_k - \overline{\pi}_\mathcal{U}\,\overline{y}_\mathcal{U}\right] = \frac{1}{\overline{\pi}_\mathcal{U}}\,\mathrm{Cov}_{\pi y} \tag{B.6}$$

with $\mathrm{Cov}_{\pi y} = (N-1)S_{\pi y}/N$ where $S_{\pi y}$ is the adjusted covariance (Appendix A). Therefore, we can also write:

$$\mathrm{B}_p\left(\overline{y}_s\right) \approx \frac{\mathrm{Cov}_{\pi y}}{\overline{\pi}_\mathcal{U}} = \frac{(N-1)}{N}\frac{R_{\pi y}\,S_\pi S_y}{\overline{\pi}_\mathcal{U}} \tag{B.7}$$

For $(N-1)/N \approx 1$, we obtain the approximate expression:

$$\mathrm{B}_p\left(\overline{y}_s\right) \approx \frac{R_{\pi y}\,S_\pi S_y}{\overline{\pi}_\mathcal{U}} \tag{B.8}$$

Appendix B.2. Fixed-size sampling design

With a fixed-size sampling design, the sample size is the constant $n = \sum_{k\in\mathcal{U}} \pi_k$, and the sample mean (B.1) can therefore be written as follows:

$$\overline{y}_s = \frac{\sum_{k\in\mathcal{U}} I_k y_k}{\sum_{k\in\mathcal{U}} \pi_k} \tag{B.9}$$

The expectation of the sample mean is then:

$$\mathrm{E}_p\left(\overline{y}_s\right) = \frac{\mathrm{E}_p\left(\sum_{k\in\mathcal{U}} I_k y_k\right)}{\sum_{k\in\mathcal{U}} \pi_k} = \frac{\sum_{k\in\mathcal{U}} \pi_k y_k}{\sum_{k\in\mathcal{U}} \pi_k} = \frac{1}{N\overline{\pi}_{\mathcal{U}}} \sum_{k\in\mathcal{U}} \pi_k y_k \tag{B.10}$$

The bias is therefore written exactly:

$$\mathrm{B}_p\left(\overline{y}_s\right) = \frac{\mathrm{Cov}_{\pi y}}{\overline{\pi}_{\mathcal{U}}} = \frac{(N-1)}{N} \frac{R_{\pi y}\, S_\pi S_y}{\overline{\pi}_{\mathcal{U}}} \tag{B.11}$$

For $(N-1)/N \approx 1$, we again obtain the approximate expression (B.8).

References

Albert, C.H., Yoccoz, N.G., Edwards, T.C., Graham, C.H., Zimmermann, N.E., Thuiller, W., 2010. Sampling in ecology and evolution — bridging the gap between theory and practice. Ecography 33, 1028–1037.

Allison, P.D., 2002. Missing data. Sage Publications, Thousand Oaks, California, USA.

Aubry, P., 2023. On the correct implementation of the Hanurav-Vijayan selection procedure for unequal probability sampling without replacement. Commun. Stat. Simul. Comput. 52, 1849–1877.

Aubry, P., Francesiaz, C., 2022. On comparing design-based estimation versus model-based prediction to assess the abundance of biological populations. Ecol. Indic. 144, 109394.

Aubry, P., Guillemain, M., Sorrenti, M., 2020. Increasing the trust in hunting bag statistics: why random selection of hunters is so important. Ecol. Indic. 117, 106522.

Aubry, P., Pontier, D., Aubineau, J., Berger, F., Léonard, Y., Mauvy, B., Marchandeau, S., 2012. Monitoring population size of mammals using a spotlight-count-based abundance index: how to relate the number of counts to the precision? Ecol. Indic. 18, 599–607.

Aubry, P., Quaintenne, G., Dupuy, J., Francesiaz, C., Guillemain, M., Caizergues, A., 2023. On using stratified two-stage sampling for large-scale multispecies surveys. Ecol. Inform. 77, 102229.

Baddeley, A., Rubak, E., Turner, R., 2016. Spatial point patterns. Methodology and applications with R. Chapman & Hall/CRC, Boca Raton, Florida, USA.

Baker, R., Brick, J.M., Bates, N.A., Battaglia, M., Couper, M.P., Dever, J.A., Gile, K.J., Tourangeau, R., 2013. Summary report of the AAPOR task force on non-probability sampling. J. Surv. Stat. Methodol. 1, 90–143.

Barnett, V., 2002. Sample survey. Principles & methods. Third edition. John Wiley & Sons, Chichester, UK.

Bethlehem, J., 1988. Reduction of nonresponse bias through regression estimation. J. Off. Stat. 4, 251–260.

Bethlehem, J., 1999. Cross-sectional research, in: Ader, H.J., Mellenbergh, G.J. (Eds.), Research methodology in the social, behavioural and life sciences. Sage Publications, Thousand Oaks, California, USA, pp. 110–142.

Bethlehem, J., 2002. Weighting nonresponse adjustments based on auxiliary information, in: Groves, R.M., Dillman, D.A., Eltinge, J.L., Little, R.J.A. (Eds.), Survey nonresponse. John Wiley & Sons, New York, New York, USA, pp. 275–287.

Bethlehem, J., 2009. Applied survey methods: a statistical perspective. John Wiley & Sons, Hoboken, New Jersey, USA.

Bethlehem, J., Cobben, F., Schouten, B., 2011. Handbook of nonresponse in household surveys. John Wiley & Sons, Hoboken, New Jersey, USA.

Boakes, E.H., McGowan, P.J.K., Fuller, R.A., Chang-qing, D., Clark, N.E., O'Connor, K., Mace, G.M., 2010. Distorted views of biodiversity: spatial and temporal bias in species occurrence data. Plos Biol. 8, e1000385.

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. Sci. Rep. 12, 11069.

Boyd, R.J., Powney, G.D., Pescott, O.L., 2023. We need to talk about nonprobability samples. Trends Ecol. Evol. 38, 521–531.

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

Brick, J.M., Montaquila, J.M., 2009. Nonresponse and weighting, in: Pfeffermann, D., Rao, C.R. (Eds.), Handbook of Statistics 29A. Sample surveys: design, methods and applications. Elsevier, Oxford, UK, pp. 163–185.

Brus, D.J., de Gruijter, J.J., 1993. Design-based versus model-based estimates of spatial means: theory and application in environmental soil science. Environmetrics 4, 123–152.

Brus, D.J., de Gruijter, J.J., 1997. Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with discussion). Geoderma 80, 1–44.

Cecconi, L., Grisotto, L., Catelan, D., Lagazio, C., Berrocal, V., Biggeri, A., 2016. Preferential sampling and Bayesian geostatistics: statistical modeling and examples. Stat. Methods Med. Res. 25, 1224–1243.

Chadwick, F.J., Haydon, D.T., Husmeier, D., Ovaskainen, O., Matthiopoulos, J., 2023. LIES of omission: complex observation processes in ecology. Trends Ecol. Evol. 39, 368–380.

Charles, E.P., 2005. The correction for attenuation due to measurement error: clarifying concepts and creating confidence sets. Psychol. Methods 10, 206–226.

Chytrý, M., 2001. Phytosociological data give biased estimates of species richness. J. Veg. Sci. 12, 439–444.

Clifford, D., Kuhnert, P., Dobbie, M., Baldock, J., McKenzie, N., Harch, B., Wheeler, I., McBratney, A., 2011. The dramatic effect of preferential sampling of spatial data on variance estimates, in: Proceedings of the 2011 European Regional Conference of The International Environmetrics Society (TIES).

Cochran, W.G., 1977. Sampling techniques. Third edition. John Wiley & Sons, New York, New York, USA.

Conn, P.B., Thorson, J.T., Johnson, D.S., 2017. Confronting preferential sampling when analysing population distributions: diagnosis and model-based triage. Methods Ecol. Evol. 8, 1535–1546.

Courbois, J.Y.P., Urquhart, N.S., 2004. Comparison of survey estimates of the finite population variance. J. Agric. Biol. Environ. Stat. 9, 236–251.

Dahms, C., Killen, S.S., 2023. Temperature change effects on marine fish range shifts: a meta-analysis of ecological and methodological predictors. Glob. Chang. Biol. 29, 4459–4479.

Delany, S., 2010. Guidance on waterbird monitoring methodology: field protocol for waterbird counting. Technical Report. Wetlands International. Wageningen, The Netherlands.

Diekmann, M., Kühne, A., Isermann, M., 2007. Random vs non-random sampling: effects on patterns of species abundance, species richness and vegetation-environment relationships. Folia Geobot. 42, 179–190.

Diggle, P.J., Menezes, R., Su, T., 2010. Geostatistical inference under preferential sampling. Appl. Stat. 32, 191–232.

Diggle, P.J., Ribeiro, P.J., 2007. Model-based geostatistics. Springer, New York, New York, USA.

Dinsdale, D., Salibian-Barrera, M., 2019. Methods for preferential sampling in geostatistics. Appl. Stat. 68, 181–198.

Dumelle, M., Higham, M., Ver Hoef, J.M., Olsen, A.R., Madsen, L., 2022. A comparison of design-based and model-based approaches for finite population spatial sampling and inference. Methods Ecol. Evol. 13, 2018–2029.

Edwards, T.C., Cutler, D.R., Zimmermann, N.E., Geiser, L., Moisen, G.G., 2006. Effects of sample survey design on the accuracy of classification tree models in species distribution models. Ecol. Modell. 199, 132–141.

Eklund, G., 1959. Studies of selection bias in applied statistics. Almqvist & Wiksell, Uppsala, Sweden.

Ekroos, J., Fox, A.D., Christensen, T.K., Petersen, I.K., Kilpi, M., Jónsson, J.E., Green, M., Laursen, K., Cervencl, A., de Boer, P., Nilsson, L., Meissner, W., Garthe, S., Öst, M., 2012. Declines amongst breeding Eider *Somateria mollissima* numbers in the Baltic/Wadden Sea flyway. Ornis Fenn. 89, 81–90.

Elzinga, C.L., Salzer, D.W., Willoughby, J.W., Gibbs, J.P., 2001. Monitoring plant and animal populations. Blackwell Science, Malden, Massachusetts, USA.

Enders, C.K., 2010. Applied missing data analysis. The Guilford Press, New York, New York, USA.

Fernández, D., Nakamura, M., 2015. Estimation of spatial sampling effort based on presence-only data and accessibility. Ecol. Modell. 299, 147–155.

Fournier, A.M.V., White, E.R., Heard, S.B., 2019. Site-selection bias and apparent population declines in long-term studies. Conserv. Biol. 33, 1370–1379.

Gelfand, A.E., Sahu, S.K., Holland, D.M., 2012. On the effect of preferential sampling in spatial prediction. Environmetrics 23, 565–578.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B., 2014. Bayesian data analysis. Third edition. CRC Press, Boca Raton, Florida, USA.

Girvetz, E.H., Greco, S.E., 2007. How to define a patch: a spatial model for hierarchically delineating organism-specific habitat patches. Landsc. Ecol. 22, 1131–1142.

Gitzen, R.A., Millspaugh, J.J., Cooper, A.B., Licht, D.S., 2012. Design and analysis of long-term ecological monitoring studies. Cambridge University Press, New York, New York, USA.

Goovaerts, P., 1997. Geostatistics for natural resources evaluation. Oxford University Press, New York, New York, USA.

Gray, E.J., Evangelou, E., 2023. A design utility approach for preferentially sampled spatial data. Appl. Stat. 72, 1041–1063.

Greenwood, J.J.D., 2003. The monitoring of british breeding birds: a success story for conservation science? Sci. Total Environ. 310, 221–230.

Gregoire, T.G., 1998. Design-based and model-based inference in survey sampling: appreciating the difference. Can. J. For. Res. 28, 1429–1447.

Grisotto, L., Consonni, D., Cecconi, L., Catelan, D., Lagazio, C., Bertazzi, P.A., Baccini, M., Biggeri, A., 2016. Geostatistical integration and uncertainty in pollutant concentration surface under preferential sampling. Geospat. Health 11, 56–61.

Groves, R.M., Fowler, F.J., Couper, M.P., Lepkowski, J.M., Singer, E., Tourangeau, R., 2004a. Survey methodology. John Wiley & Sons, Hoboken, New Jersey, USA.

Groves, R.M., Presser, S., Dipko, S., 2004b. The role of topic interest in survey participation decisions. Public Opin. Q. 68, 2–31.

de Gruijter, J.J., ter Braak, C.J.F., 1990. Model-free estimation from spatial samples: a reappraisal of classical sampling theory. Math. Geol. 22, 407–415.

Hájek, J., 1964. Asymptotic theory of rejective sampling with varying probabilities from a finite population. Ann. Math. Stat. 35, 1491–1523.

Hájek, J., 1981. Sampling from a finite population. Marcel Dekker, New York, New York, USA.

Hall, L.S., Krausman, P.R., Morrison, M.L., 1997. The habitat concept and a plea for standard terminology. Wildl. Soc. Bull. 25, 173–182.

Hankin, D.G., Mohr, M.S., Newman, K.B., 2019. Sampling theory for the ecological and natural resource sciences. Oxford University Press, Oxford, UK.

Haziza, D., Beaumont, J.F., 2017. Construction of weights in surveys: a review. Stat. Sci. 32, 206–226.

Hedayat, A.S., Sinha, B.K., 1991. Design and inference in finite population sampling. John Wiley & Sons, New York, New York, USA.

Hobbs, N.T., Hooten, M.B., 2015. Bayesian models. A statistical primer for ecologists. Princeton University Press, Princeton, New Jersey, USA.

Illian, J., Penttinen, A., Stoyan, H., Stoyan, D., 2008. Statistical analysis and modelling of spatial point patterns. John Wiley & Sons, Chichester, UK.

Irvine, K.M., Rodhouse, T.J., Wright, W.J., Olsen, A.R., 2018. Occupancy modeling species-environment relationships with non-ignorable designs. Ecol. Appl. 28, 1616–1625.

Johnson, N.L., Kemp, A.W., Kotz, S., 2005. Univariate discrete distributions. Third edition. John Wiley & Sons, Hoboken, New Jersey, USA.

Johnson, N.L., Kotz, S., Balakrishnan, N., 1997. Discrete multivariate distributions. John Wiley & Sons, New York, New York, USA.

Johnston, A., Matechou, E., Dennis, E.B., 2023. Outstanding challenges and future directions for biodiversity monitoring using citizen science data. Methods Ecol. Evol. 14, 103–116.

Journel, A.G., 1994. Resampling from stochastic simulations (with discussion). Environ. Ecol. Stat. 1, 63–91.

Kalton, G., Maligalig, D., 1991. A comparison of methods of weighting adjustment for nonresponse, in: Anderson-Brown, M. (Ed.), Proceedings of the U.S. Bureau of the Census 1991 Annual Research Conference, U.S. Department of Commerce, Bureau of the census (Washington, District of Columbia, USA). pp. 409–428.

Kellner, K., Swihart, R., 2014. Accounting for imperfect detection in ecology: a quantitative review. PLoS ONE 9, e111436.

Koop, J.C., 1951. A note on the bias of the ratio estimate. Bull Int. Stat. Inst. 33, 141–146.

Koop, J.C., 1972. On the derivation of expected value and variance of ratios without the use of infinite series expansions. Metrika 19, 156–170.

Kotz, S., Balakrishnan, N., Read, C., Vidakovic, B., Johnson, N., 2006. Bias, in: Encyclopedia of statistical sciences. Second edition. Volume 1. John Wiley & Sons, Hoboken, New Jersey, USA, pp. 483–484.

Kruskal, W.H., Mosteller, F., 2006. Representative sampling, in: Kotz, S., Balakrishnan, N., Read, C.B., Vidakovic, B., Johnson, N.L. (Eds.), Encyclopedia of statistical sciences. Second edition. Volume 11. John Wiley & Sons, Hoboken, New Jersey, USA, pp. 7203–7207.

Lawless, J.F., 1997. Likelihood and pseudo likelihood estimation based on response-biased observation, in: Basawa, I.V., Godambe, V.P., Taylor, R.L. (Eds.), Selected Proceedings of the Symposium on Estimating Functions. Inst. Math. Stat., Hayward, California, USA, pp. 43–55.

Lawless, J.F., Kalbfleisch, J.D., Wild, C.J., 1999. Semiparametric methods for response-selective and missing data problems in regression. J. R. Stat. Soc. Ser. B Methodol. 61, 413–438.

Lehikoinen, A., Jaatinen, K., Vähätalo, A.V., Clausen, P., Crowe, O., Deceuninck, B., Hearn, R., Holt, C.A., Hornman, M., Keller, V., Nilsson, L., Langendoen, T., Tománková, I., Wahl, J., Fox, A.D., 2013. Rapid climate driven shifts in wintering distributions of three common waterbird species. Glob. Chang. Biol. 19, 2071–2081.

Lehikoinen, E., Sparks, T.H., 2010. Changes in migration, in: Moller, A.P., Fiedler, W., Berthold, P. (Eds.), Effects of climate change on birds. Oxford University Press, Oxford, UK, pp. 89–112.

Lenoir, J., Svenning, J.C., 2013. Latitudinal and elevational range shifts under contemporary climate change, in: Levin, S. (Ed.), Encyclopedia of Biodiversity. Second edition. Volume 4. Elsevier, Amsterdam, The Netherlands, pp. 599–611.

Lenoir, J., Svenning, J.C., 2015. Climate-related range shifts — a global multidimensional synthesis and new research directions. Ecography 38, 15–28.

Lepš, J., Šmilauer, P., 2007. Subjectively sampled vegetation data: don't throw out the baby with the bath water. Folia Geobot. 42, 169–178.

Little, R.J.A., Rubin, D.B., 2019. Statistical analysis with missing data. Third edition. John Wiley & Sons, Hoboken, New Jersey, USA.

Maclean, I.M.D., Austin, G.E., Rehfisch, M.M., Blew, J., Crowe, O., Delany, S., Devos, K., Deceuninck, B., Günther, L., Laursen, K., Van Roomen, M., Wahl, J., 2008. Climate change causes rapid changes in the distribution and site abundance of birds in winter. Glob. Chang. Biol. 14, 2489–2500.

McCarty, J.P., 2001. Ecological consequences of recent climate change. Conserv. Biol. 15, 320–331.

McClure, C.J.W., Rolek, B.W., 2023. Pitfalls arising from site selection bias in population monitoring defy simple heuristics. Methods Ecol. Evol. 14, 1489–1499.

Mentges, A., Blowes, S.A., Hodapp, D., Hillebrand, H., Chase, J.M., 2021. Effects of site-selection bias on estimates of biodiversity change. Conserv. Biol. 35, 688–698.

Merckx, B., Steyaert, M., Vanreusel, A., Vincx, M., Vanaverbeke, J., 2011. Null models reveal preferential sampling, spatial autocorrelation and overfitting in habitat suitability modelling. Ecol. Modell. 222, 588–597.

Midzuno, H., 1950. An outline of the theory of sampling systems. Ann. Inst. Stat. Math. 1, 149–156.

Molenberghs, G., Fitzmaurice, G., Kenward, M.G., Tsiatis, A., Verbeke, G., 2015. Handbook of missing data methodology. CRC Press, Boca Raton, Florida, USA.

Nichols, J.D., Williams, B.K., 2006. Monitoring for conservation. Trends Ecol. Evol. 21, 668–673.

Oh, H.L., Scheuren, F.J., 1983. Weighting adjustment for unit nonresponse, in: Madow, W.G., Olkin, I., Rubin, D.B. (Eds.), Incomplete data in sample surveys. Volume 2: Theory and bibliographies. Academic Press, San Diego, California, USA, pp. 143–184.

Olea, R.A., 2021. Revisiting the declustering of spatial data with preferential sampling. Comput. Geosci. 157, 104946.

O'Neill, B., 2014. Some useful moment results in sampling problems. Am. Stat. 68, 282–296.

Orlóci, L., 1975. Multivariate analysis in vegetation research. Springer, Dordrecht, The Netherlands.

Overton, W.S., Stehman, S.V., 1995. Design implications of anticipated data uses for comprehensive environmental monitoring programmes. Environ. Ecol. Stat. 2, 287–303.

Palmer, M.D., 1993. Potential biases in site and species selection for ecological monitoring. Environ. Monit. Assess. 26, 277–282.

Pan, Y., Ren, X., Gao, B., Liu, Y., Gao, Y., Hao, X., Chen, Z., 2015. Global mean estimation using a self-organizing dual-zoning method for preferential sampling. Environ. Monit. Assess. 187, 121.

Parmesan, C., 2006. Ecological and evolutionary responses to recent climate change. Annu. Rev. Ecol. Evol. Syst. 37, 637–669.

Parmesan, C., 2019. Range and abundance changes, in: Lovejoy, T.E., Hannah, L. (Eds.), Biodiversity and climate change. Transforming the biosphere. Yale University Press, New Haven, Connecticut, USA, pp. 25–38.

Pati, D., Reich, B.J., Dunson, D.B., 2011. Bayesian geostatistical modelling with informative sampling locations. Biometrika 98, 35–48.

Pavlacky, D.C., Lukacs, P.M., Blakesley, J.A., Skorkowsky, R.C., Klute, D.S., Hahn, B.A., Dreitz, V.J., George, T.L., Hanni, D.J., 2017. A statistically rigorous sampling design to integrate avian monitoring and management within bird conservation regions. PLoS ONE 12, e0185924.

Pennino, M.G., Paradinas, I., Illian, J.B., Muñoz, F., Bellido, J.M., López-Quílez, A., Conesa, D., 2019. Accounting for preferential sampling in species distribution models. Ecol. Evol. 9, 653–663.

Perret, J., Besnard, A., Charpentier, A., Papuga, G., 2023. Plants stand still but hide: imperfect and heterogeneous detection is the rule when counting plants. J. Ecol. 111, 1483–1496.

Podani, J., 1984. Spatial processes in the analysis of vegetation: theory and review. Acta Bot. Hung. 30, 75–118.

Pokallus, J.W., Pauli, J.N., 2015. Population dynamics of a northern-adapted mammal: disentangling the influence of predation and climate change. Ecol. Appl. 25, 1546–1556.

Press, W.H., Flannery, B.P., Teukolsky, S.A., Vetterling, W.T., 1989. Numerical recipes in Pascal. The art of scientific computing. Cambridge University Press, New York, New York, USA.

Roleček, J., Chytrí, M., Hájek, M., Lvončík, S., Tichý, L., 2007. Sampling design in large-scale vegetation studies: do not sacrifice ecological thinking to statistical purism! Folia Geobot. 42, 199–208.

Särndal, C.E., Lundström, S., 2005. Estimation in surveys with nonresponse. John Wiley & Sons, Chichester, UK.

Särndal, C.E., Swensson, B., Wretman, J.H., 1992. Model assisted survey sampling. Springer, New York, New York, USA.

Schuster, B.M., 2020. The contribution of terminology research to the understanding of science communication, in: Leßmöllmann, A., Dascal, M., Glonings, T. (Eds.), Science communication. De Gruyter, Berlin, Germany, pp. 167–186.

Scott, A.J., Wild, C.J., 2011. Fitting binary regression models with response-biased samples. Can. J. Stat. 39, 519–536.

da Silva Ferreira, G., 2020. Geostatistics under preferential sampling in the presence of local repulsion effects. Environ. Ecol. Stat. 27, 549–570.

Smith, A.N.H., Anderson, M.J., Pawley, M.D.M., 2017. Could ecologists be more random? straightforward alternatives to haphazard spatial sampling. Ecography 40, 1251–1255.

ter Steege, H., Haripersaud, P.P., Banki, O.S., Schieving, F., 2011. A model of botanical collectors' behavior in the field: never the same species twice. Am. J. Bot. 98, 31–37.

Stuart, A., 1984. The ideas of sampling. Third edition. Charles Griffin, London, UK.

Swacha, G., Botta-Dukát, Z., Kacki, Z., Pruchniewicz, D., Zolnierz, L., 2017. A performance comparison of sampling methods in the assessment of species composition patterns and environment — vegetation relationships in species-rich grasslands. Acta Soc. Bot. Pol. 86, 3561.

Tillé, Y., 2006. Sampling algorithms. Springer, New York, New York, USA.

Tillé, Y., 2020. Sampling and estimation from finite populations. John Wiley & Sons, Hoboken, New Jersey, USA.

Tripathi, R.C., Gupta, R.C., Gurland, J., 1994. Estimation of parameters in the beta binomial model. Ann. Inst. Stat. Math. 46, 317–331.

van Turnhout, C.A.M., Willems, F., Plate, C., van Strien, A., Teunissen, W., van Dijk, A., Foppen, R., 2008. Monitoring common and scarce breeding birds in the netherlands: applying a post-hoc stratification and weighting procedure to obtain less biased population trends. Rev. Catalana Ornitol. 24, 15–29.

Vallecillo, D., Gauthier-Clerc, M., Guillemain, M., Vittecoq, M., Vandewalle, P., Roche, B., Champagnon, J., 2021. Reliability of animal counts and implications for the interpreation of trends. Ecol. Evol. 11, 2249–2260.

Vedensky, D., Parker, P.A., Holan, S.H., 2023. A look into the problem of preferential sampling through the lens of survey statistics. Am. Stat. 77, 313–322.

Vos, P., Meelis, E., ter Keurs, W.J., 2000. A framework for the design of ecological monitoring programs as a tool for environmental and nature management. Environ. Monit. Assess. 61, 317–344.

Wang, Y.H., 1993. On the number of successes in independent trials. Stat. Sin. 3, 295–312.

Watson, J., Zidek, J.V., Shaddick, G., 2019. A general theory for preferential sampling in environmental networks. Ann. Appl. Stat. 13, 2662–2700.

White, G.C., 2005. Correcting wildlife counts using detection probabilities. Wildl. Res. 32, 211–216.

Wiens, J.A., 1976. Population responses to patchy environments. Annu. Rev. Ecol. Syst 7, 81–120.

van Wilgenburg, S.L., Mahon, C.L., Campbell, G., McLeod, L., Campbell, M., Evans, D., Easton, W., Francis, C., Haché, S., Machtans, C.S., Mader, C., Pankratz, R.F., Russell, R., Smith, A.C., Thomas, P., Toms, J.D., Tremblay, J.A., 2020. A cost efficient spatially balanced hierarchical sampling design for monitoring boreal birds incorporating access costs and habitat stratification. PLoS ONE 15, e0234494.

Yoccoz, N.G., Nichols, J.D., Boulinier, T., 2001. Monitoring of biological diversity in space and time. Trends Ecol. Evol. 16, 446–453.

Yoccoz, N.G., Nichols, J.D., Boulinier, T., 2003. Monitoring of biological diversity — a response to Danielsen et al. Oryx 37, 410.

Young, L., Young, J., 1998. Statistical Ecology. A population perspective. Springer, New York, New York, USA.

Zarkovich, S.S., 1966. Quality of statistical data. FAO, Rome, Italy.

Zidek, J.V., Shaddick, G., Taylor, C.G., 2014. Reducing estimation bias in adaptively changing monitoring networks with preferential site selection. Ann. Appl. Stat. 8, 1640–1670.