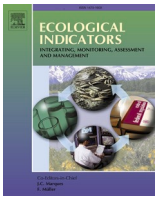
Ecological Indicators 144 (2022) 109394

Contents lists available at ScienceDirect

Ecological Indicators

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



Original Articles

On comparing design-based estimation versus model-based prediction to assess the abundance of biological populations

Philippe Aubry [a,*], Charlotte Francesiaz [b]

[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

ARTICLE INFO

Keywords:
Abundance estimation
Superpopulation
Delta-lognormal distribution
Monte Carlo simulation
Vanellus vanellus
Pluvialis apricaria

ABSTRACT

Assessing the abundance of biological populations is a central technical challenge in ecology, whether fundamental or applied. Although the question faced is simple, it is actually a complicated topic that has produced a vast array of methods. The first step is always sampling the geographical space, ideally by means of a randomized selection process. In that case, with a frequentist interpretation of probability, abundance can be estimated by taking into account randomization, or predicted conditionally on the sample at hand, by specifying a statistical model. This leads to a choice between so-called design-based estimation and model-based prediction. The goal of this methodological article is to contribute to the understanding of fundamental notions regarding these two statistical frameworks by the targeted audience, mainly quantitative ecologists. For this purpose, we illustrate the comparison between design-based estimation in the case of simple random sampling without replacement (SRSWOR) and model-based prediction. As an example, we model count data with a delta-lognormal distribution and rely on uniformly minimum-variance unbiased estimators (UMVUEs) for the prediction of abundance. We investigate the robustness of the predictor by contaminating the delta-lognormal distribution using actual count data. Data from a survey concerning wintering populations in France of two wader species, namely, northern lapwing (*Vanellus vanellus*) and European golden plover (*Pluvialis apricaria*) serve as illustrative examples. By means of Monte Carlo simulations, we highlight the lack of robustness of the predictor based on the delta-lognormal distributional model, in terms of both actual bias and precision. We organize the discussion around the illustrative examples in the context of the sampling design, the model and the data.

1. Introduction

The question of determining the abundance of individuals of a species at a given spatial scale remains central to many ecological studies and monitoring programs. Ecology is sometimes even defined as the study of the causes of variations in abundance in space and time (Krebs, 2014). In particular, this is one of the definitions of ecology that is embodied by the ecological indicator field (Niemi and McDonald, 2004). From population dynamics to management decisions (exotic species regulation, rewilding, etc.), biological population size is the most basic — yet not necessarily the easiest — information to know. For example, the International Union for Conservation of Nature (IUCN) status of species is defined by the species' abundance trends. Many management decisions, such as hunting bag limits or conservation actions, depend on IUCN status. Beyond species conservation, the rationale for biological monitoring rests on the fact that living organisms integrate the impact of many variables and that their abundance and other population parameters can provide an indication of the overall health of the ecosystems of which they are a part (Spellerberg, 2005). Abundance — whether absolute or relative — is a key variable for finding indicator species (e.g., Dufrêne and Legendre, 1997) or designing indicators of environmental disturbance (e.g., Meire and Dereu, 1990; Trenkel and Rochet, 2003; Hiddink, 2005).

While abundance assessment is a central question, it remains difficult to address in practice because there is no reliable and inexpensive way to determine the abundance of any species at any spatial scale. This is a particularly difficult challenge for large spatial scales (and even more so for species with high dispersal capacities). As it is impossible to count all

* Corresponding author.
E-mail addresses: philippe.aubry@ofb.gouv.fr (P. Aubry), charlotte.francesiaz@ofb.gouv.fr (C. Francesiaz).

https://doi.org/10.1016/j.ecolind.2022.109394
Received 21 June 2022; Accepted 28 August 2022
Available online 18 September 2022

individuals, one must estimate (or predict) total abundance, with all the statistical uncertainty this implies. This results in an extensive literature devoted to the single issue of determining abundance over a bounded territory during a limited period of time. This is particularly the case in animal ecology, where abundance assessment is a central technical topic, leading to books (e.g., Norton-Griffiths, 1978; Seber, 1982; Buckland et al., 2001; Buckland et al., 2004; Rivoirard et al., 2000; Borchers et al., 2002) and reviews (e.g., Seber, 1986; Seber, 1992; Schwarz and Seber, 1999; Iijima, 2020) documenting perpetually evolving methods.

In this methodological article, we address some fundamental concerns regarding the type of statistical framework that can be used by quantitative ecologists aiming to assess biological population size. Our article is not intended to be a review of abundance estimation methods or to offer novel insights for improving population size estimates or to discuss abundance-based ecological indicators. The purpose of this article is rather to help document the comparison between two approaches (estimation and prediction) in the case of a finite population of spatial sampling units in the field of abundance assessment, to allow discussion of their strengths, limits and challenges in this context.

Let us denote $\mathscr{D}$, a spatial domain — considered as two-dimensional, not necessarily connex — and $\mathscr{T}$, a time period concerned by the study at hand. Determining the abundance of a given species over $\mathscr{D} \times \mathscr{T}$ necessarily requires some form of spatial and temporal sampling because it is impossible to count organisms at all points of space, at any time. We put aside here the issue of temporal sampling by considering that counting is performed during a period of demographic stability (i.e., neither recruitment of adults nor significant mortality has occurred). We also consider that during this period, immigration and emigration movements in relation to the domain's borders can be neglected. Thus, at the time of the survey, we consider that $\mathscr{D}$ contains an approximately closed biological population, both demographically and geographically.

1.1. Finite population of spatial sampling units

Spatial sampling can be performed by considering geographic space either as continuous or discretized. In this article, we adopt the second perspective by considering in particular that $\mathscr{D}$ is partitioned into spatial sampling units, i.e., we define a finite population $\mathscr{U}$ of nonoverlapping subdomains u_i whose set completely covers $\mathscr{D}$ (Fig. 1).

We consider that the spatial sampling units u_i are unambiguously identifiable by integer labels $(1, 2, \ldots, i, \ldots, N)$. The list of units (labels) constitutes a sampling frame (Fig. 2). As it will not be necessary here to refer to the physical nature of the units, to simplify the notation, we directly use the indices $(1, 2, \ldots, i, \ldots, N)$ to designate the units, in lieu of

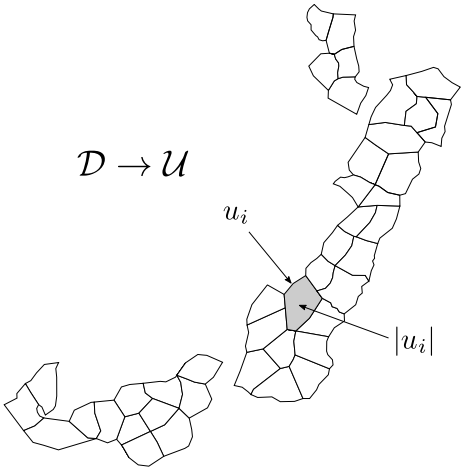


Fig. 1. Example of a (non connex) spatial domain $\mathscr{D}$ partitioned into a finite set $\mathscr{U}$ of subdomains u_i with area $|u_i|$ such that $\sum|u_i| = |\mathscr{D}|$.

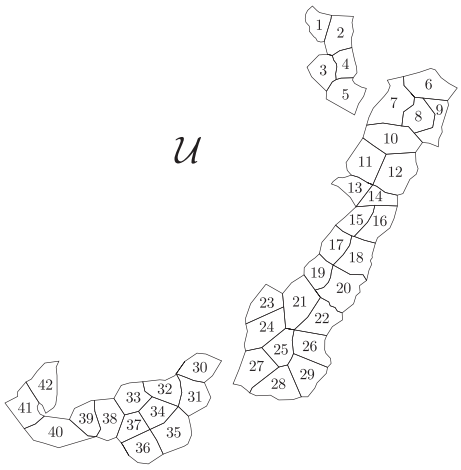


Fig. 2. Example of a finite population $\mathscr{U}$ with $N = 42$ spatial sampling units partitioning a domain. The units are individually identified by *labels* whose list is known (*sampling frame*).

$(u_1, u_2, \ldots, u_i \ldots, u_N)$.

1.2. Assessing the abundance of biological populations

Abundance is determined via two steps: (i) sampling the population of spatial sampling units $\mathscr{U}$, which leads to a sample s of size n; and (ii) counting the individuals in the units of s. The second step can also be viewed formally as a sampling step due to *imperfect detection* (e.g., White, 2005; Kellner and Swihart, 2014).

In step (i), one has the opportunity to control the sampling process applied to $\mathscr{U}$, in particular using probability sampling. There are many possibilities in this area, including stratified sampling, multistage sampling, multiphase sampling and spatially balanced sampling; we refer the reader, for example, to Hankin et al. (2019) and Tillé (2020). These multiple possibilities can be combined in more or less complex ways.

For step (ii), the sampling process that leads to the number of individuals counted (the variable of interest, which we denote as y) is not under control. At best, inclusion probabilities of individuals in the set of the counted individuals can only be estimated or modeled, often in a global manner (i.e., on average for all individuals).

In this article, we are interested in only the first step, which concerns a finite population of spatial units, considering that probability sampling can be implemented.

After counting the individuals in each unit of the spatial sample, we obtain a dataset $\{y_i, i \in s\}$ from which we try to assess a quantity θ defined on the statistical population $\mathscr{U}$; we speak here of the mean or total number of individuals in $\mathscr{D}$. The statistical assessment of θ requires that a probabilistic link be made between the units of s and those belonging to the unknown part of the sampled population $r = \mathscr{U} - s$ (r stands for *remaining*). This link can be made in several ways, which potentially offers several theoretical frameworks. In this article, we endeavor to contribute to the comparison between two frequentist paradigms known as *design-based* and *model-based* approaches. For the notations used in this article, the reader is referred to Appendix A.

2. Design-based versus model-based dichotomy

The contrast between design-based and model-based paradigms appeared explicitly in the survey sampling literature at the end of the 1970s with the article by Särndal (1978), then in the environmental sciences during the 1990s with, in particular De Gruijter and Ter Braak (1990) in the Earth sciences, Brus and De Gruijter (1993) in soil sciences, Gregoire (1998) in forestry, and more modestly in applied ecology with

Edwards (1998) citing Olsen et al. (1999). As the distinction between design-based and model-based approaches is not necessarily well known to the target audience, in this section, we start by explaining the fundamentals about these two paradigms.

2.1. Design-based estimation

Consider samples of fixed size n drawn without replacement from a finite population of size N; the number of *distinct* samples (see Hedayat and Sinha, 1991, p. 2) that can be formed is the number of combinations $\binom{N}{n}$. Even if this number becomes extremely large when N increases, the set of all distinct samples $\mathscr{S}$ constitutes a finite set of which we can — at least in principle — number all the elements in lexicographic order. For example, with the very small population $\mathscr{U} = \{1, 2, 3\}$ and the sample size $n = 2$, we have $\mathscr{S} = \{s_1 = \{1,2\}, s_2 = \{1,3\}, s_3 = \{2,3\}\}$. Let us denote $r(s)$ the lexicographic rank of a sample $s \in \mathscr{S}$, for example, in this case, $r(\{1,3\}) = r(s_2) = 2$. By assigning a selection probability $p(s)$ to each sample $s \in \mathscr{S}$, we obtain a probability mass function (discrete probability distribution). For example, in this case, $p(s_1) = 1/4$, $p(s_2) = 1/4$ and $p(s_3) = 1/2$. The specification of $p(s) \geqslant 0$ for all $s \in \mathscr{S}$ defines a (probability) sampling design, with:

$$\sum_{s \in \mathscr{S}} p\left(s\right) = 1 \tag{1}$$

Designs are implemented by sampling algorithms, whose theory is now very advanced (see Tillé, 2006).

In design-based estimation, the expectation and variance of an estimator $\widehat{\theta}$ are defined with respect to the discrete probability distribution $p(s)$, which gives (for simplicity of notation, in what follows, the variable of interest y is implied):

$$\mathrm{E}_p\left(\widehat{\theta}\right) = \sum_{s \in \mathscr{S}} p\left(s\right) \widehat{\theta}\left(s\right) \tag{2}$$

$$\mathrm{V}_p(\widehat{\theta}) = \sum_{s \in \mathscr{S}} p(s) \left[\widehat{\theta}(s) - \mathrm{E}_p(\widehat{\theta})\right]^2 \tag{3}$$

The use of p as a subscript for mathematical operators (expectation, variance, etc.) refers to the sampling design and is intended to avoid confusion about the underlying source of stochasticity. Similarly, we will use the terms p-expectation, p-variance, etc.

The probabilities that single units and pairs of units are part of a sample (*first- and second-order inclusion probabilities*), respectively, are defined as:

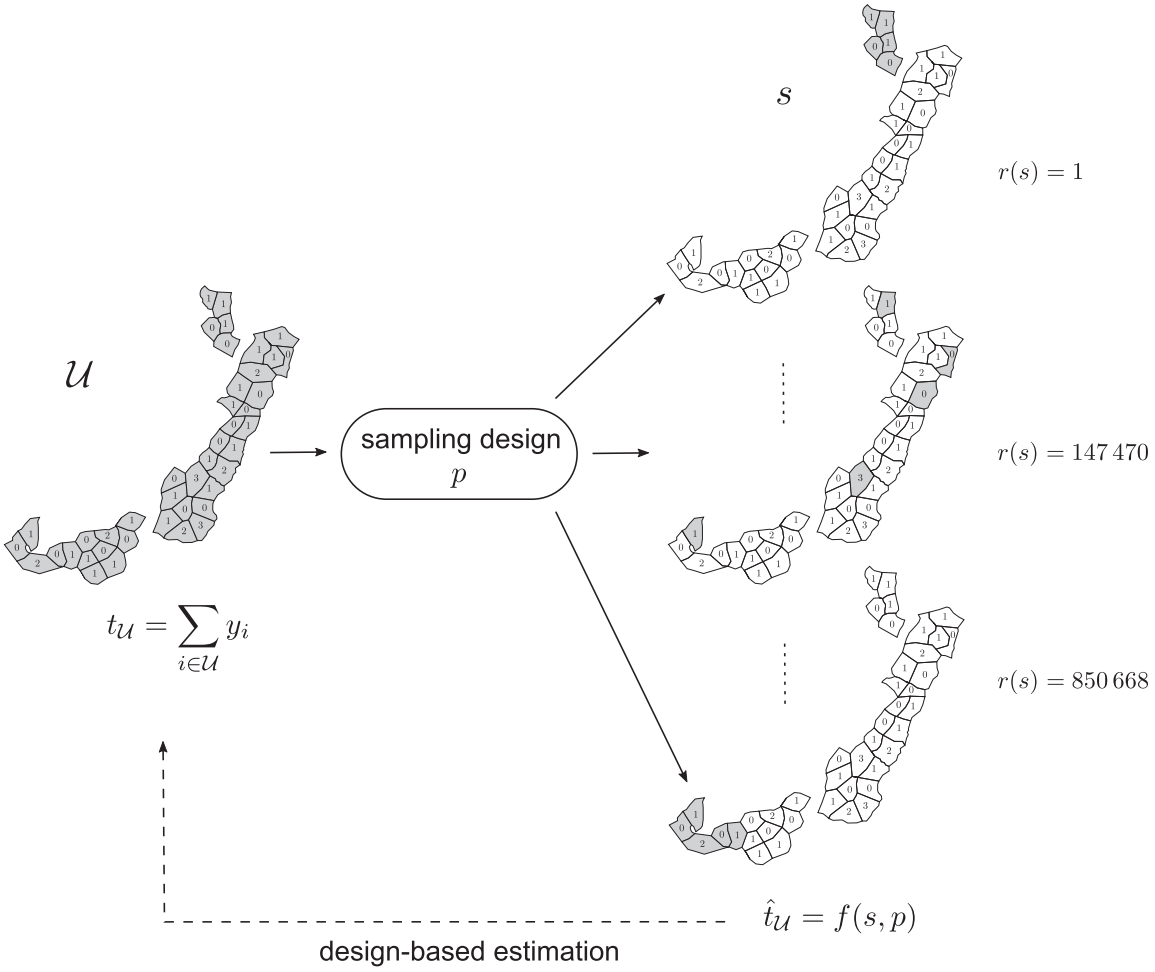


Fig. 3. Schematic representation of design-based estimation. The total $t_{\mathscr{U}}$ is a fixed value that we want to estimate. Sampling the population $\mathscr{U}$ by the means of a sampling design leads to a sample s (for example, the sample of lexicographic rank $r(s) = 147\,470$) among all those that can be formed (here $\binom{N}{n} = \binom{42}{5} = 850\,668$ samples) with a known probability $p(s)$. The estimator $\widehat{t}_{\mathscr{U}}$ is a function of sample s and sampling design p.

$$\pi_i = \sum_{s \ni i} p\left(s\right) \tag{4}$$

$$\pi_{ij} = \sum_{s \ni (i,j)} p\left(s\right) \tag{5}$$

Note that in expressions (4) and (5), the symbol "$\ni$" stands for "contains" and the sums are therefore on all samples s that contain unit i or a pair of units (i,j), respectively.

Knowledge of inclusion probabilities is sufficient to allow statistical estimation of the finite population parameters. In particular, when $\pi_i > 0$ and $\pi_{ij} > 0$ for all $i,j \in \mathscr{U}$, unbiased estimators for the mean (or total) and its sampling variance may be defined without requiring any assumptions about the spatial autocorrelation or shape of the statistical distribution for the variable of interest, which may be arbitrarily complex.

To summarize (Fig. 3), (i) the quantity of interest is a fixed value (the population is fixed); (ii) a probability sampling design assigns a selection probability $p(s)$ to each of the samples that can be formed ($s \in \mathscr{S}$); and (iii) the estimator of the parameter of interest depends on the inclusion probabilities of the units, which in turn depend on the selection probabilities of the samples. The design-based approach relies on the frequentist interpretation of probability.

2.2. Model-based prediction

If we remain in a frequentist framework, then the model-based approach considers that the y_i ($i \in \mathscr{U}$) are random variables of joint distribution ξ. Thus, the finite population under study is viewed itself as a random sample drawn from a *superpopulation* (infinite population) of model ξ. It follows that a function θ of y_i ($i \in \mathscr{U}$) is a random variable. Under a model ξ, the expectation and variance of θ are written as integrals defined over an infinity of realizations, which can be simply denoted as:

$$\mathrm{E}_\xi\left(\theta\right) = \int \theta d\xi \tag{6}$$

$$\mathrm{V}_\xi(\theta) = \int \left[\theta - \mathrm{E}_\xi(\theta)\right]^2 d\xi \tag{7}$$

The use of ξ as a subscript for mathematical operators refers to the model and is intended to avoid confusion about the underlying source of stochasticity. Similarly, we will use the terms ξ-expectation, ξ-variance, ξ-covariance, etc.

In the strict sense, under the model, the value of θ for the finite

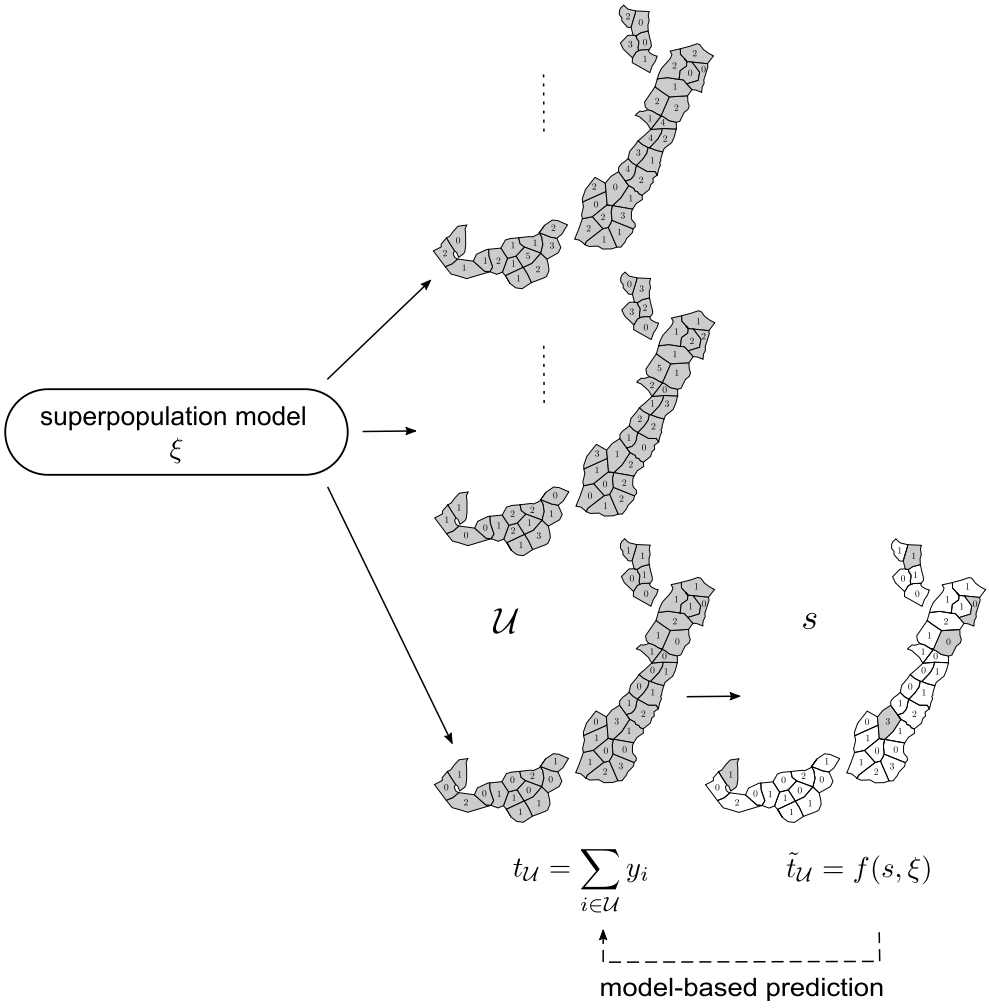


Fig. 4. Schematic representation of model-based prediction. The population $\mathscr{U}$ under study is viewed as a realization of an infinite set of populations (a superpopulation) following a model ξ. The total $t_\mathscr{U}$ is a random variable whose value is to be predicted. A sample s is drawn from the population $\mathscr{U}$ according to an ignorable selection process. The predictor $\tilde{t}_\mathscr{U}$ is a function of sample s and model ξ.

population under study is not estimated (it is not a fixed value) but predicted (it is a realization of a random variable), using a predictor $\widetilde{\theta}$.

In the superpopulation model approach, the population is random but the sample s is fixed; in particular, it is not assumed to have been selected by a probability sampling design. Nevertheless, the selection process of the sample is assumed to be ignorable or made ignorable (see, for example, Sugden and Smith, 1984 or Pfeffermann, 1993). The statistical properties of a predictor $\widetilde{\theta}$ thus depend on only the model ξ.

To summarize (Fig. 4), (i) the quantity of interest is viewed as a realization of a random variable (the population is random); (ii) the sample s is fixed; and (iii) the predictor of the quantity of interest depends on knowledge of the superpopulation model ξ from which the population is viewed as a random sample.

2.3. Interpreting the superpopulation concept

The model-based prediction introduces the concept of superpopulation through the sequence of the three nested sets:

$$\underset{(\infty)}{\text{superpopulation}}\ \xi\ \rightarrow\ \underset{(N)}{\text{population}}\ \mathscr{U}\ \rightarrow\ \underset{(n)}{\text{sample}}\ s \tag{8}$$

The term *superpopulation* seems to have been coined by Fisher (1932) and later introduced by Deming and Stephan (1941) in the field of survey sampling, but the idea of a finite population that is itself regarded as a random sample from an infinite population seems to have originated in the beginning of the 19th century with Laplace's work (see Pearson, 1928; Cochran, 1978).

As a matter of course, the term *superpopulation* is more widely encountered in statistics (especially in survey sampling theory) than in ecology (but see Eberhardt and Thomas, 1991 or Aubry and Debouzie, 2000; Aubry and Debouzie, 2001 for such a use). A different meaning of the term *superpopulation* is also encountered in statistical ecology in the field of capture-recapture methods (e.g., Nichols et al., 2005; Royle et al., 2014; McCrea and Morgan, 2014), where *"a 'superpopulation' is defined as the total number of animals that were alive and available to be captured during at least one sampling period of the study and is thus composed of the recruits to the population across all sampling periods"* (Wen et al., 2011). However, it is somewhat unfortunate that the term *superpopulation* has been used to define a distinct concept in statistical ecology in light of its already long history in statistics; nevertheless, the two meanings are easily distinguishable according to the context. In this article, we clearly refer to the first acception of the term *superpopulation*.

In general, statisticians do not dwell on the question of the interpretation of the concept of superpopulation, notable exceptions being Cassel et al. (1977, pp. 81–82) and Nathan (2011) who enumerate several alternative interpretations. With a frequentist definition of probability in mind, let us mention two main interpretations: (i) the superpopulation is modeled to describe a situation in which the finite population may be considered to have been generated by a real-world stochastic process (the population is then regarded as having been drawn by 'Nature'); and (ii) the superpopulation model is regarded as a purely mathematical device, useful in deriving results.

In the first interpretation, the estimation of the superpopulation parameters may be the goal of the study such that the underlying process leading to the observed values can be characterized. In this case, the survey data are used for analytic purposes. Conversely, for descriptive purposes, the survey data are simply used to assess certain characteristics defined on a given fixed population. In this case, the estimation of superpopulation parameters is merely a technical intermediate step toward, for instance, predicting total abundance. Briefly stated, the analytic questions are "why?" and "how?" whereas the descriptive question is merely "how many?" (Deming, 1953).

For model-based prediction, both interpretations (i) and (ii) of the concept of superpopulation can be used. If one refers in particular to the first interpretation (i.e., a hypothetical data-generating process), model-based prediction may be (a) concerned with the here and now of what has occurred, focusing on the status of the actual population rather than on a superpopulation corresponding to the biological and ecological processes involved in its production and (b) intended to make generalizations or predictions beyond the sampling units that comprise the finite population that has actually been sampled. To summarize, when referring to the superpopulation concept, ecologists must be clear regarding whether the population of interest is the (real) finite population (i.e., a particular realization of the invoked superpopulation model), or whether it is the (hypothetical) superpopulation itself.

In this article, we focus on the spatial finite population at hand, and the superpopulation is merely viewed as a mathematical device useful in making abundance predictions.

2.4. Superpopulation model typology

It is useful to develop a typology of superpopulation models that is intelligible to an audience of ecologists, as those proposed by statisticians may be too complex (e.g., Cassel et al., 1977, p. 90, Table 4.1). For this purpose, we propose a very simple typology that depends on the nature of the information that can be taken into account to build a superpopulation model. For a univariate situation (i.e., a single variable of interest, here the abundance for one species), in a frequentist framework, we distinguish essentially three categories of information: (i) a possible spatiotemporal autocorrelation ($\rho(d,t)$) — or only spatial $\rho(d)$ or temporal $\rho(t)$; (ii) possible auxiliary variables $\mathbf{X}$ in relation — linear or nonlinear — to the variable of interest ($f(\mathbf{X})+\boldsymbol{\varepsilon}$); and (iii) the assumption of a statistical distribution ($f(y)$). By crossing the presence-absence of these three types of information, we define eight major types of superpopulation models that appear immediately intelligible from an operational perspective (Table 1).

2.5. Comparing the design-based vs. model-based approaches

Comparing the two approaches for abundance assessment clearly makes sense only when the data are gathered from a sample selected by a probability sampling design. For a gentle introduction to this issue in the field of ecology, we consider *simple random sampling without replacement* (SRSWOR). This without-replacement design is of fixed size (the sample size n is predetermined). Moreover, SRSWOR is ignorable because of equal inclusion probabilities (self-weighted sampling design). It can therefore lead to samples that can be used directly and simply in a model-based approach.

In what follows, we consider a practical situation in which neither auxiliary variables nor an exploitable spatial autocorrelation structure are available; we therefore consider only superpopulation models of Types I and IV (Table 1).

Table 1
Typology of superpopulation models according to the nature of the information taken into account: spatiotemporal autocorrelation ($\rho(d,t)$); auxiliary variables ($f(\mathbf{X})+\boldsymbol{\varepsilon}$); statistical distribution ($f(y)$). In this article, we consider only Types I and IV models (grayed rows).

Type	$\rho(d,t)$	$f(\mathbf{X})+\boldsymbol{\varepsilon}$	$f(y)$
I	no	no	no
II	✓	no	no
III	no	✓	no
IV	no	no	✓
V	✓	✓	no
VI	✓	no	✓
VII	no	✓	✓
VIII	✓	✓	✓

3. Count data distribution models

The general characteristic of count data ($y \geqslant 0$) is that they may exhibit a high proportion of zero values. To take this into account in a sufficiently flexible manner, one approach is to use a two-component mixture model. There are two possibilities: (i) increasing the probability of zero counts from a distribution defined for $y \geqslant 0$ (*zero-inflated distributions*); and (ii) introducing a dichotomy between $y = 0$ and $y > 0$ in a mixture model with two separately estimable parts (*hurdle-at-zero, conditional, two-part, and delta distributions* designate the same thing). If the second possibility is adopted, then a suitable superpopulation model for count data ($y \geqslant 0$) is written as:

$$G(y; p_0, \boldsymbol{\theta}) = \begin{cases} p_0 & y = 0 \\ p_0 + (1 - p_0) F(y; \boldsymbol{\theta}) & y > 0 \end{cases} \tag{9}$$

where p_0 is the probability of obtaining a zero count and $F(y; \boldsymbol{\theta})$ is a cumulative distribution with parameters $\boldsymbol{\theta}$, corresponding to a positive distribution, either discrete (e.g., zero-truncated Poisson distribution) or continuous (e.g., lognormal distribution). In what follows, as an example, we consider the lognormal distribution.

3.1. Lognormal distribution

The lognormal distribution is often used to model abundance data (e.g., Clark and Bjørnstad, 2004; Dennis et al., 2006). Let z be a random variable distributed according to the standard normal distribution (i.e., $z \sim \mathrm{Norm}(0, 1)$) of probability density:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2} z^2 \right] \qquad z \in \mathbb{R} \tag{10}$$

Then, $y = \exp(\mu + \sigma z)$ follows a lognormal distribution that is completely specified by the mean (μ) and variance (σ^2) of log-transformed abundances. Parameters μ and σ^2 are estimated without bias from the subset of positive count data (denoted as s_+), respectively, by $\bar{y}_{\ln}$ and $s^2_{\ln}$ (Aitchison and Brown, 1969, p. 39):

$$\bar{y}_{\ln} = \frac{1}{n_+} \sum_{i \in s_+} \ln y_i \tag{11}$$

$$s^2_{\ln} = \frac{1}{n_+ - 1} \sum_{i \in s_+} (\ln y_i - \bar{y}_{\ln})^2 \tag{12}$$

with n_+ the size of s_+. The probability density function of the lognormal distribution is written as (Aitchison and Brown, 1969, p. 8, Eq. 2.5; Shimizu et al., 1988, p. 2, Eq. 2.1):

$$\begin{aligned} f(y; \mu, \sigma^2) &= \frac{1}{y\sigma} \phi((\ln y - \mu)/\sigma) \\ &= \frac{1}{y\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2} \left(\frac{\ln y - \mu}{\sigma} \right)^2 \right] \qquad y \in \mathbb{R}_+^* \end{aligned} \tag{13}$$

The expectation α and variance β^2 are given by (Aitchison and Brown, 1969, p. 8, Eqs. 2.7 and 2.8):

$$\alpha = \exp\left(\mu + \frac{1}{2} \sigma^2 \right) \tag{14}$$

$$\beta^2 = \alpha^2 \left(\exp(\sigma^2) - 1 \right) \tag{15}$$

3.2. Delta-lognormal distribution

The lognormal distribution is no longer appropriate when zero counts must be accounted for. Consequently, in this article, we use as a superpopulation model the delta-lognormal distribution (Shimizu et al., 1988), often also called the Δ-*distribution* (Aitchison and Brown, 1969).

The delta-lognormal distribution results from a mixture of a Dirac mass at 0 with probability p_0 and a lognormal distribution with probability $(1 - p_0)$, that is (e.g., Dennis et al., 1988, p. 325, Eq. 4.2):

$$g(y; p_0, \mu, \sigma^2) = p_0 \delta(y) + (1 - p_0) f(y; \mu, \sigma^2) \tag{16}$$

where $\delta(y)$ is a Dirac distribution that concentrates a unit mass at 0. By applying the calculation rules for the expectation and variance of a mixture distribution (see Appendix A.2), expectation κ_1 and variance κ_2 (the first two cumulants) of the distribution (16) are written as:

$$\kappa_1 = (1 - p_0) \alpha \tag{17}$$

$$\kappa_2 = (1 - p_0)(p_0 \alpha^2 + \beta^2) \tag{18}$$

4. Estimation and prediction

Before introducing the context of prediction that involves both superpopulation sampling and finite population sampling, we first recall the estimation of the finite population mean and of the superpopulation expectation. We also must consider the variance estimation (but we exclude the issue of the sampling variance of the variance estimators).

In the context of a finite population, we can indifferently consider the mean $\bar{y}_{\mathscr{U}}$ or the total $t_{\mathscr{U}} = N\bar{y}_{\mathscr{U}}$ as parameters of interest. For a superpopulation (infinite population), the parameter of interest is the expectation $\mathrm{E}_\xi(y) = \kappa_1$.

4.1. Finite population parameter estimation

Let s be a sample drawn from $\mathscr{U}$ by an SRSWOR of size n out of N. The design-unbiased (or p-unbiased) estimators of $\bar{y}_{\mathscr{U}}$ and $S^2_{\mathscr{U}}$ are, respectively, $\bar{y}_s$ and S^2_s (e.g., Cochran, 1977, pp. 21, 26). The p-variance of the mean estimator is written as (e.g., Cochran, 1977, p. 23, Eq. 2.8):

$$\mathrm{V}_p\left(\bar{y}_s \right) = \left(1 - \frac{n}{N} \right) \frac{S^2_{\mathscr{U}}}{n} \tag{19}$$

The total $t_{\mathscr{U}}$ is thus estimated by $\widehat{t}_{\mathscr{U}} = N\bar{y}_s$ (the so-called *expansion estimator*), and its sampling variance is naturally written as $\mathrm{V}_p(\widehat{t}_{\mathscr{U}}) = N^2 \mathrm{V}_p(\bar{y}_s)$. These results apply regardless of the statistical distribution of y in $\mathscr{U}$.

4.2. Infinite population parameter estimation

4.2.1. Unknown shape of the distribution

Let s be a sample of size n drawn by random sampling from an infinite population of unknown shape. The model-unbiased (or ξ-unbiased) nonparametric estimators of κ_1 and κ_2 are $\bar{y}_s$ and S^2_s, respectively. The ξ-variance of the sample mean is written as (e.g., Kendall, 1945, p. 206, Eq. 9.7):

$$\mathrm{V}_\xi\left(\bar{y}_s \right) = \frac{\kappa_2}{n} \tag{20}$$

4.2.2. Known shape of the distribution

As mentioned above, as an example, in this article we consider the delta-lognormal distribution. The expectation κ_1 and variance κ_2 of this distribution can be estimated by means of *uniformly minimum-variance unbiased estimators* (UMVUEs), that is, unbiased estimators that has lower variance than any other unbiased estimators for all possible values of the parameters of interest. The UMVUEs of κ_1 and κ_2 (denoted as $\widehat{\kappa}_1$ and $\widehat{\kappa}_2$, respectively) were given by Aitchison and Brown (1969, p. 97, Eqs. 9.54 and 9.55) and the exact variance $\mathrm{V}_\xi(\widehat{\kappa}_1)$ was provided by Shimizu and Iwase (1981, Remark 3.1), Shimizu et al. (1988, p. 50) and Smith (1988, Eq. 6). For his part, Pennington (1983, Eq. 4) has provided the UMVUE for the variance $\mathrm{V}_\xi(\widehat{\kappa}_1)$. We refer the reader interested in the formal and computational details to Aubry (2022).

4.3. Prediction of the finite population total

This section is essentially devoted to demystifying the formulas regarding predictors and variances of prediction error, as we surmise that the intended audience may be not aware of prediction theory. Therefore, the different steps of the reasoning process are provided rather than directly providing the final formulas without an understandable rationale.

By denoting $r = \mathscr{U} - s$, the total abundance can be written as:

$$t_{\mathscr{U}} = \sum_{i \in \mathscr{U}} y_i = \underbrace{\sum_{i \in s} y_i}_{t_s} + \underbrace{\sum_{i \in r} y_i}_{t_r} \tag{21}$$

that is, the sum of the totals defined over the sample (t_s) and the remaining part in the population (t_r). In the framework of a superpopulation model, $t_{\mathscr{U}}, t_s$ and t_r are random variables. The prediction issue for $t_{\mathscr{U}}$ can give rise to predictors of two forms (Särndal and Wright, 1984; Firth and Bennett, 1998):

(i) the *projective form*, which is the sum of the predicted values for the whole population, that is:

$$\widetilde{t}_{\mathscr{U}}^{\text{pro}} = \sum_{i \in \mathscr{U}} \widetilde{y}_i \tag{22}$$

(ii) the *predictive form*, which predicts t_r only, that is:

$$\widetilde{t}_{\mathscr{U}}^{\text{pre}} = \underbrace{\sum_{i \in s} y_i}_{t_s} + \underbrace{\sum_{i \in r} \widetilde{y}_i}_{\widetilde{t}_r} \tag{23}$$

In situations where these two predictors differ, $\widetilde{t}_{\mathscr{U}}^{\text{pre}}$ is preferred to $\widetilde{t}_{\mathscr{U}}^{\text{pro}}$ (Firth and Bennett, 1998). Indeed, when n tends toward N, with the predictive form, the predictor tends toward $t_{\mathscr{U}}$, and for $n = N$ (a census case), we obtain $\widetilde{t}_{\mathscr{U}}^{\text{pre}} = t_{\mathscr{U}}$, a property that makes sense (*finite population consistency*, see Cochran, 1977, Section 2.4, p. 21 or Hankin et al., 2019, p. 325). In the case of the projective form, even if we have a census, we still have variability in the estimators of the superpopulation parameters and this translates into uncertainty in the predicted total.

Whatever the form of a predictor $\widetilde{t}_{\mathscr{U}}$, the variance of the prediction error $(\widetilde{t}_{\mathscr{U}} - t_{\mathscr{U}})$ is written as the variance of the difference of two random variables, that is (e.g., Kendall, 1945, p. 226, Eq. 9.60):

$$\mathrm{V}_\xi\left(\widetilde{t}_{\mathscr{U}} - t_{\mathscr{U}}\right) = \mathrm{V}_\xi\left(\widetilde{t}_{\mathscr{U}}\right) + \mathrm{V}_\xi\left(t_{\mathscr{U}}\right) - 2\mathrm{Cov}_\xi\left(\widetilde{t}_{\mathscr{U}}, t_{\mathscr{U}}\right) \tag{24}$$

The covariance $\mathrm{Cov}_\xi(\widetilde{t}_{\mathscr{U}}, t_{\mathscr{U}})$ is positive and tends to 0 when N increases to ∞.

In the absence of auxiliary variables that would be related to the variable of interest, the optimal predictor — in the sense of mean squared error minimization — can be written as (Chambers and Clark, 2012, p. 21):

$$\widetilde{t}_{\mathscr{U}}^{*} = \sum_{i \in s} y_i + \underbrace{\mathrm{E}_\xi(t_r | \{y_i, i \in s\})}_{\text{conditional expectation}} \tag{25}$$

that is, the predictor of the total t_r is its expectation given the data at hand $\{y_i, i \in s\}$.

In the absence of obvious spatial structure that could be exploited (i.e., a trend and autocorrelation), we consider that the random variables y_i ($i = 1, \ldots, N$) have the same expectations and variances and are not correlated, that is:

$$\mathrm{E}_\xi\left(y_i\right) = \kappa_1 \tag{26a}$$

$$\mathrm{V}_\xi\left(y_i\right) = \kappa_2 \tag{26b}$$

$$\mathrm{Cov}_\xi\left(y_i, y_j\right) = 0 \qquad (i \neq j) \tag{26c}$$

There is no universally adopted terminology for designing model (26) (see Cassel et al., 1977, Table 4.1, p. 90; Bolfarine and Zacks, 1992, p. 9; Chambers and Clark, 2012, p. 20), but it may be referred to as a *mean model* (Gregoire, 1998; Hankin et al., 2019, p. 119).

For model (26), the optimal predictor (25) is written as (Chambers and Clark, 2012, p. 21):

$$\widetilde{t}_{\mathscr{U}}^{*} = \sum_{i \in s} y_i + \left(N - n\right) \mathrm{E}_\xi\left(y_i\right) \tag{27}$$

The empirical predictor (predictive form) is then written as:

$$\widetilde{t}_{\mathscr{U}} = \sum_{i \in s} y_i + \underbrace{(N - n)\widehat{\mathrm{E}}_\xi(y_i)}_{\widetilde{t}_r} \tag{28}$$

by substituting an estimator for the corresponding parameter (expectation). Predictor (28) is ξ-unbiased; that is, if the model is correct and the expectation estimator is unbiased, then $\mathrm{E}_\xi(\widetilde{t}_{\mathscr{U}} - t_{\mathscr{U}}) = 0$. The variance of the prediction error $(\widetilde{t}_{\mathscr{U}} - t_{\mathscr{U}})$ is obtained as:

$$\mathrm{V}_\xi\left(\widetilde{t}_{\mathscr{U}} - t_{\mathscr{U}}\right) = \mathrm{V}_\xi\left(\widetilde{t}_r - t_r\right) = \mathrm{V}_\xi\left(\widetilde{t}_r\right) + \mathrm{V}_\xi(t_r) - \underbrace{2\mathrm{Cov}_\xi\left(\widetilde{t}_r, t_r\right)}_{0} \tag{29}$$

The ξ-covariance between $\widetilde{t}_r$ and t_r is zero since: (i) $\widetilde{t}_r$ is a function of the set of values in s ($\{y_i, i \in s\}$), not of the set of values in r ($\{y_i, i \in r\}$) which we do not know; and (ii) the two sets of values $\{y_i, i \in s\}$ and $\{y_i, i \in r\}$ are uncorrelated under the model (see 26c).

4.3.1. Distribution of unknown shape

In our typology (Table 1), when the shape of the distribution $f(y)$ is not known (or known but not taken into account), model (26) corresponds to a Type I model. Then, predictor (28) can be written as:

$$\widetilde{t}_{\mathscr{U}} = \sum_{i \in s} y_i + \left(N - n\right)\overline{y}_s = n\overline{y}_s + \left(N - n\right)\overline{y}_s = N\overline{y}_s \tag{30}$$

which corresponds to the expansion estimator $\widehat{t}_{\mathscr{U}}$ in the SRSWOR case (Section 4.1), although obtained in a different framework. We can therefore designate predictor (30) as an *expansion predictor* (Bolfarine and Zacks, 1992). Note that the projective form is also $N\overline{y}_s$ (*predictive-projective equivalence*; see Firth and Bennett, 1998). From relation (29), the prediction error variance of $(\widetilde{t}_{\mathscr{U}} - t_{\mathscr{U}})$ is obtained as:

$$\mathrm{V}_\xi\left(\widetilde{t}_{\mathscr{U}} - t_{\mathscr{U}}\right) = (N - n)^2\left[\mathrm{V}_\xi\left(\overline{y}_s\right) + \mathrm{V}_\xi\left(\overline{y}_r\right)\right] \tag{31}$$

$$= (N - n)^2\left(\frac{1}{n} + \frac{1}{N - n}\right)\kappa_2 \tag{32}$$

$$= N^2\left(1 - \frac{n}{N}\right)\frac{\kappa_2}{n} \tag{33}$$

a well-known result from Cochran (1939, Eq. 1). For prediction error variance (33), a ξ-unbiased estimator is obtained by substituting S_s^2 for κ_2, which leads to the same variance estimator as for the expansion estimator. The predictor of the mean is $\overline{y}_s$, and its prediction error variance is $\mathrm{V}_\xi(\overline{y}_s - \overline{y}_{\mathscr{U}}) = (1 - n/N)\kappa_2/n$.

4.3.2. Distribution of known shape

In our typology (Table 1), when the shape of the distribution $f(y)$ is (assumed to be) known, model (26) corresponds to a Type IV model. The predictive form is written as:

$$\widetilde{t}_{\#}^{\text{pre}} = \sum_{i \in s} y_i + \left(N - n\right)\widehat{\kappa}_1 \tag{34}$$

which differs from the projective form $N\widehat{\kappa}_1$, which we will not consider further in this article for predicting $t_{\#}$. According to relation (29), the prediction error variance $(\widetilde{t}_{\#}^{\text{pre}} - t_{\#})$ is obtained as:

$$\mathrm{V}_{\xi}\left(\widetilde{t}_{\#}^{\text{pre}} - t_{\#}\right) = (N-n)^2\left[\mathrm{V}_{\xi}\left(\widehat{\kappa}_1\right) + \mathrm{V}_{\xi}\left(\overline{y}_r\right)\right] \tag{35}$$

$$= (N-n)\left[(N-n)\mathrm{V}_{\xi}\left(\widehat{\kappa}_1\right) + \kappa_2\right] \tag{36}$$

The mean predictor is $\widetilde{\overline{y}}_{\#}^{\text{pre}} = N^{-1}\widetilde{t}_{\#}^{\text{pre}}$ (e.g., Smith, 1990, Eq. 3), and its prediction error variance is $\mathrm{V}_{\xi}(\widetilde{\overline{y}}_{\#}^{\text{pre}} - \overline{y}_{\#}) = N^{-2}\mathrm{V}_{\xi}(\widetilde{t}_{\#}^{\text{pre}} - t_{\#})$. For prediction error variance (36), a ξ-unbiased estimator is obtained by substituting estimators $\widehat{\mathrm{V}}_{\xi}(\widehat{\kappa}_1)$ and $\widehat{\kappa}_2$ for $\mathrm{V}_{\xi}(\widehat{\kappa}_1)$ and κ_2, respectively.

5. Case studies

Theory matters but what interests the practitioner most is knowing whether one should prefer an abundance estimate based on a design or on a model. To illustrate this type of questioning in a concrete manner, we rely on count data collected on a random sample of spatial sampling units that concern wintering populations in France of two wader species: (a) northern lapwing (*Vanellus vanellus*); (b) European golden plover (*Pluvialis apricaria*). With reference to the delta-lognormal distribution for prediction using $\widetilde{t}_{\#}^{\text{pre}}$ (34), we seek to answer the following operational question: in terms of bias and efficiency, how does the model-based predictor behave with these data?

Northern lapwing and European golden plover were counted in six regions in northwestern France during the winter of 2004–2005. At this time of the year, individuals of these species exhibit gregarious behavior and are distributed in open habitats, facilitating their counting. Both species frequent more or less the same types of open environments (agricultural plains and meadows), enabling a common monitoring. We present both species in this study as they were counted on the same spatial sample but correspond to different status of abundance. Indeed, northern lapwings are more abundant than European golden plovers (number of zero counts differs greatly). Notably, France is one of the few countries where these two species are hunted. The IUCN status categories are “near threatened” (”vulnerable” on the European scale) and “least concern” for northern lapwing and Eurasian golden plover, respectively.

5.1. Datasets

$N = 9\,312$ communes (the *commune* is the smallest administrative unit in France) located in the Bretagne, Pays de la Loire, Basse-Normandie, Haute-Normandie, Centre and Poitou–Charentes regions (the *region* is the greater administrative subdivision in France) formed the spatial population of sampling units. Among the 786 communes selected by SRSWOR, $n = 784$ communes were surveyed; the nonresponse concerning approximately 0.25% of the sample is considered ignorable. Approximately 280 observers were mobilized to conduct the counts by car, mainly between January 5 and January 12, 2005, with instructions to visit each sampled commune to count the birds from the car, with binoculars if necessary. In the case of rural communes in hedged farmland where most of the surface is likely to host both species, it was requested to cover almost all the commune roads.

At the end of the survey, we obtained two datasets (one for each species) with a high proportion of zero counts (the numbers of zero counts are $n_0 = 490$ and $n_0 = 681$ for northern lapwing and European golden plover, respectively) (see Appendix B). Thus, we estimate the probability of zero counts as $\widehat{p}_0 = 0.625$ for northern lapwing and $\widehat{p}_0 \simeq 0.869$ for European golden plover. The skewness of the distribution of count data is very high, with $\widehat{\gamma}_1 \simeq 9.9$ and $\widehat{\gamma}_1 \simeq 17.6$ for northern lapwing and European golden plover, respectively (see also Fig. 5). The counts are not correlated with the surface areas of the communes. Moreover, at the spatial scale of communes, the counts do not show any spatial autocorrelation structure (unpublished results).

We computed probability-probability plots (PP-plots) against the normal distribution for n_+ log-transformed positive count data; after sorting the data in ascending order, we plotted points with abscissa $\Phi((y_i - \overline{y}_{\ln})/s_{\ln})$ (*fitted cumulative probabilities*) and ordinate $p_i = (i - 0.375)/(0.25 + n_+)$ (*empirical cumulative probabilities*), for $i = 1, \ldots, n_+$, with $\Phi(\cdot)$ as the distribution function of the standard normal distribution, approximated numerically as in Abramowitz and Stegun (1972, p. 932, Eq. 26.2.17). The plotting positions p_i that we used were proposed by Blom (1958, p. 145, Eq. 1) for the special case of a normal distribution. The obtained PP-plots show that the decision to model the positive counts by means of a lognormal distribution is not a priori unreasonable (Fig. 6).

5.2. Estimation and prediction

We seek to compare design-based estimation in the case of SRSWOR and model-based prediction under model (26), either in the absence of assumptions about the shape of the distribution (Type I model) or by assuming that the empirical distribution is adequately described by a delta-lognormal distribution (Type IV model). The estimation under SRSWOR is formally identical to the prediction under the Type I model (Section 4.3.1). In addition, we again find this formal link in defining the efficiency index devoted to comparing the p-variance of the expansion estimator on the average under the ξ-model and the ξ-variance of the prediction error for $\widetilde{t}_{\#}^{\text{pre}}$:

$$\frac{\mathrm{V}_{\xi}\left(\widetilde{t}_{\#}^{\text{pre}} - t_{\#}\right)}{\mathrm{E}_{\xi}\left[\mathrm{V}_p\left(\widehat{t}_{\#}\right)\right]} = \frac{\mathrm{V}_{\xi}\left(\widetilde{t}_{\#}^{\text{pre}} - t_{\#}\right)}{\mathrm{V}_{\xi}\left(\widetilde{t}_{\#} - t_{\#}\right)} = \text{eff} \tag{37}$$

since $\mathrm{E}_{\xi}\left[\mathrm{V}_p\left(\widehat{t}_{\#}\right)\right]$ is identical to (33):

$$\mathrm{E}_{\xi}\left[\mathrm{V}_p\left(\widehat{t}_{\#}\right)\right] = \mathrm{E}_{\xi}\left[N^2\left(1 - \frac{n}{N}\right)\frac{S^2_{\text{残}}}{n}\right] = N^2\left(1 - \frac{n}{N}\right)\frac{1}{n}\mathrm{E}_{\xi}\left(S^2_{\#}\right) \tag{38}$$

$$= N^2\left(1 - \frac{n}{N}\right)\frac{\kappa_2}{n} \tag{39}$$

Note that the efficiency is defined here in the same way as that used by Aitchison and Brown (1969, p. 99, Eq. 9.62) (see also Aubry, 2022).

The estimates obtained from the data are reported in Tables (2.a) and (2.b) for northern lapwing and European golden plover, respectively. In both cases, the value obtained for the predictor (prediction based on the Type IV model) is higher than that obtained with the expansion estimator/predictor (SRSWOR-based estimation or prediction based on the Type I model). The same holds for the estimates of the prediction error variance vs. the estimation variance, and for the respective coefficients of variation. For these datasets, it is clear that we will not use the results obtained in the case of the Type IV model assuming that the counts follow a delta-lognormal distribution. However, to document the impact of the model deviation on the bias and efficiency of the predictor, we must proceed slightly further, relying on a Monte Carlo study.

5.3. Monte Carlo study

Recall that in the framework of probability sampling, we denote π_i as the inclusion probability of a spatial unit $i \in s$, associated with the number of individuals $y_i \geqslant 0$. By definition, the sampling weight is $w_i = \pi_i^{-1}$. Considering only the subset s_+ of positive counts (for $i \in s_+$ we have $y_i > 0$), we define the total weight $W = \sum_{i \in s_+} w_i$ and $P_i = w_i / W$ normalized weights in the sense that $\sum_{i \in s_+} P_i = 1$.

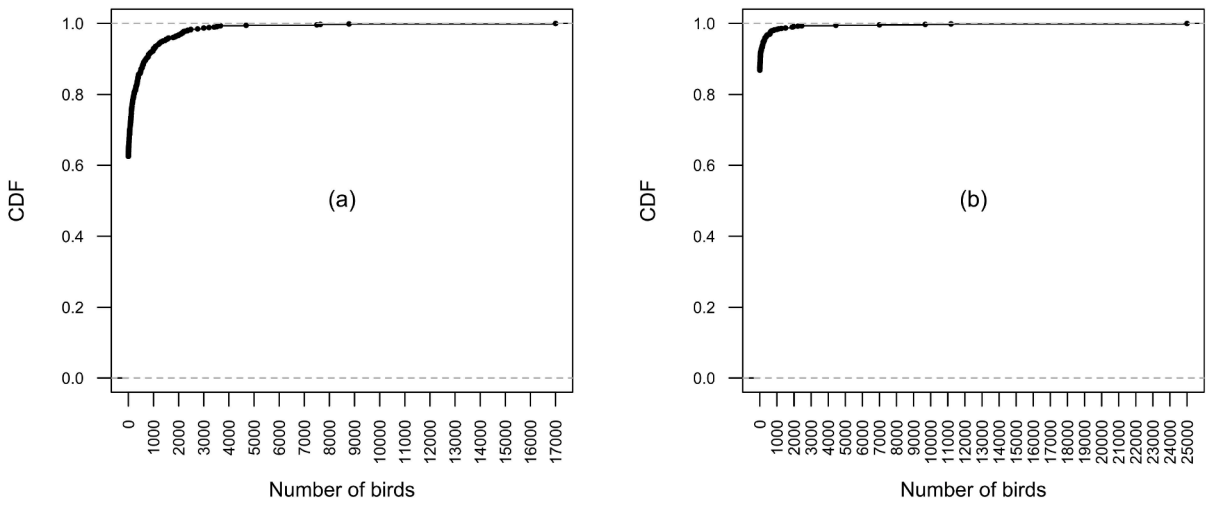


Fig. 5. Empirical cumulative distribution functions (CDFs) for the two datasets. (a) northern lapwing. (b) European golden plover.

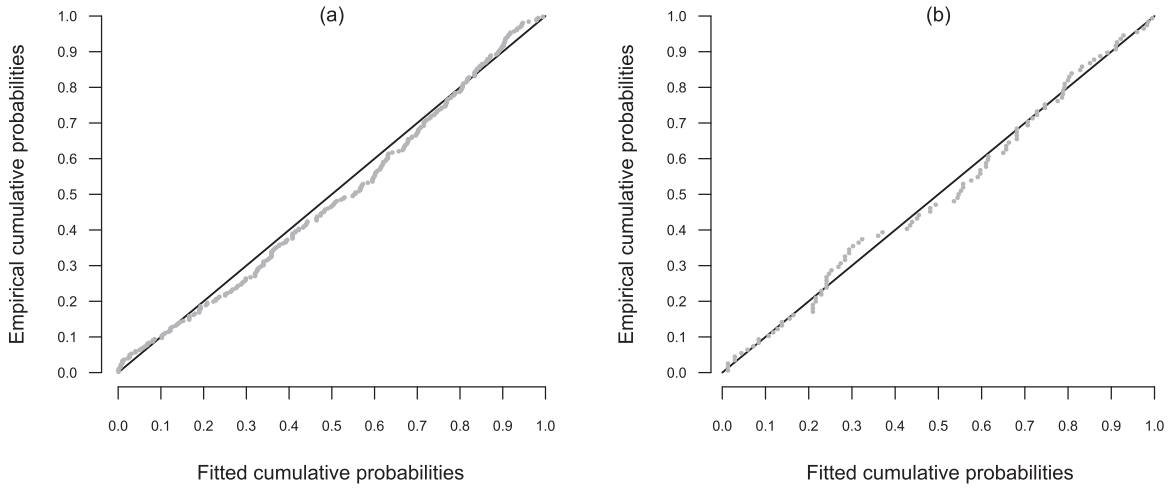


Fig. 6. Probability-probability plots with respect to the normal distribution of n_+ log-transformed positive count data. (a) Northern lapwing ($n_+ = 294, \widehat{\mu} \simeq 5.41243$, $\widehat{\sigma} \simeq 1.72896$). (b) European golden plover ($n_+ = 103, \widehat{\mu} \simeq 4.70569, \widehat{\sigma} \simeq 2.11696$). Details in the text.

Table 2
Estimated mean abundance by commune, based on the expansion estimator ($N^{-1}\widehat{t}_{/\!/}$) and on the predictor assuming the delta-lognormal distribution ($N^{-1}\widetilde{t}^{\text{pre}}_{/\!/}$).

(a) Northern lapwing

	Estimate	Variance	CV (%)
Estimator	264.51	1036.18	12.17
Predictor	361.20	3638.07	16.70

(b) European golden plover

	Estimate	Variance	CV (%)
Estimator	113.31	1399.00	33.01
Predictor	126.54	2271.91	37.67

5.3.1. *Obtaining a superpopulation by resampling*

To define a ξ-superpopulation corresponding to the observed positive values without referring to a parametric model, it is sufficient to consider sampling with replacement of y_i with probabilities P_i ($i \in s_+$). The corresponding ξ-superpopulation can then be written as:

$$h\left(y; \left\{\left(P_i, y_i\right), i \in s_+\right\}\right) = \sum_{i \in s_+} P_i \delta\left(y - y_i\right) \tag{40}$$

where $\delta(y - y_i)$ is a Dirac distribution that concentrates a unit mass at y_i. This superpopulation is nothing more than a mixture of n_+ Dirac distributions that associate a weight P_i to each count $y_i > 0$.

The sampling design considered in the definition of h (40) is unequal probability sampling with replacement. For the implementation of this design, the reader is referred to Särndal et al. (1992, p. 91, Section 3.6.3, p. 97) or to Aubry (2021, Remark 6). In our case — where the sample of spatial units was selected by SRSWOR — the design implemented for resampling is just simple random sampling with replacement (SRSWR). The ξ-superpopulation thus defined reproduces, on average, the observed data; hence, in particular we have $\mathrm{E}_\xi(y) = \overline{y}_{s_+}$ (denoted as a) and $\mathrm{V}_\xi(y) = S^2_{s_+}$ (denoted as b^2).

5.3.2. *Contamination of the theoretical distribution*

The Monte Carlo study proposed in this article consists of contaminating the positive values of the delta-lognormal distribution by the superpopulation h, with a given contamination rate. The contaminated

distribution is thus written as a three-part mixing model:

$$g^{'}\left(y;p_0,\in,\mu,\sigma^2,\{(P_i,y_i),i\in s_+\}\right)= \begin{array}{l} p_0\,\delta(y) \\ +(1-p_0)[\in h(y;\{(P_i,y_i),i\in s_+\}) \\ +(1-\in)f(y;\mu,\sigma^2)\,] \end{array} \tag{41}$$

At first glance, this expression may look complicated but is actually simple: p_0 is the probability of obtaining a zero count; $\in$ is the contamination rate of the lognormal distribution f by the superpopulation h defined by resampling the positive data; μ and σ^2 are the parameters specifying the lognormal part of the model; $\{(P_i,y_i),i\in s_+\}$ is the set of pairs with probability P_i for resampling the positive abundance y_i.

By applying the calculation rules for the expectation and variance of a mixture distribution (see Appendix A.2), the first two cumulants of (41) are written as:

$$\kappa_1=(1-p_0)[\in a+(1-\in)\alpha] \tag{42}$$

$$\kappa_2=(1-p_0)\left\{\in\left(a^2+b^2\right)+(1-\in)\left(\alpha^2+\beta^2\right)-(1-p_0)[\in a+(1-\in)\alpha]^2\right\} \tag{43}$$

For $\in=0$, we obtain the delta-lognormal distribution and we find expressions (17) and (18) as particular instances of (42) and (43), respectively. For $\in=1$, the superpopulation h defined by resampling the positive data entirely replaces the lognormal distribution f, and we obtain:

$$\kappa_1=(1-p_0)\,a \tag{44}$$

$$\kappa_2=\left(1-p_0\right)\left(p_0\,a^2+b^2\right) \tag{45}$$

Thus, by varying $\in$, we can appreciate the impact of the gradual replacement of the delta-lognormal distribution f by the superpopulation h on the statistical performance of the predictor $\tilde{t}_{\mathscr{U}}^{\text{pre}}$ (34).

5.3.3. Simulation algorithm

Let K be the number of simulations to be performed (Monte Carlo effort) to approximate the quantities of interest by an average of the form $\bar{v}^{(j)}=K^{-1}\sum_{k=1}^{K}v_k^{(j)}$. The Monte Carlo simulation proceeds as follows:

1. for a given contamination rate $\in$
2. let $k\leftarrow 0$
3. increment $k\leftarrow k+1$
4. generate a fictitious population $\mathscr{U}$ of size N according to the mixing model (41)
5. compute and memorize the population total $v_k^{(0)}\leftarrow t_{\mathscr{U}}$
6. sample $\mathscr{U}$ by SRSWOR to obtain a sample s
7. compute the UMVUE $\widehat{\kappa}_1$ (see Aubry, 2022)
8. compute the expansion estimator $\widehat{t}_{\mathscr{U}}=N\bar{y}_s$
9. compute the predictor $\tilde{t}_{\mathscr{U}}^{\text{pre}}$ according to expression (34)
10. compute the deviations (errors) $e_1\leftarrow\widehat{t}_{\mathscr{U}}-t_{\mathscr{U}}$ and $e_2\leftarrow\tilde{t}_{\mathscr{U}}^{\text{pre}}-t_{\mathscr{U}}$
11. memorize $v_k^{(1)}\leftarrow e_1, v_k^{(2)}\leftarrow e_2, v_k^{(3)}\leftarrow e_1^2$ and $v_k^{(4)}\leftarrow e_2^2$
12. if $k<K$, then go to step 3; otherwise, go to the next step
13. compute $\widehat{\mathrm{E}}_{\xi}(t_{\mathscr{U}})\leftarrow\bar{v}^{(0)}$, $\widehat{\mathrm{E}}_{\xi}(\widehat{t}_{\mathscr{U}}-t_{\mathscr{U}})\leftarrow\bar{v}^{(1)}$ and $\widehat{\mathrm{E}}_{\xi}(\tilde{t}_{\mathscr{U}}^{\text{pre}}-t_{\mathscr{U}})\leftarrow\bar{v}^{(2)}$
14. compute relative biases $B_1\leftarrow\widehat{\mathrm{E}}_{\xi}(\widehat{t}_{\mathscr{U}}-t_{\mathscr{U}})/\widehat{\mathrm{E}}_{\xi}(t_{\mathscr{U}})$ and $B_2\leftarrow\widehat{\mathrm{E}}_{\xi}\left(\tilde{t}_{\mathscr{U}}^{\text{pre}}-t_{\mathscr{U}}\right)\Big/\widehat{\mathrm{E}}_{\xi}(t_{\mathscr{U}})$
15. compute $\widehat{\mathrm{MSE}}_{\xi}(\widehat{t}_{\mathscr{U}}-t_{\mathscr{U}})\leftarrow\bar{v}^{(3)}$ and $\widehat{\mathrm{MSE}}_{\xi}\left(\tilde{t}_{\mathscr{U}}^{pre}-t_{\mathscr{U}}\right)\leftarrow\bar{v}^{(4)}$
16. compute the relative efficiency $E\leftarrow\widehat{\mathrm{MSE}}_{\xi}(\tilde{t}_{\mathscr{U}}^{\text{pre}}-t_{\mathscr{U}})/\widehat{\mathrm{MSE}}_{\xi}(\widehat{t}_{\mathscr{U}}-t_{\mathscr{U}})$

For this article, we performed the simulations for $\in=0.0(0.05)1.0$ with $K=10^6$ simulations for each value of $\in$.

5.3.4. Results

As expected, the relative bias B_1 (defined for the expansion estimator) is essentially zero regardless of the value of $\in$ (results not shown). The relationship between the values of the relative bias B_2 and the $\in$-values is fitted by a nonlinear model of the form:

$$B_2(\in)=\exp\left(A\in+B\in^C\right)-1 \tag{46}$$

The relative bias increases approximately linearly with $\in$ (Fig. 7.a and 8.a), with a much steeper slope for northern lapwing than for European golden plover. We plotted the relative bias thresholds at 5 and 10 percent, considering that below 5 percent, the relative bias can be viewed as small, between 5 and 10 percent it is medium, and above 10 percent it is of concern. For both species, the results show that the relative bias of the predictor assuming the delta-lognormal distribution is much too high for $\in=1$, and it is more than twice as high for northern lapwing than it is for European golden plover.

The relationship between the values of the relative efficiency E and the $\in$-values is fitted by a nonlinear model of the form:

$$E(\in)=\exp\left(A\in+B\in^C\right)-(1-D) \tag{47}$$

where $D=\text{eff}$ is the value of $E(\in)$ for $\in=0$, exactly computed as:

$$D=\frac{n\left[(N-n)\mathrm{V}_{\xi}\left(\widehat{\kappa}_1\right)+\kappa_2\right]}{N\kappa_2} \tag{48}$$

We obtain $D\simeq 0.4675706$ for northern lapwing and $D\simeq 0.2538185$ for European golden plover. We plot the efficiency limit at 1, the value above which the predictor is less efficient than the expansion estimator/predictor. This threshold is reached for $\in\simeq 0.50$ in the case of northern lapwing and for $\in\simeq 0.90$ in the case of European golden plover (Fig. 7.b and 8.b).

5.3.5. A simple predictive check

Among possible predictive checks and before embarking on the simulation described in Section 5.3.3, as illustrated by Gelman et al. (2014, pp. 189–190), one should first examine whether the model is able to correctly predict the sample total. This check is simply a matter of predicting the sample total using the projective form, that is:

$$\tilde{t}_s^{\text{pro}}=\sum_{i\in s}\tilde{y}_i \tag{49}$$

As illustrated by Fig. 9 and Fig. 10, for $\in=1$, by definition of the superpopulation model (41), we have the ability to predict the sample total (that is, $t_s=207\,379$ for northern lapwing and $t_s=88\,834$ for European golden plover), which corresponds to the mean of the predictor distribution.

For $\in$ tending to 0 — that is, when tending to the pure delta-lognormal model — the ability to predict the sample total decreases gradually, with the behavior depending on the species (Fig. 10).

For $\in=0$ (delta-lognormal model) and 10^6 simulated samples, for northern lapwing, when $t_s=207\,379$, the median is 281 456, the 90% prediction interval is $[206\,249, 419\,010]$, and the largest prediction is 5 347 363, indicating strong overestimation. For European golden plover, when $t_s=88\,834$, the median is 87 804, the 90% prediction interval is $[44\,917, 221\,691]$ and the largest prediction is 15 479 805. Comparison of the two species shows that while for northern lapwing, the distribution is globally shifted to the right (Fig. 10.a, for $\in=0.0$), for European golden plover, the median is very close to the sample total, but the spread to the right is excessive, with a very long right tail of the simulated predictor distribution (Fig. 10.b, for $\in=0.0$).

6. Discussion

The issue of abundance estimation in the case of a finite population of spatial sampling units falls under the theory of probability sampling. Despite regular calls to rely on this theory in the field of ecology (Albert

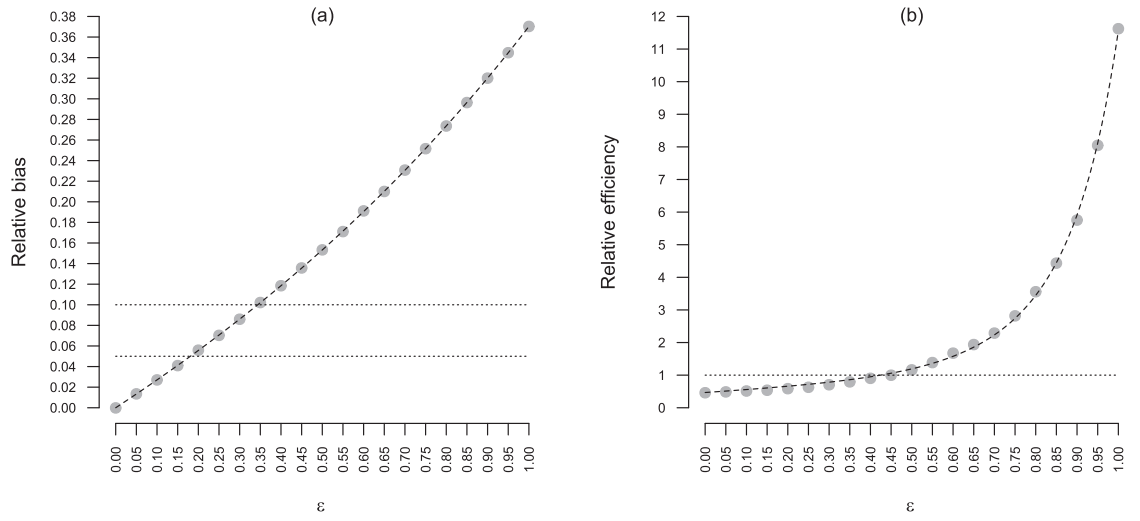


Fig. 7. For northern lapwing, Monte Carlo simulation results for the three-part mixing model. (a) Relative bias, modeled as $\exp(0.26426 \times \epsilon + 0.05072 \times \epsilon^{2.27724}) - 1$. (b) Relative efficiency, modeled as $\exp(0.8654 \times \epsilon + 1.6268 \times \epsilon^{3.8535}) - (1 - 0.4675706)$. Details in the text.

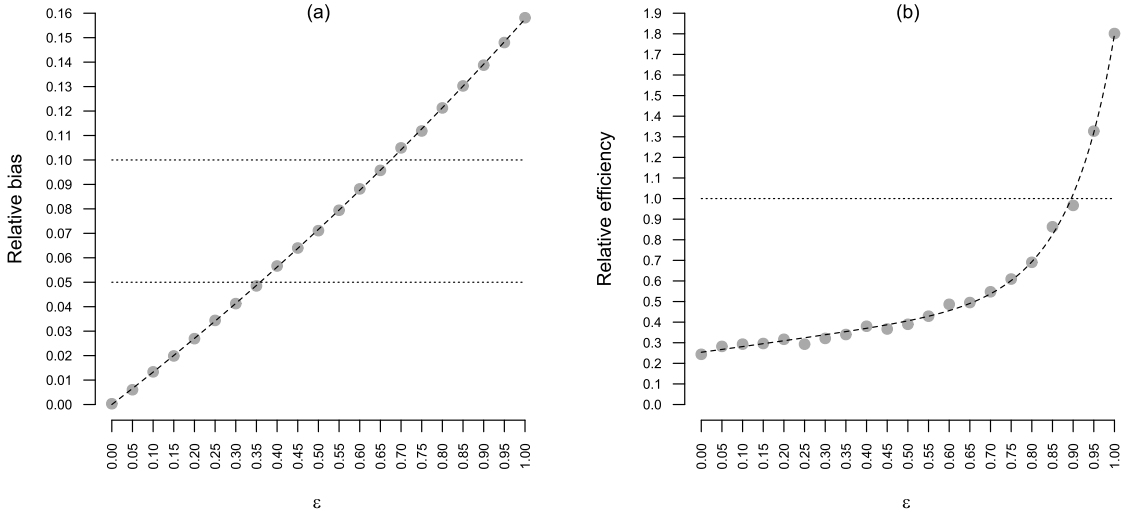


Fig. 8. For European golden plover, Monte Carlo simulation results for the three-part mixing model. (a) Relative bias, modeled as $\exp(0.12986 \times \epsilon + 0.01649 \times \epsilon^{1.96212}) - 1$. (b) Relative efficiency, modeled as $\exp(0.2717 \times \epsilon + 0.6614 \times \epsilon^{6.7483}) - (1 - 0.2538185)$. Details in the text.

et al., 2010; Smith et al., 2017; Aubry et al., 2020), it remains relatively unknown because it is not yet sufficiently taught (Hankin et al., 2019, Preface; Webb, 2021, p. 817) and is rarely present in the literature devoted to animal abundance assessment (Borchers et al., 2002, p. 48). Furthermore, as noted by Hahn (1969); Whitmore (1986) and Lin and Liao (2008) (for instance), the concept of prediction is also not often taught in college courses and rarely present in general books on applied statistics, apart from a presentation in the context of simple linear regression (e.g., Janke and Tinsley, 2005). It follows that the comparison between the estimation and prediction of abundance remains largely to be investigated in our field.

As a first contribution, we were interested here in the potential gain in precision that can be achieved by using a distribution model for count data compared to using the sample mean. As an example, we examined the use of the delta-lognormal distribution.

6.1. *The use of the delta-lognormal distribution*

Pennington (1983) introduced the use of the delta-lognormal distribution in marine biology to estimate mean abundance more efficiently than can be achieved with the sample mean in the case of highly skewed distributions. The article by Pennington (1983) is widely cited in the literature in this field, which may give the impression that this approach is restricted to marine organisms (e.g., Dennis et al., 1988, Section 4.2). The lack of widespread diffusion for a method outside the field in which it was introduced is more a matter of epistemology than related to the nature of the organism populations that one seeks to count. The use of a given distribution can be justified in three main ways: (i) it is easy to manipulate mathematically and/or computationally; (ii) it corresponds to a data generating process that we know is actually at work (for the lognormal distribution, see, for instance, Koch, 1966; Koch, 1969; Parkin et al., 1992, pp. 198–200; Hilborn and Mangel, 1997, pp. 73–76 and Shimizu et al., 1988, Section 3); and (iii) it provides an adequate summary of the data at hand. We have implicitly

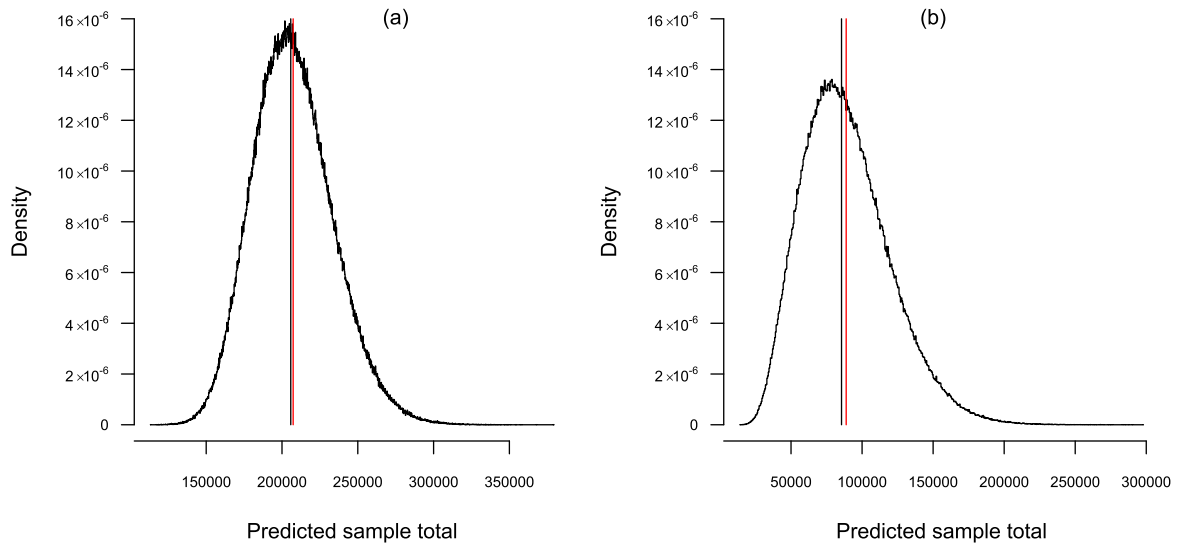


Fig. 9. Empirical distribution of the predicted sample total for $\epsilon = 1$ and 10^6 simulations (black line: median; red line: mean). (a) Northern lapwing. (b) European golden plover.

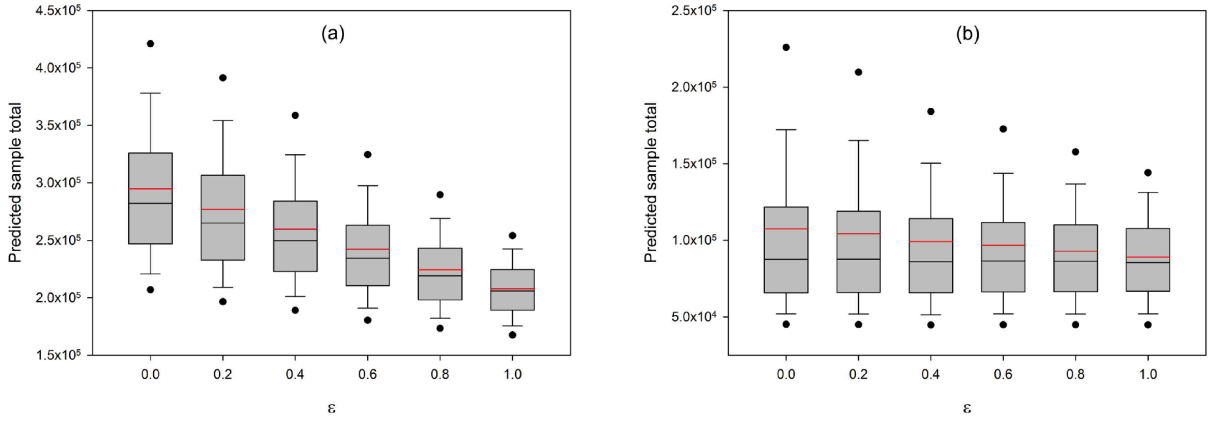


Fig. 10. Box-plot of the predicted sample total as a function of the contamination rate $\epsilon = 0(0.2)1.0$ for 10^4 simulations (dots: 5th/95th percentiles; black line: median; red line: mean). (a) Northern lapwing. (b) European golden plover.

considered the third perspective in this article. In doing so, the use of a continuous distribution in the context of discrete data such as counts can only be justified because it provides an adequate approximation of the empirical distribution of the data, as explained, for example, by Myers and Pepin (1990). This is the case when counts can take many different values within a wide range, and one is not interested in the probability mass of each of them. Under these conditions, it is possible to use this distribution for any type of organisms, as long as it approximates the distribution of the observed data. Thus, in this article, it was a priori reasonable to use the delta-lognormal distribution in the case of counts for the two wader species (Section 5.1).

6.2. *The infinite or finite nature of the sampled population*

The population that can be sampled by controlling for inclusion probabilities is the population of spatial units (the statistical population of interest). Beyond the nature of the distribution used for modeling the counts, there is an issue related to the finite or infinite nature of the statistical population of interest. If we can neglect the fact that the population is finite, then the aims of population mean prediction and superpopulation expectation estimation align. In general, when the spatial sampling fraction is small compared to 1, it is implicitly assumed that the population can be considered infinite (Myers and Pepin, 1990). To the best of our knowledge, the example chosen for this article has rarely been addressed in the context of a spatial population whose finite character could not be ignored (however, see Smith, 1990).

If the data distribution conforms to a delta-lognormal distribution, using UMVUE $\widehat{\kappa}_1$ for estimation (superpopulation case) or for building a predictor (finite population case) results in a substantive precision gain as σ^2 increases. As documented by Aubry (2022), the effect of the finite nature of the sampled population essentially translates into a decrease in the precision gain and a modification of the speed of convergence to its asymptotic value. These theoretical results demonstrate the interest in relying on the UMVU predictor if the data distribution strictly conforms to a delta-lognormal distribution. Of course, in practice, the empirical distribution of the data deviates from a pure delta-lognormal distribution, which entails concerns about the robustness of the prediction based on this superpopulation model.

6.3. *A matter of robustness*

In fields other than ecology, the lack of robustness that could arise from the use of the lognormal distribution when the empirical distribution deviates from the theoretical model have been reported for several decades (e.g., Link and Koch, 1975). For the prediction of abundance, the use of the delta-lognormal distribution thus also

legitimately raises concerns about the robustness both of the unbiasedness and efficiency of the predictor when the empirical distribution deviates from the model.

At least since Tukey and Olkin (1960), this type of question has been studied by referring to contamination of the assumed model by another distribution. This is formalized as a mixture of distributions with a contamination rate ϵ that can be varied. This is how Myers and Pepin (1990) proceed, considering as contaminating distributions the gamma distribution, the Weibull distribution with two parameters and the transformation of the normal distribution given by Johnson and Kotz (1970, p. 268, Eq. 50). The use of arbitrary distributions or under conditions that are not necessarily realistic can, quite legitimately, be criticized (Pennington, 1991; Pennington, 1996; Syrjala, 2000). To avoid this pitfall, in this article, we studied the effect of contamination in the context of the actual datasets considered. We have indeed proposed that the contaminating superpopulation be defined directly from the data, without reference to a parametric model of a distribution (Section 5.3.1). Thus, we can progressively move from the assumed model to a situation corresponding to the data at hand. To our knowledge, such an approach is original. It provides a general way of answering questions asked in the context of actual datasets, which is immediately meaningful to the target audience. By proceeding this way, in the case of our datasets on the numbers of northern lapwings and European golden plovers counted in a portion of French territory during the winter of 2004–2005, we showed that the predictor assuming the delta-lognormal distribution presented significant bias and lower efficiency than the estimator/predictor by expansion and that it should therefore not be used in either of these two cases (Section 5.3.4).

The conclusion reached for the two examples discussed in this article is consistent with those of Myers and Pepin (1990), Myers and Pepin (1991), Syrjala (2000) and Christman (2019, p. 45) and conflicts with Pennington and Stromme (1998), who claim that Pennington's method (Pennington, 1983) is robust to departures from lognormality. However, our conclusion about the case studies does not mean that the use of the delta-lognormal distribution should be rejected in any situation. The problem is that in practice, we need to know whether the use of the delta-lognormal distribution is justified, in the sense that a gain in precision is obtained while maintaining a low bias. As a matter of fact, the use of statistical tests is not very relevant for answering this question.

6.4. Goodness-of-fit is not sufficient

Normality tests (Shapiro–Wilk, Jarque–Bera, Anderson–Darling, Cramér-von Mises, Lilliefors) applied to log-transformed positive counts of northern lapwings ($n_+ = 294$) all lead to rejection of the null hypothesis (resp. $p \simeq 0.0002$, $p < 0.0001$, $p \simeq 0.0008$, $p \simeq 0.0037$, $p \simeq 0.0099$), a result consistent with our findings, which show how catastrophic the use of the delta-lognormal distribution is. However, the results are also not favorable for European golden plover ($n_+ = 103$), yet none of the tests indicate rejection of the null hypothesis (resp. $p \simeq 0.68$, $p \simeq 0.95$, $p \simeq 0.52$, $p \simeq 0.39$, $p \simeq 0.47$). As correctly reported by Syrjala (2000) "*Failure to reject a goodness-of-fit test is not sufficient to ensure that the data are adequately close to a lognormal distribution to obtain reasonably unbiased estimates*". Thus, testing the distribution of positive data, as suggested by Trenkel and Rochet (2003) or Greenstreet et al. (2010), is not sufficient to answer the genuine question. Actually, the question we are concerned with is not whether the data truly follow a delta-lognormal distribution (or some other distribution), which we know to be false without having to conduct a statistical test (Chambers et al., 1983, p. 192), but rather how to decide that the empirical data distribution is sufficiently close to a delta-lognormal distribution — in a way to be agreed upon by statisticians — to warrant use in a prediction context. Hence, as written by Gelman et al. (2014, p. 190) "*the problem* [⋯] *is not an inability of the models to fit the data, but an inherent inability of the data to distinguish between alternative models that have different implications for* [prediction] *of the population total* [⋯]".

One approach would be to use an adequate measure of dissimilarity between the empirical and theoretical distributions (distance or divergence, for a review see Cha, 2007; Basseville, 2013) and to know beyond which value of this dissimilarity one should not use the delta-lognormal distribution in the context of prediction; however, choosing (or designing) a dissimilarity measure that accounts for all aspects of the problem is anything but obvious. Moreover, this question goes beyond the specific case of the delta-lognormal distribution, which implies that one should be able to answer the question correctly for any distribution, which again seems to be a very difficult (or even impossible) goal to achieve.

As suggested by Zipkin et al. (2014), beyond goodness-of-fit assessment, a Monte Carlo simulation study can help in selecting a suitable distribution. The Monte Carlo simulation described in Section (5.3) answers the question, while being general in principle. The use of contamination by a superpopulation directly defined from the data at hand goes beyond the scope of this article and can be used to investigate the robustness of any method that assumes a parametric distribution model, including delta-generalized linear models (delta-GLMs) (Stefánsson, 1996; Fletcher et al., 2005) and delta-generalized additive models (delta-GAMs) (Li et al., 2011; Berg et al., 2014; Rubec et al., 2016) (Type VII models, Table 1) or geostatistical delta-generalized linear mixed models (delta-GLMMs) (Thorson et al., 2015) (Type VIII models, Table 1).

6.5. Choosing the type of superpopulation model

In the absence of auxiliary variables and spatial (or temporal or spatiotemporal) autocorrelation, we either ignore the question of the shape of the distribution (Type I model), in which case — if sampling is ignorable — the prediction gives the same result as the SRSWOR-based estimate, or we take the shape of the distribution into account (Type IV model). The main question then is whether this latter choice is acceptable both in terms of actual bias and efficiency. Notably, in the field of parameter estimation from a finite population, to our knowledge, survey statisticians do not venture to make such a strong distributional assumption, so Type IV models are rarely used (Ståhl, 2016). More often than not, the assumed model is too restrictive to describe the dataset, which leads to severe biases (Chen et al., 2004). The two examples documented in this article are good illustrations of this point (Section 5).

Faced with the difficulties posed by predictions based on a Type IV model, at least two nonexclusive attitudes are possible: (a) assess the robustness of the predictions (bias, efficiency) for the chosen distribution, or (b) seek an approach that is robust to model departure. In regard to using a Type IV model with a given distribution model, we advocate relying on Monte Carlo simulation predictive checks such as those described in Section 5.3.

7. Perspectives

7.1. Tackling the robustness issue

The use of a distribution alone (Type IV model) raises several related questions that deserve further consideration. (i) With which distribution (s) can the predictions prove sufficiently robust to be useful in practice? (ii) For a given distribution, is there a difference in robustness between using UMVUE (if available) and the maximum likelihood estimator (MLE)? (iii) How can empirical distributions that often have extreme values be fit? More generally, selecting a model that assumes a statistical distribution (Type IV, VI, VII or VIII models, Table 1) remains a challenge because of the influential data in the right tail of the distribution (Favre-Martinoz et al., 2021). Again quoting Gelman et al. (2014, p. 190), "*In order to* [predict] *the total (or the mean), not only do we need a model that reasonably fits the observed data, but we also need a model that provides realistic extrapolations beyond the region of the data.* [Predictions] *of* [the total] *depend strongly on the upper extreme of the distribution* [⋯]".

The question of influential values also concerns estimation based on a sampling design (e.g., Hulliger, 1995; Beaumont et al., 2013). However, questions (i), (ii) and (iii) are not relevant in a design-based framework when only the total (or mean) and the sampling variance are to be estimated. From this perspective, design-based estimation has a decisive advantage over model-based prediction in terms of objectivity, both in terms of bias and variance. It is for this reason that we also find in the literature the terminology opposing the *model-free* (or *assumption-free*) nature of the design-based approach to the *model-dependent* nature of the superpopulation-based approach (see Hansen et al., 1983; Särndal, 1985 and references therein). When coping with model-based prediction, it is crucial to acknowledge that *"If no estimate is made of the* [actual] *bias in the model-based* [predictor]*, only partial information is available on the* [predictor]*'s mean squared error.* [⋯] *In consequence the* [predictor] *suffers the serious limitation of lacking a true measure of its precision. There is the danger that the model-based standard error will be interpreted as an overall measure of precision, a procedure which may attribute to the* [predictor] *far greater precision than is appropriate."* (Kalton, 1983, p. 180).

7.2. Comparing design-based vs. model-based inferences

We have considered thus far only the estimation/prediction of the parameter of interest (the mean or total) and an intermediate step, that is, the variance estimation for the estimation/prediction error, but to make an inference, it is necessary to consider interval estimation by means of a confidence interval (design-based approach) or a prediction interval (superpopulation model-based approach). The design-based approach retains its model-free nature in regard to estimating the sampling variance, but in the strict sense, it loses this nature in terms of obtaining confidence intervals, an aspect aptly noted by Borchers et al. (2002, p. 49). A common misconception is to think that the design-based framework leads to unbiased confidence intervals (e.g., Albert et al., 2010, p. 1029; McGarvey et al., 2016, p. 244). Except for a possible miswording from the authors, they refer to $(1-\alpha)$ confidence intervals for which the probability of covering false values of the parameter of interest is less than or equal to $(1-\alpha)$ (Graybill, 1976, pp. 87–88); to our knowledge, there is no such guarantee under the design-based approach. Actually, what interests the practitioner most is that the actual coverage probability should be very close to its nominal level and that noncoverage probabilities on both sides of the interval are well balanced. Regarding design-based confidence intervals, for statistical/mathematical fundamentals, we refer the reader, for instance, to Bellhouse (2001); Prášková and Sen (2009); Knottnerus (2009). Let us simply note informally that (i) one has to adopt an adequate asymptotic framework for sampling without replacement of finite populations, which leads to a specific version of the central limit theorem (CLT); (ii) for a particular statistic (or family of statistics) — for example, the expansion estimator — asymptotic properties depend on the design; and (iii) properties at finite distance may be very different from asymptotic properties (slow convergence) when the shape of the statistical distribution of the variable of interest departs substantially from symmetry. As one can imagine, the subject is of a rather complex theoretical nature. On a practical level, in the field considered in this article, using the normal distribution does not necessarily lead to confidence intervals with right and left noncoverage probabilities close to their prescribed nominal levels $(\alpha/2)$; the required confidence interval must be asymmetric, and this asymmetry may not be negligible. This issue requires a thorough examination that could not be conducted within the scope of this article and will be addressed on other occasions.

We consider that there is a need to first document the issues regarding the computation of accurate confidence and prediction intervals before being able to conduct a proper and fair comparison of design-based and model-based inferences.

7.3. Taking into account imperfect detection

We have not addressed the issue of imperfect detection, which either implicitly assumes that there is no observational error — in the sense of a discrepancy between true and observed abundance — which is not realistic, or considers that a relative abundance (abundance index) rather than an absolute abundance is estimated. When addressing the consideration of observational error in addition to sampling error — sampling error exists whether the approach is design-based or model-based — it is possible to treat both at the same time within the formalism of each approach. For example, we can generalize the estimator by expansion to take into account estimated detection probabilities (e.g., Steinhorst and Samuel, 1989; Thompson and Seber, 1994). However, it is essential to understand that the use of the (probability) sampling theory formalism when the sample has not actually been selected by a design is not, in the strict sense, a design-based approach. In other words, it is actually a probability modeling of the observation process, which is combined with the design-based approach within the same formalism. At this stage, it is worth recalling that a formula is not sufficient for full understanding of the underlying conceptual framework, as illustrated by the case of the expansion estimator/predictor coincidence. In this regard, this sort of formal coincidence often occasions considerable confusion (Brewer et al., 2009, p. 13).

As a matter of fact, the observation error cannot be taken into account in a purely design-based framework if we do not control the sampling of the individuals counted (Borchers et al., 2002, p. 48). This does not necessarily require being able to list them but at least being able to observe them with exactly known probabilities. In contrast, both types of errors can be taken into account in a purely model-based framework. It is also possible using a hybrid design-model-based approach, for instance, when estimating an estimator's variance that takes into account both sampling and observation errors (e.g., Aubry et al., 2012).

7.4. The need for further studies

In absolute terms, the model-based approach encompasses more than just the superpopulation approach in the frequentist sense. Indeed, in this article, we have considered a global frequentist framework, while a Bayesian approach is of course also possible (e.g., Ghosh and Meeden, 1997; Fieberg et al., 2013; Mendoza et al., 2021; Gelman et al., 2014, Section 8.3).

Contending with the fundamental topic addressed in this article involves being able to answer a sequence of questions. (i) Is the type of approach frequentist or not? (ii) In the case of a frequentist approach, is it dependent or not on a model? (iii) In the case of an approach that depends on a model, what type of model should be used? (iv) For a well-identified model type, what specification should be adopted? and so on. The subject is therefore extremely rich and impossible to tackle adequately in just a few pages (Hankin et al., 2019, p. 3). In this respect, the section by Borchers et al. (2002, Section 3.2) devoted to the *design-based* and *model-based* comparison, although welcome, is insufficient to do justice to the subject.

There are many possibilities of hybridization to various degrees between the use of probability sampling and that of models, whether at the design or estimation stage (e.g., Kalton, 1983; Särndal et al., 1992; Tillé, 2020, Chapter 13), and the future undoubtedly lies in the intelligent use of the two frameworks, rather than in a dogmatic confrontation, such as may have occurred in the past among statisticians (Iachan, 1984; Little, 2004; Sterba, 2009). From an operational perspective it is at least what seems most desirable in ecology; as written by Williams and Brown (2019), "*ecologists must understand the strengths and limitations of each approach in order to tailor designs and analyses to specific questions and produce unbiased inferences from survey data*". In the context of abundance estimation, to meet this salutary injunction, the use of a design- and/or model-based approach should be further documented by

repeating and extending the kind of investigations conducted in this article with other types of superpopulation models (Table 1), other datasets, and other contexts (e.g., with small populations of spatial units).

We have considered a situation in which there is no exploitable autocorrelation nor auxiliary variables. In this default situation, the choice of a probability sampling design without replacement naturally leads to the SRSWOR rather than to a more complex design that would require more information. However, the comparison between design-based estimation and model-based prediction should also be made in situations involving more complex sampling designs (stratified, unequal probability, multistage, multiphase, balanced, etc.).

As one can imagine, the topic addressed in this article is very vast. It raises fundamental aspects that need to be properly discussed at a sufficient level of detail. The result should be operational guidance for ecologists and population biologists whose primary goal is to obtain abundance estimates they can trust. We hope that by providing such syntheses, more scientists and/or managers will be motivated to think more carefully about sampling design and ways to assess biological population size.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The datasets used in this article were collected under the direction of Bertrand Trolliet (retired) at the Office National de la Chasse et de la Faune Sauvage - ONCFS (now Office Français de la Biodiversité - OFB); the first author thanks him for all the methodological issues raised during the time of their collaboration between 2005 and 2010. We thank Dr. Clément Calenge for bringing our attention to Gelman et al. (2014, pp. 189–190) and Pr. Nigel Yoccoz for his commentaries on a previous version of this article. We also thank American Journal Experts (AJE) for helping us to improve the English of this article.

Appendix A. Notation conventions

A.1. Random variables

Random variables are typically written in upper case, whereas their realizations are written in lower case (e.g., Edwards and Auger-Méthé, 2019); this causes difficulties in the context of this article and represents an obstacle to the expression. Another possibility is to use the so-called *Dutch notation*, which consists in underlining symbols for the random counterparts of fixed quantities (Hemelrijk, 1966). Although this convention for distinguishing between fixed and random quantities is more generally applicable than the previous one, it has the disadvantage of making the notation considerably more cumbersome. We choose to follow Chambers and Clark (2012, p. 7), who do not distinguish random variables from their realizations using notation but simply refer to the context.

A.2. Cumulants, moments and skewness

We note the mean of y on a finite set A:

$$\overline{y}_A = \frac{1}{|A|}\sum_{i\in A} y_i = \frac{1}{|A|} t_A \tag{A.1}$$

where $|A|$ is the cardinality of A and t_A is the total of y on A. In the context of finite populations, the mean population $\overline{y}_{/\!/}$ is the first cumulant. We note the variance of y on A:

$$S_A^2 = \frac{1}{|A|-1}\sum_{i\in A}(y_i - \overline{y}_A)^2 \tag{A.2}$$

With this notation, the population variance $S_{/\!/}^2$ (second cumulant) has $N-1$ as a denominator (Thompson, 1997, p. 27, Eq. 2.60) and follows the convention in use in survey sampling theory (see Cochran, 1977, p. 23). It also makes sense in that the population variance is then an unbiased estimator of the superpopulation variance from which the population is drawn at random (O'Neill, 2014, p. 283).

At the level of an infinite population of model ξ, κ_r is the cumulant of order r (e.g., Kendall, 1945). The expectation and variance in the ξ-model are the first two cumulants and are thus denoted as $\mathrm{E}_\xi(y) = \kappa_1$ and $\mathrm{V}_\xi(y) = \kappa_2$. The central moment of order r is written as:

$$\mu_r = \mathrm{E}_\xi[(y-\kappa_1)^r] \tag{A.3}$$

In particular, we have $\mu_2 = \kappa_2$ and $\mu_3 = \kappa_3$. Even if it is not necessarily the best possible definition, in this article, skewness is classically defined as:

$$\gamma_1 = \mathrm{E}_\xi\left[\left(\frac{y-\kappa_1}{\kappa_2^{1/2}}\right)^3\right] = \frac{\kappa_3}{\kappa_2^{3/2}} \tag{A.4}$$

Consider a mixture of n random variables y_i with normalized weights w_i (in the sense that $\sum_i w_i = 1$) and expectations μ_i. By linearity of the expectation operator, for the resulting random variable y, we obtain:

$$\mathrm{E}_\xi[y] = \mu = \sum_{i=1}^{n} w_i \mu_i \tag{A.5}$$

The central moment of order r can be written as:

$$\mathrm{E}_{\xi}[(y-\mu)^r] = \sum_{i=1}^{n} w_i \sum_{k=0}^{r} \binom{r}{k} (\mu_i - \mu)^{r-k} \mathrm{E}_{\xi}\left[(y_i - \mu_i)^k\right] \tag{A.6}$$

From (A.6), for $r = 2$, we obtain the variance as:

$$\mathrm{E}_{\xi}\left[(y-\mu)^2\right] = \sigma^2 = \sum_{i=1}^{n} w_i \left(\mu_i^2 + \sigma_i^2\right) - \mu^2 \tag{A.7}$$

Appendix B. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.ecolind.2022.109394.

References

Abramowitz, M., Stegun, I., 1972. Handbook of mathematical functions. In: Tenth printing, with corrections. Dover Publications, New York, USA.

Aitchison, J., Brown, J., 1969. The lognormal distribution (reprinted from 1957 edition). Cambridge University Press, Cambridge, UK.

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

Aubry, P., 2021. On the correct implementation of the Hanurav-Vijayan selection procedure for unequal probability sampling without replacement. Commun. Stat. - Simul. Comput.

Aubry, P., 2022. On evaluating the efficiency of the delta-lognormal mean estimator and predictor. MethodsX 9.

Aubry, P., Debouzie, D., 2000. Geostatistical estimation variance for the spatial mean in two-dimensional systematic sampling. Ecology 81, 543–553.

Aubry, P., Debouzie, D., 2001. Estimation of the mean from a two-dimensional sample: the geostatistical model-based approach. Ecology 82, 1484–1494.

Aubry, P., Guillemain, M., Sorrenti, M., 2020. Increasing the trust in hunting bag statistics: why random selection of hunters is so important. Ecol. Ind. 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. Ind. 18, 599–607.

Basseville, M., 2013. Divergence measures for statistical data processing. An annotated bibliography. Signal Process. 93, 621–633.

Beaumont, J., Haziza, D., Ruiz-Gazen, A., 2013. A unified approach to robust estimation in finite population sampling. Biometrika 100, 555–569.

Bellhouse, D., 2001. The central limit theorem under simple random sampling. Am. Stat. 55, 352–357.

Berg, C., Nielsen, A., Kristensen, K., 2014. Evaluation of alternative age-based methods for estimating relative abundance from survey data in relation to assessment models. Fish. Res. 151, 91–99.

Blom, G., 1958. Statistical estimates and transformed Beta-variables. Almqvist & Wicksell, Stockholm, Sweden.

Bolfarine, H., Zacks, S., 1992. Prediction theory for finite populations. Springer, New York, USA.

Borchers, S., Buckland, S., Zucchini, W., 2002. Estimating animal abundance. Closed populations, Springer, London, UK.

Brewer, K., Gregoire, T., 2009. Introduction to survey sampling. In: Pfeffermann, D., Rao, C.R. (Eds.), Handbook of Statistics 29B. Sample surveys: inference and analysis. Elsevier, Oxford, UK, pp. 9–37.

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

Buckland, S., Anderson, D., Burnham, K., Laake, J., Borchers, D., Thomas, L., 2001. Introduction to distance sampling. Estimating abundance of biological populations. Oxford University Press, Oxford, UK.

Buckland, S., Anderson, D., Burnham, K., Laake, J., Borchers, D., Thomas, L., 2004. Advanced distance sampling. Estimating abundance of biological populations. Oxford University Press, Oxford, UK.

Cassel, C., Särndal, C., Wretman, J., 1977. Foundations of inference in survey sampling. John Wiley & Sons, London, UK.

Cha, S., 2007. Comprehensive survey on distance/similarity measures between probability density functions. Int. J. Math. Models Methods Appl. Sci. 4, 300–307.

Chambers, J., Cleveland, W., Kleiner, B., Tukey, P., 1983. Graphical methods for data analysis. Wadsworth & Brooks/Cole, Monterey, California, USA.

Chambers, R., Clark, R., 2012. An introduction to model-based survey sampling with applications. Oxford University Press, Oxford, UK.

Chen, J., Thompson, M., Wu, C., 2004. Estimation of fish abundance indices based on scientific research trawl surveys. Biometrics 60, 116–123.

Christman, M., 2019. Review of estimation methods for parameters of the delta-lognormal distribution. Technical Report. MCC Statistical Consulting LLC. Gainesville, Florida, USA.

Clark, J., Bjørnstad, O., 2004. Population time series: process variability, observation errors, missing values, lags, and hidden states. Ecology 85, 3140–3150.

Cochran, W., 1939. The use of analysis of variance in enumeration by sampling. J. Am. Stat. Assoc. 34, 492–510.

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

Cochran, W., 1978. Laplace's ratio estimator. In: David, H.A. (Ed.), Contributions to survey sampling and applied statistics. Papers in the honor of H.O. Hartley. Academic Press, New York, USA, pp. 3–10.

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

Deming, W., 1953. On the distinction between enumerative and analytic surveys. J. Am. Stat. Assoc. 48, 244–255.

Deming, W., Stephan, F., 1941. On the interpretation of censuses as samples. J. Am. Stat. Assoc. 36, 45–49.

Dennis, B., Patil, G., 1988. Applications in ecology. In: Crow, E.L., Shimizu, K. (Eds.), Lognormal distributions: theory and applications. Marcel Dekker, New York, USA, pp. 303–330.

Dennis, B., Ponciano, J., Lele, S., Taper, M., Staples, D., 2006. Estimating density dependence, process noise, and observation error. Ecol. Monogr. 76, 323–341.

Dufrêne, M., Legendre, P., 1997. Species assemblages and indicator species: the need for a flexible asymmetrical approach. Ecol. Monogr. 67, 345–366.

Eberhardt, L., Thomas, J., 1991. Designing environmental field studies. Ecol. Monogr. 61, 53–73.

Edwards, A., Auger-Méthé, M., 2019. Some guidance on using mathematical notation in ecology. Methods Ecol. Evol. 10, 92–99.

Edwards, D., 1998. Issues and themes for natural resources trend and change detection. Ecol. Appl. 8, 323–325.

Favre-Martinoz, C., Haziza, D., Beaumont, J., 2021. Efficient nonparametric estimation for skewed distributions. Can. J. Stat. 49, 471–496.

Fieberg, J., Alexander, M., Tse, S., St. Clair, K., 2013. Abundance estimation with sightability data: a Bayesian data augmentation approach, Methods Ecol. Evol. 4, 854–864.

Firth, D., Bennett, K., 1998. Robust models in probability sampling. J. R. Stat. Soc. Ser. B 60, 3–21.

Fisher, R., 1932. Inverse probability and the use of likelihood. Math. Proc. Cambridge Philos. Soc. 28, 257–261.

Fletcher, D., Mackenzie, D., Villouta, E., 2005. Modelling skewed data with many zeros: a simple approach combining ordinary and logistic regression. Environ. Ecol. Stat. 12, 45–54.

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

Ghosh, M., Meeden, G., 1997. Bayesian methods for finite population sampling. Chapman & Hall, London, UK.

Graybill, F., 1976. Theory and application of the linear model. Wadsworth & Brooks/Cole, Pacific Grove, California, USA.

Greenstreet, S., Holland, G., Guirey, E., Armstrong, E., Fraser, H., Gibb, I., 2010. Combining hydroacoustic seabed survey and grab sampling techniques to assess "local" sandeel population abundance. ICES J. Mar. Sci. 67, 971–984.

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

Hahn, G., 1969. Factors for calculating two-sided prediction intervals for samples from a normal distribution. J. Am. Stat. Assoc. 64, 878–888.

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

Hansen, M., Madow, W., Tepping, B., 1983. An evaluation of model-dependent and probability-sampling inferences in samples surveys (with discussion). J. Am. Stat. Assoc. 78, 776–793.

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

Hemelrijk, J., 1966. Underlining random variables. Stat. Neerl. 20, 1–7.

Hiddink, J., 2005. Implications of Liebig's law of the minimum for the use of ecological indicators based on abundance. Ecography 28, 264–271.
Hilborn, R., Mangel, M., 1997. The ecological detective: confronting models with data. Princeton University Press, Princeton, New Jersey, USA.
Hulliger, B., 1995. Outlier robust Horvitz-Thompson estimators. Survey Methodol. 21, 79–87.
Iachan, R., 1984. Sampling strategies, robustness and efficiency: the state of the art. Int. Stat. Rev. 52, 209–218.
Iijima, H., 2020. A review of wildlife abundance estimation models: comparison of models for correct application. Mammal Study 45, 177–188.
Janke, S., Tinsley, F., 2005. Introduction to linear models and statistical inference. John Wiley & Sons, Hoboken, New Jersey, USA.
Johnson, N., Kotz, S., 1970. Continuous univariate distributions – 2. John Wiley & Sons, New York, USA.
Kalton, G., 1983. Models in the practice of survey sampling. Int. Stat. Rev. 51, 175–188.
Kellner, K., Swihart, R., 2014. Accounting for imperfect detection in ecology: a quantitative review. Plos ONE 9, e111436.
Kendall, M., 1945. The advanced theory of statistics, vol. I. Second edition, Charles Griffin, London, UK.
Knottnerus, P., 2009. On asymptotic distributions in random sampling from finite populations. Technical Report. Statistics Netherlands. The Hague, The Netherlands.
Koch, A., 1966. The logarithm in biology. I. Mechanisms generating the log-normal distribution exactly. J. Theor. Biol. 12, 276–290.
Koch, A., 1969. The logarithm in biology. II. Distributions simulating the log-normal. J. Theor. Biol. 23, 251–268.
Krebs, C., 2014. Ecology: the experimental analysis of distribution and abundance, Sixth edition. Pearson Education, Harlow, UK.
Li, Y., Jiao, Y., He, Q., 2011. Decreasing uncertainty in catch rate analyses using Delta-AdaBoost: an alternative approach in catch and bycatch analyses with high percentage of zeros. Fish. Res. 107, 261–271.
Lin, T., Liao, C., 2008. Prediction intervals for general balanced linear random models. J. Stat. Plann. Inference 138, 3164–3175.
Link, R., Koch, G., 1975. Some consequences of applying lognormal theory to pseudolognormal distributions. Math. Geol. 7, 117–128.
Little, R., 2004. To model or not to model? Competing modes of inference for finite population sampling. J. Am. Stat. Assoc. 99, 546–556.
McCrea, R., Morgan, B., 2014. Analysis of capture-recapture data. Chapman & Hall/CRC, Boca Raton, Florida, USA.
McGarvey, R., Burch, P., Matthews, J., 2016. Precision of systematic and random sampling in clustered populations: habitat patches and aggregating organisms. Ecol. Appl. 26, 233–248.
Meire, P., Dereu, J., 1990. Use of the abundance/biomass comparisons method for detecting environmental stress: some considerations based on intertidal macrozoobenthos and bird communities. J. Appl. Ecol. 27, 210–223.
Mendoza, M., Contreras-Cristan, A., Gutierrez-Pena, E., 2021. Bayesian analysis of finite populations under simple random sampling. Entropy 23, 318.
Myers, R., Pepin, P., 1990. The robustness of lognormal-based estimators of abundance. Biometrics 46, 1185–1192.
Myers, R., Pepin, P., 1991. Rejoinder to the letter to the editors from M. Pennington, "On testing the robustness of lognormal-based estimators". Biometrics 47, 1623–1624.
Nathan, G., 2011. Superpopulation models in survey sampling. In: Lovric, M. (Ed.), International encyclopedia of statistical science. Springer, Berlin, Germany, pp. 1575–1576.
Nichols, J., 2005. Modern open-population capture-recapture models. In: Amstrup, S.C., McDonald, T.L., Manly, B.F.J. (Eds.), Handbook of capture-recapture analysis. Princeton University Press, Princeton, New Jersey, USA, pp. 88–123.
Niemi, G., McDonald, M., 2004. Application of ecological indicators. Annu. Rev. Ecol. Syst. 35, 89–111.
Norton-Griffiths, M., 1978. Counting animals. African Wildlife Leadership Foundation, Nairobi, Kenya.
Olsen, A., Sedransk, J., Edwards, D., Gotway, C., Liggett, W., Rathbun, S., Reckhow, K., Young, L., 1999. Statistical issues for monitoring ecological and natural resources in the United States. Environ. Monit. Assess. 54, 1–45.
O'Neill, B., 2014. Some useful moment results in sampling problems. Am. Stat. 68, 282–296.
Parkin, T., Robinson, J., 1992. Analysis of lognormal data, in: Stewart, B.A. (Ed.), Advances in soil science. Volume 20. Springer, New York, USA, pp. 193–235.
Pearson, K., 1928. On a method of ascertaining limits to the actual number of marked members in a population of given size from a sample. Biometrika 20A, 149–174.
Pennington, M., 1983. Efficient estimators of abundance, for fish and plankton surveys. Biometrics 39, 281–286.
Pennington, M., 1991. On testing the robustness of lognormal-based estimators. Biometrics 47, 1623.
Pennington, M., 1996. Estimating the mean and variance from highly skewed marine data. Fish. Bull. 98, 498–505.
Pennington, M., Stromme, T., 1998. Surveys as a research tool for managing dynamic stocks. Fish. Res. 37, 97–106.
Pfeffermann, D., 1993. The role of sampling weights when modeling survey data. Int. Stat. Rev. 61, 317–337.
Prášková, Z., Sen, P., 2009. Asymptotics in finite population sampling. In: Pfeffermann, D., Rao, C.R. (Eds.), Handbook of statistics 29B. Sample surveys: inference and analysis. Elsevier, Oxford, UK, pp. 489–522.
Rivoirard, J., Simmonds, J., Foote, K., P., F., Bez, N., 2000. Geostatistics for estimating fish abundance. Blackwell Science, Oxford, UK.
Royle, J., Chandler, R., Sollmann, R., Gardner, B., 2014. Spatial capture-recapture. Elsevier/Academic Press, Waltham, Massachusetts, USA.
Rubec, P., Kiltie, R., Leone, E., Flamm, R., McEachron, L., Santi, C., 2016. Using delta-generalized additive models to predict spatial distributions and population abundance of juvenile Pink Shrimp in Tampa Bay, Florida. Marine and Coastal Fisheries: Dynamics, Management, and Ecosystem Science 8, 232–243.
Särndal, C., 1978. Design-based and model-based inference in survey sampling (with discussion). Scand. J. Stat. 5, 27–52.
Särndal, C., 1985. How survey methodologists communicate. J. Off. Stat. 1, 49–63.
Särndal, C., Swensson, B., Wretman, J., 1992. Model assisted survey sampling. Springer, New York, USA.
Särndal, C., Wright, R., 1984. Cosmetic forms of estimators in survey sampling. Scand. J. Stat. 11, 146–156.
Schwarz, C., Seber, G., 1999. Estimating animal abundance: review III. Stat. Sci. 14, 427–456.
Seber, G., 1982. The estimation of animal abundance and related parameters, Second edition. Charles Griffin, London, UK.
Seber, G., 1986. A review of estimating animal abundance. Biometrics 42, 267–292.
Seber, G., 1992. A review of estimating animal abundance II. Int. Stat. Rev. 60, 129–166.
Shimizu, K., 1988. Point estimation. In: Crow, E.L., Shimizu, K. (Eds.), Lognormal distributions: theory and applications. Marcel Dekker, New York, USA, pp. 27–86.
Shimizu, K., Crow, E., 1988. History, genesis, and properties. In: Crow, E.L., Shimizu, K. (Eds.), Lognormal distributions: theory and applications. Marcel Dekker, New York, USA, pp. 1–25.
Shimizu, K., Iwase, K., 1981. Uniformly minimum variance unbiased estimation in lognormal and related distributions. Commun. Stat. – Theory Methods 10, 1127–1147.
Smith, A., Anderson, M., Pawley, M., 2017. Could ecologists be more random? Straightforward alternatives to haphazard spatial sampling. Ecography 40, 1251–1255.
Smith, S., 1988. Evaluating the efficiency of the Δ-distribution mean estimator. Biometrics 44, 485–493.
Smith, S., 1990. Use of statistical models for the estimation of abundance from groundfish trawl survey data. Can. J. Fish. Aquat. Sci. 47, 894–903.
Spellerberg, I., 2005. Monitoring ecological change. Cambridge University Press, Cambridge, UK.
Ståhl, O., 2016. Point estimation using tail modelling for right skew populations. J. Stat. Comput. Simul. 86, 2073–2088.
Stefánsson, G., 1996. Analysis of groundfish survey abundance data: combining the GLM and delta approaches. ICES J. Mar. Sci. 53, 577–588.
Steinhorst, R., Samuel, M., 1989. Sightability adjustment methods for aerial surveys of wildlife populations. Biometrics 45, 415–425.
Sterba, S., 2009. Alternative model-based and design-based frameworks for inference from samples to populations: from polarization to integration. Multivar. Behav. Res. 44, 711–740.
Sugden, R., Smith, T., 1984. Ignorable and informative designs in survey sampling inference. Biometrika 71, 495–506.
Syrjala, S., 2000. Critique on the use of the delta distribution for the analysis of trawl survey data. ICES J. Mar. Sci. 57, 831–842.
Thompson, M., 1997. Theory of sample surveys. Chapman & Hall, London, UK.
Thompson, S., Seber, G., 1994. Detectability in conventional and adaptive sampling. Biometrics 50, 712–724.
Thorson, J., Shelton, A., Ward, E., Skaug, H., 2015. Geostatistical delta-generalized linear mixed models improve precision for estimated abundance indices for west coast groundfishes. ICES J. Mar. Sci. 72, 1297–1310.
Tillé, Y., 2006. Sampling algorithms. Springer, New York, USA.
Tillé, Y., 2020. Sampling and estimation from finite populations. John Wiley & Sons, Hoboken, New Jersey, USA.
Trenkel, V., Rochet, M., 2003. Performance of indicators derived from abundance estimates for detecting the impact of fishing on a fish community. Can. J. Fish. Aquat. Sci. 60, 67–85.
Tukey, J., 1960. A survey of sampling from contaminated distributions. In: Olkin, I. (Ed.), Contributions to probability and statistics: essays in honor of Harold Hotelling. Stanford University Press, Stanford, California, USA, pp. 448–485.
Webb, S., 2021. Book review: Sampling theory for the ecological and natural resource sciences. David G. Hankin, Michael S. Mohr, and Kenneth B. Newman. 2019. Oxford University Press, Oxford, United Kingdom. 368 pp. Journal of Wildlife Management 85, 816–817.
Wen, Z., Pollock, K., Nichols, J., Waser, P., 2011. Augmenting superpopulation capture-recapture models with population assignment data. Biometrics 67, 691–700.
White, G., 2005. Correcting wildlife counts using detection probabilities. Wildlife Res. 32, 211–216.
Whitmore, G., 1986. Prediction limits for a univariate normal observation. Am. Stat. 40, 141–143.
Williams, B., Brown, E., 2019. Sampling and analysis frameworks for inference in ecology. Methods Ecol. Evol. 10, 1832–1842.
Zipkin, E., Leirness, J., Kinlan, B., O'Connell, A., Silverman, E., 2014. Fitting statistical distributions to sea duck count data: implications for survey design and abundance estimation. Stat. Methodol. 17, 67–81.