Environ Ecol Stat
DOI 10.1007/s10651-016-0342-2

Imperfect observations in ecological studies

Hideyasu Shimadzu[1,2] · **Scott D. Foster**[3] · **Ross Darnell**[4]

Received: 3 March 2015 / Revised: 7 October 2015

Abstract Every ecological data set is the result of sampling the biota at sampling locations. Such samples are rarely a census of the biota at the sampling locations and so will inherently contain biases. It is crucial to account for the bias induced by sampling if valid inference on biodiversity quantities is to be drawn from the observed data. The literature on accounting for sampling effects is large, but most are dedicated to the specific type of inference required, the type of analysis performed and the type of survey undertaken. There is no general and systematic approach to sampling. Here, we explore the unification of modelling approaches to account for sampling. We focus

Handling Editor: Pierre Dutilleul.

Electronic supplementary material The online version of this article (doi:10.1007/s10651-016-0342-2) contains supplementary material, which is available to authorized users.

✉ Hideyasu Shimadzu
hs50@st-andrews.ac.uk; H.Shimadzu@lboro.ac.uk

Scott D. Foster
scott.foster@csiro.au

Ross Darnell
ross.darnell@csiro.au

[1] Centre for Biological Diversity and Scottish Oceans Institute, University of St Andrews, Dyers Brae House, St Andrews, Fife KY16 9TH, UK

[2] Department of Mathematical Sciences, Loughborough University, Loughborough, Leicestershire LE11 3TU, UK

[3] Commonwealth Scientific and Industrial Research Organisation, GPO Box 1538, Hobart, TAS 7001, Australia

[4] Commonwealth Scientific and Industrial Research Organisation, PO Box 2583, Brisbane, QLD 4001, Australia

Published online: 08 April 2016

on individuals in ecological communities as the fundamental sampling element, and show that methods for accounting for sampling at the species level can be equated to individual sampling effects. Particular emphasis is given to the case where the probability of observing an individual, when it is present at the site sampled, is less than one. We call these situations ‘imperfect observations’. The proposed framework is easily implemented in standard software packages. We highlight some practical benefits of this formal framework: the ability of predicting the true number of individuals using an expectation that conditions on the observed data, and designing appropriate survey plans accounting for uncertainty due to sampling. The principles and methods are illustrated with marine survey data from tropical northern Australia.

Keywords Compound distributions · Detection probability · Ecological modelling · Marine surveys · Sampling · Species distribution models (SDMs)

1 Introduction

One of the long-standing challenges in quantitative ecology is to model biodiversity quantities, such as a species’ presence/absence, abundance or biomass. In the last decade or so there has been rapid methodological developments and a wide range of applications (Guisan and Zimmermann 2000; Ferrier and Guisan 2006; Pitcher et al. 2007; Elith and Leathwick 2009; Gattone and Battista 2009; Lozier et al. 2009; Franklin 2010; Bax 2011). The core of the modelling challenge is typically a regression-type problem: how to relate biodiversity to a set of descriptors (covariates) such as environmental or anthropogenic variables. It involves describing the variability in the data into parts that are common to all data (the signal), and a part that remains unexplained (the noise). There are two types of variance that can adversely affect the model: one is due to sampling biodiversity (considered in this article), and the other is due to sampling/predicting the covariates (e.g. Foster et al. 2012; Stoklosa et al. 2015).

It is tempting, when modelling biodiversity attributes, to ignore any variance in the data due to sampling (including imperfect observations). This assumes that the manner in which data were collected is unimportant, or that it will simply add to the random part of the model and not the signal about biodiversity. However, this requires assumptions that are unlikely to be met. The unfortunate implication is that sampling issues can affect inferences. Accounting for sampling effects within a model requires careful consideration as it tends to vary from one survey to the next.

One sampling effect, which is often ignored, arises when the data are a sample (not a census) of the biological material at a sample location. We refer to these data as *imperfect observations*, and they are the central topic of the work presented here. An example of where imperfect observations occur is in marine surveys, where a large amount of biotic material is obtained (too much for scientific processing of all the material). The practical method to quantify all of the biotic material is to sample the different catch (a process called sub-sampling). Sub-sampling can take a number of forms—sample from all the biotic material as a single group, or the sample different broad taxonomic or size strata. This process adds another layer of variability into the data.

The methods presented here have broader application. The effect of imperfect observations is explored in this work and simple methods for adjusting statistical models for analysis of these types of data are presented. The model unifies many of the disparate research areas that consider imperfect observations, which has not been done before. One key point of distinction in previous approaches for accounting for imperfect observations is whether the focus is on sampling an individual organism (*individual detectability*), or a species (*species detectability*). For example, species abundances and biomass are related with individual detectability and has been studied as 'ascertainment' (Fisher 1934; Fisher et al. 1943; Rao 1965), 'detection probability' (Borchers et al. 2002; Buckland et al. 2004), or as 'attenuation' (Shimadzu and Darnell 2015). In contrast, species presence/absence and richness are more related with species detectability, and it has been studied as 'rarefaction' (Sanders 1968; Hurlbert 1971; Simberloff 1972; Heck et al. 1975) and 'occupancy' (MacKenzie et al. 2002). However, these approaches concentrate on species-level data and fail to exploit the fact that for the species to be detected at least one of the individuals needs to be sampled. So individual detection must play a pivotal role in understanding species detection.

Our approach utilises a compound distribution of the possible number of individuals at a site. It highlights the precise data needed to disentangle the number of individuals and the probability of sampling. We present a modelling framework, which is trivially implemented in software packages, to handle imperfect observations (including detectability issues and sub-sampling). The formal modelling framework has some practical benefits as well—predictions of the true number of individuals at any sampling site can be made through predictive distributions and the effect of imprecise observations can be incorporated easily when designing surveys. The principles and methods are illustrated throughout the manuscript with two marine data sets from tropical northern Australia. Both of these examples are for a particular case of imperfect observations, namely subsampling. However, we note that the methods presented are for a wider class of applications—any situations where the probability of observation is less than one.

2 Conceptual framework for imperfect observation

Every ecological data set is the result of sampling from a population of interest and every ecological datum can likewise be thought of as a sample (sometimes a census) of the population at a site. Here and elsewhere we use the statistical term 'population' to mean the individuals at a sampling site, as this is our prime interest. We note however, that much of the suggested framework could extend to a broader definition of population. For notation convenience, we omit possible site subscripts. Formally, the principle of sampling is the random partition of a population $\mathcal{P} = \{\omega_i\}_{i=1}^{M_0}$, of size $M_0 = |\mathcal{P}|$, into two disjoint categories: the sample $\mathcal{S}$, and the remainder of the population $\mathcal{S}^c = \mathcal{P} \setminus \mathcal{S}$. These two sets are disjoint, $\mathcal{S}^c \cap \mathcal{S} = \emptyset$. Each element of the population, ω_i, is typically an individual organism but it may also be a colony or a family in certain situations (e.g. corals and sponges). We shall use the 'individual' nomenclature to describe all possibilities. Note that we treat M_0 as random throughout the paper, in order to dealiniate the extent to which the expected abundance, $\mathrm{E}[M_0]$, responds to different environment conditions; more details will be discussed in the later sections.

For ease of exposition, we introduce a random variable Z_i to indicate whether the population's i-th element is in the sample $\mathcal{S}$ or not; it is defined as $Z_i = I(\omega_i \in \mathcal{S})$. This simple variable enables us to efficiently describe key biodiversity measures such as species abundance, presence/absence and biomass of the population under study.

Abundance Species abundance in a sample, $M_1 = |\mathcal{S}|$, is given as the compound form

$$M_1 = \begin{cases} \sum_{i=1}^{M_0} Z_i, & (M_0 > 0); \\ 0, & (M_0 = 0). \end{cases} \tag{1}$$

Species presence/absence Species presence/absence in a sample Y_1 can be defined by using an indicator function $I(\cdot)$ as

$$Y_1 = I(M_1 > 0). \tag{2}$$

Biomass Species biomass V_1 in a sample can also be defined in a compound form as an extension of abundance. Let W_i be the weight of the i-th organism then biomass becomes

$$V_1 = \sum_{i=1}^{M_0} W_i Z_i.$$

Note that $V_1 = 0$ when $M_0 = 0$ as in Eq. (1). If individual weight W_i's are independent gamma random variables then biomass, V_1, follows a distribution called Poisson–gamma distribution. This formulation has been exploited previously (Foster and Bravington 2013) as a special case of the Tweedie distribution (Jørgenson 1997; Dunn and Smyth 2005).

These descriptions provide a natural basis to handle imperfect observations. When the observations are imperfect, the probability of an individual being sampled is less than one, that is $\Pr(Z_i = 1) < 1$. We shall refer to this probability as the *individual detection probability*. This probability can be assumed to be common, or varying, over all individuals, $\{\omega_i\}_{i=1}^{M_0}$. The situation when the individual detection probability is constant over individuals is classically described as simple random sampling. Simple random sampling and immediate extensions will be discussed in the following sections.

Another useful construct is the *species detection probability*, which is defined as one minus the probability that none of the individuals of the species is observed: $1 - \prod_{i=1}^{M_0} \{1 - \Pr(Z_i = 1)\}$; this obviously assumes individual independence. The species detection probability is a function of individual detection probability, $\Pr(Z_i = 1)$, and species abundance, M_0.

3 Models for imperfect observation

3.1 Compound distributions

Since we can only deal with the sample (observable) abundance M_1, its probability mass function, $f(m_1|\boldsymbol{x})$ say, plays the pivotal role in modelling. Here, $\boldsymbol{x}$ represents auxiliary information/variables that act as covariates in the regression model. The simple expression for abundance, Eq. (1), gives the mechanism for defining the probability mass function. It becomes a compound distribution (Feller 1968), also known as a generalised distribution (Gurland 1957) and a stopped-sum distribution (Johnson et al. 1992)

$$
\begin{aligned}
f(m_1|\boldsymbol{x}) &= \sum_{m_0=0}^{\infty} f(m_1, m_0|\boldsymbol{x}) \\
&= \sum_{m_0=0}^{\infty} f(m_0|\boldsymbol{x}) f(m_1|m_0),
\end{aligned}
\tag{3}
$$

where conditional independence ($M_1 \perp\!\!\!\perp \boldsymbol{X}|M_0$) has been assumed. This assumption implies that the auxiliary variables play no role in the process of sampling the population. The first term in Eq. (3), $f(m_0|\boldsymbol{x})$, describes the way that the *true*, and not the *sampled* abundance varies with the covariates. This is appealing as a model defined on the sampled abundance cannot give inferences about the true abundance, which is what is ideally sought. The second term, $f(m_1|m_0)$, defines the sampling process. The sampling process does not depend on covariates; it merely relates how the individuals are drawn from the population. Note that this assumption may not match reality, and the sampling process may be dependent on other auxiliary variables. It is possible to relax this assumption, by specifying the joint conditional distribution as $f(m_0|\boldsymbol{x}) f(m_1|m_0, \boldsymbol{x})$. However, we keep this assumption in this work for ease of exposition of our framework.

3.2 Sampling mechanisms

In general, sampling mechanisms are survey specific and can explicitly be described by $f(m_1|m_0)$ in Eq. (3). Two major sampling procedures in ecological studies are: *simple random sampling* and *stratified sampling* (Cochran 1977). We consider these two cases as examples.

Simple random sampling Under the simple random sampling scheme, the probability of observing a sample $\mathcal{S}$ consisting of m_1 individuals from the population $\mathcal{P}$ of m_0 individuals is a multiplication of the probability of each individual being sampled, $\Pr(\omega_i \in \mathcal{S})$, or being not sampled, $1 - \Pr(\omega_i \in \mathcal{S})$. If it is assumed that the probability of being sampled is common among individuals, $\Pr(\omega_i \in \mathcal{S}) = r$ say, then the probability of sampling m_1 individuals from a population with size m_0 is a binomial distribution $\mathsf{Bi}(m_1; m_0, r)$,

$$f(m_1|m_0) = \binom{m_0}{m_1} r^{m_1}(1-r)^{m_0-m_1}.$$

Simple random sampling assumes that the probability of individual detection is homogeneous, which can be inappropriate for some surveys. The idea of stratified sampling, discussed presently, allows the assumption of population homogeneity to be relaxed. This is achieved by defining strata, within which the individuals have a homogeneous probability of being sampled. The sampling probability between strata may vary, inducing population heterogeneity.

Stratified sampling Let $\mathcal{U}_j$ be the j-th stratum in stratified sampling. The population $\mathcal{P}$ then consists of k strata, $\mathcal{P} = \cup_{j=1}^{k}\mathcal{U}_j, \mathcal{U}_j \cap \mathcal{U}_{j'} = \emptyset\ (j \neq j')$, and the individuals in the population are partitioned into the strata. The number of individuals (m_0 say) are randomly partitioned as $\boldsymbol{m}_0 = (m_{01}, m_{02}, \ldots, m_{0k}), m_{0j} = |\mathcal{U}_j|$. This partitioning mechanism can be described by a multinomial distribution, $\mathsf{Mn}(\boldsymbol{m}_0; m_0, \boldsymbol{p})$, with parameters $\boldsymbol{p} = (p_1, p_2, \ldots, p_j)$ giving the probability of belonging to each multinomial class. The samples are then randomly drawn from each stratum j with sampling fraction r_j. This draw is independent between the strata so each strata's sampling process can be described by a binomial distribution with simple random sampling, as above. In terms of the imperfect observation model (3) the probability mass function of the stratified sample is

$$\begin{aligned} f(\boldsymbol{m}_1|m_0) &= f(\boldsymbol{m}_1|\boldsymbol{m}_0) f(\boldsymbol{m}_0|m_0) \\ &= \left\{\prod_{j=1}^{k}\binom{m_{0j}}{m_{1j}} r_j^{m_{1j}}(1-r_j)^{m_{0j}-m_{1j}}\right\}\binom{m_0}{m_{01},\ldots,m_{0k}}\prod_{j=1}^{k} p_j^{m_{0j}}, \end{aligned}$$

where $\sum p_j = 1$.

3.3 Marginal distributions

In theory, the probability mass function $f(m_0|\boldsymbol{x})$ in Eq. (3) can take any plausible form. In this paper, we focus our attention to the commonly used Poisson models although other distributions could be used. That is, $M_0 \sim \mathsf{Po}(\lambda(\boldsymbol{x}))$ where the Poisson probability function is

$$f(m_0|\boldsymbol{x}) = \frac{\lambda(\boldsymbol{x})^{m_0}}{m_0!} e^{-\lambda(\boldsymbol{x})}. \tag{4}$$

The idea behind this is that there is a systematic component, $\lambda(\boldsymbol{x})$, which drives the expected number of individuals according to different environmental conditions, $\boldsymbol{x}$. When this is coupled with a sampling model, such as simple random sampling or stratified sampling, the distribution of the sampled species abundance M_1 is marginalised over the true abundance, m_0, and becomes another Poisson, $\mathsf{Po}(m_1; \alpha\lambda(\boldsymbol{x}))$:

$$f(m_1|\boldsymbol{x}) = \frac{(\alpha\lambda(\boldsymbol{x}))^{m_1}}{m_1!} e^{-\alpha\lambda(\boldsymbol{x})}, \tag{5}$$

where $0 < \alpha \leq 1$ is the individual detection probability, or sampling fraction. For simple random sampling, $\alpha = r$ so that $f(m_1|\boldsymbol{x}) = \mathsf{Po}(m_1; r\lambda(\boldsymbol{x}))$. For stratified sampling, α depends on each stratum j so it becomes $\alpha_j = p_j r_j$ as

$$f(\boldsymbol{m}_1|\boldsymbol{x}) = \prod_{j=1}^{k} \mathsf{Po}(m_{1j}; p_j r_j \lambda(\boldsymbol{x})). \tag{6}$$

See Appendix for the detailed derivations. Clearly the effect of sampling with these sampling schemes is to reduce the expected abundance by a constant amount.

In general, conventional ecological modelling can be regarded as a mix of design-based and model-based approaches. From a design-based aspect, as we have discussed, it leads to a general model (Eq. 5) that plays a key role in species abundance and presence/absence modelling as we will see in the following Sect. 3.4. The other aspect, model-based one, can also be vital because the assumption that the individual detection probability, α, is fixed by design may sound unreasonable for some cases. It assumes the detection probability as an unknown function of other variable t as $\alpha(t)$ which needs to be estimated, for example by maximum likelihood. The component of estimation is thus model-based. Although the formulation allows more flexibility to cope with heterogeneity induced by different types of sampling, such as observer error and species rarity, for example, it requires an extra care, since with the Poisson model the individual detection probability, α, cannot be dis-entangled with the abundance expectation from the data alone. Sprott (1965) studied the condition of the probability generating function and identified this model as being inestimable, amongst other compound distributions. Further information about the sampling mechanism or the population's rate are required.

As noted, the probability mass function $f(m_0|\boldsymbol{x})$ in Eq. (3) can take any plausible form, and a negative binomial distribution could also be used. When a negative binomial distribution $\mathsf{NB}(m_0; s, t)$ is, instead of a Poisson distribution, coupled with a binomial sampling distribution, we still obtain an equivalent result. The marginal distribution is a compound negative binomial distribution and its form is explicitly written as $\mathsf{NB}(m_1; s, t/\{1-(1-\alpha)(1-t)\})$ (see Appendix for the detailed derivation).

There is a close link between Eq. (5) and a model class, namely N-mixture models, by Royle (2004). When the sampling replication is one, Eq. (5) gives the exact analytical expression of the N-mixture model, although Royle (2004) suggested a numerical approximation, calculating instead a finite summation over m_0 up to a reasonably large number. We note that a recent study (Dennis et al. 2015) has pointed out that the choice of the arbitrary large number in the numerical calculation can result in underestimation of abundance.

3.4 Modelling biological responses

We now discuss how to incorporate the sampling effect into modelling. Incorporation can be easily done using existing software that fit common models, such as generalised linear models (GLMs) (McCullagh and Nelder 1989) and their many extensions, including generalised additive models (GAMs) (Hastie and Tibshirani 1990).

Species abundance The sampled abundance m_1 has the expected value, from Eq. (5), as

$$\mu = \mathrm{E}[M_1] = \alpha\lambda(\boldsymbol{x}).$$

This form suggests that the effect of sampling can be treated as an offset using the log link function,

$$\log(\mu) = \log(\alpha) + \eta(\boldsymbol{x}),$$

where $\eta(\boldsymbol{x})$ is the (non-)linear predictor of environment variables. So, to convert a Poisson model for the sampled abundance into a model for the population abundance, all one has to do is to include $\log(\alpha)$ as an offset.

Using an offset term with the log link function in modelling is also commonplace for normalisation, calculating an expected abundance *per unit* of sampling space such as area, time duration and the length of a transect. This appears to be a parallel to the modelling approach above, as the size of sampling space often varies among surveys. However, we stress here that $0 < \alpha \leq 1$ is the individual detection probability, whereas the offset term for normalisation is not related with imperfect observation due to sampling.

Note the fact that the model formulation presented for biological responses is still valid for the model-based approach, dealing with the detection probability as an unknown function $\alpha(t)$, but the model fitting algorithm required will no longer be simple as adopting `glm` or `gam` provided in R, unless $\log(\alpha(t))$ forms a linear function. Also note that it is inestimable when the expected abundance and the sampling probability are confounded, such as situations when there are no repeat visits to a survey site, since these cannot be disentagled as we have noted in Sect. 3.3.

Species presence/absence From Eq. (2), the distribution of sampled presence/absence, Y_1, is the binarisation of a Poisson random variable:

$$f(y_1|\boldsymbol{x}) = \left(1 - e^{-\alpha\lambda(\boldsymbol{x})}\right)^{y_1} \left(e^{-\alpha\lambda(\boldsymbol{x})}\right)^{(1-y_1)}. \tag{7}$$

It has the expected value

$$\mu = \mathrm{E}[Y_1] = 1 - e^{-\alpha\lambda(\boldsymbol{x})}.$$

and is equal to the probability of the Poisson random variable M_1 to take non-zero values. In the GLM and GAM contexts, a Bernoulli model for the presence/absence variable is also easily implemented, using the complementary log-log link function, viz

$$\log(-\log(1 - \mu)) = \log(\alpha) + \eta(\boldsymbol{x}). \tag{8}$$

The sampling effect α, again is an offset term. The early idea of the complementary log-log link can be found in Fisher (1922) for a dilution assay study and it is more formally stated by McCullagh and Nelder (1989).

We note here a link to a modelling framework widely used for species presence/absence, the occupancy model (MacKenzie et al. 2002, 2006), that deals with species, not individual, detectability. In fact, the occupancy model can be interpreted as an approximate model of Eq. (8) when the species has low abundance, low probability of occupancy in other words. The occupancy model, equivalent to Eq. (7), is expressed as a zero-inflated Bernoulli model as

$$f(y_1|\boldsymbol{x}) = \psi p^{y_1}(1-p)^{1-y_1} + (1-\psi)I(y_1 = 0), \tag{9}$$

where $0 < \psi \leq 1$ is the species occupancy (true presence) probability and $0 < p \leq 1$ is the species detection probability, each of which is modelled in the logit form, such as $\text{logit}(\psi) = \xi(\boldsymbol{x})$. Since its expected value is $\mu = \text{E}[Y_1] = \psi p$, it follows

$$\log(\alpha) + \eta(\boldsymbol{x}) = \log(-\log(1-\mu)) \approx \log(p) + \xi(\boldsymbol{x}) - \psi.$$

This suggests that for a species with low probability of occupancy, with a small ψ, the occupancy model approximates Eq. (8), utilising species detectability. Note that we here used the facts that $\log(-\log(1-b)) \approx \log(b)$ and $\text{logit}(b) \approx \log(b) + b$ for a small b.

3.5 The effect of binarisation

As we have shown, species' presence/absence data can be treated as the binarisation of species abundance (Eq. 7), which allows us to deal with sampling effects. This also implies that species' presence/absence model is able to predict the species abundance, by gaining the information from the estimated intensity function $\hat{\lambda}(\boldsymbol{x})$, as suggested by Royle and Nichols (2003) for example. Although this is appealing, the cost is wider standard errors for the estimates. This is due to the loss of information by binarisation. Let $\hat{\lambda}(M_1; \alpha)$ and $\hat{\lambda}(Y_1; \alpha)$ be the likelihood estimators based on the species abundance and presence/absence data, respectively. As the variance of the estimators are given by the Fisher information, noting Eqs. (5) and (7), the efficiency of $\hat{\lambda}(Y_1; \alpha)$ becomes

$$\text{eff}(\hat{\lambda}(Y_1; \alpha)) = \frac{\text{Var}[\hat{\lambda}(M_1; \alpha)]}{\text{Var}[\hat{\lambda}(Y_1; \alpha)]} = \frac{\alpha\hat{\lambda}}{e^{\alpha\hat{\lambda}} - 1}, \quad (0 < \alpha \leq 1).$$

This suggests that the variance of the parameter estimated from species presence/absence data increases in the exponential order according to the mean abundance, λ. For even moderate abundances, this is likely to be substantial—enough to suggest that in many situations modelling abundance from presence-absence data is a risky practice. This is formal verification of an intuitively unsurprising result. We also note Howard et al. (2014) as a recent study reporting an empirical evidence of this issue.

3.6 Predictions of biological responses

The estimated intensity function $\hat{\lambda}(\boldsymbol{x})$ reveals how the biological responses are related to the environment factors $\boldsymbol{x}$. It allows us to make predictions of unobserved biological responses: species population abundance, M_0, and presence/absence, Y_0. We describe here two types of predictions. The first is *marginal prediction* which is the unconditional expectation of the observation ($\mathrm{E}[M_0]$ or $\mathrm{E}[Y_0]$) and is directly derived from the distribution $f(m_0; \hat{\lambda}(\boldsymbol{x}))$, Eq. (4). The other is *conditional prediction* ($\mathrm{E}[M_0|M_1 = m_1]$ or $\mathrm{E}[Y_0|Y_1 = y_1]$) calculated from the distribution, $f(m_0|m_1; \hat{\lambda}(\boldsymbol{x}))$. Note that $\hat{\lambda}(\boldsymbol{x})$ is used as a plug-in estimate. In a Bayesian analysis, one would incorporate uncertainty in this estimate into the predictive distribution. Due to the dependence on sample data, the conditional distribution (and predictions) are only available at previously sampled locations. However, at other locations one could define the conditional prediction to coincide with the marginal prediction—note though that there is no extra data to inform the process.

The conditional predictive distribution is derived by using Bayes' theorem and the distributions $f(m_1|m_0)$, $f(m_0|\boldsymbol{x})$ and $f(m_1|\boldsymbol{x}) = \sum_{m_0} f(m_1|m_0)f(m_0|\boldsymbol{x})$. The first two terms are already specified in the earlier sections. The conditional distribution is given as

$$\begin{aligned} f(m_0|m_1) &= \frac{f(m_1|m_0)f(m_0)}{f(m_1|\boldsymbol{x})} \\ &= \frac{\{(1-\alpha)\lambda(\boldsymbol{x})\}^{m_0-m_1}}{(m_0-m_1)!} e^{-(1-\alpha)\lambda(\boldsymbol{x})}. \end{aligned} \tag{10}$$

Note that the distribution (10) suggests that the difference $(m_0 - m_1)$, which is the amount of abundance that should have been observed but degenerated by sampling, also follows the Poisson distribution with the rate parameter $(1-\alpha)\lambda(\boldsymbol{x})$. A closed form for the conditional expectation is not generally available but it is here due to the Poisson assumption for $f(m_1|\boldsymbol{x})$. If other assumptions are made then it is likely that the conditional predictive distribution will have to be calculated numerically.

Species abundance The predictors are derived respectively from Eq. (4) and (10) as

$$\begin{aligned} \mathrm{E}[M_0] &= \lambda(\boldsymbol{x}), \quad \text{and} \\ \mathrm{E}[M_0|M_1 = m_1] &= \sum_{m_0=0}^{\infty} m_0 f(m_0|m_1) \\ &= (1-\alpha)\lambda(\boldsymbol{x}) + m_1 \\ &= \mathrm{E}[M_0] - \mathrm{E}[M_1] + m_1. \end{aligned}$$

This is the observed abundance plus the difference between expected true and expected observed abundance. In essence, it takes the observation and adjusts it for what is likely to be missed through sampling.

Presence/absence Viewing presence/absence data as the binarisation of a Poisson variable, the predictors are

$$
\begin{aligned}
\mathrm{E}[Y_0] &= 1 - e^{-\lambda(\boldsymbol{x})}, \quad \text{and} \\
\mathrm{E}[Y_0|Y_1 = y_1] &= \sum_{y_0=0}^{1} y_0 f(y_0|y_1) \\
&= \sum_{y_0=0}^{1} y_0 \left\{ \left(1 - e^{-(1-\alpha)\lambda(\boldsymbol{x})}\right)^{y_0} \left(e^{-(1-\alpha)\lambda(\boldsymbol{x})}\right)^{(1-y_0)} I(y_1 = 0) \right. \\
&\qquad \left. + y_0 I(y_1 = 1) \right\} \\
&= \left(1 - e^{-(1-\alpha)\lambda(\boldsymbol{x})}\right) I(y_1 = 0) + I(y_1 = 1) \\
&= \frac{\mathrm{E}[Y_0] - \mathrm{E}[Y_1]}{1 - \mathrm{E}[Y_1]} I(y_1 = 0) + I(y_1 = 1).
\end{aligned}
$$

The conditional expectation, $\mathrm{E}[Y_0|Y_1 = y_1]$, is 1 if the species is observed and a non-zero probability if the species is not observed. The non-zero probability reflects the difference in expectation between the true and the observed presence/absence record.

4 Data analysis

We analyse two ecological data sets from the marine realm. It is common, but not ubiquitous, in marine surveys to sub-sample the biological content at a survey location as the volume of biological material can be large. Sub-sampling (Heales et al. 2000) is performed to reduce the processing time and storage requirements for the biological material. The sub-sampling process sometimes divides the full ecological sample into strata, for example taxa groups or size classes, and then takes a proportion from each stratum. Often the sampling proportion changes between strata. We use these data sets to highlight the ideas and methods introduced in the previous sections. One data set exhibits simple random sampling and the other is generated from stratified sampling (Fig. 1).

4.1 Carnarvon Shelf data

4.1.1 Data and sampling method

Data were collected in a seabed mapping survey of the Carnarvon Shelf offshore (Fig. 1b) from central Western Australia (Brooke et al. 2009). The aim of the survey was to acquire physical and biological data to enable a range of environmental parameters to be tested as surrogates of benthic biodiversity patterns. A Smith-McIntyre grab was deployed at 142 sites. For each grab, a sediment sample (∼50 ml) was retained for analysis of textural characteristics while the remaining sediments were processed for infauna. The infauna samples were separated by washing sediments through a 500 μm

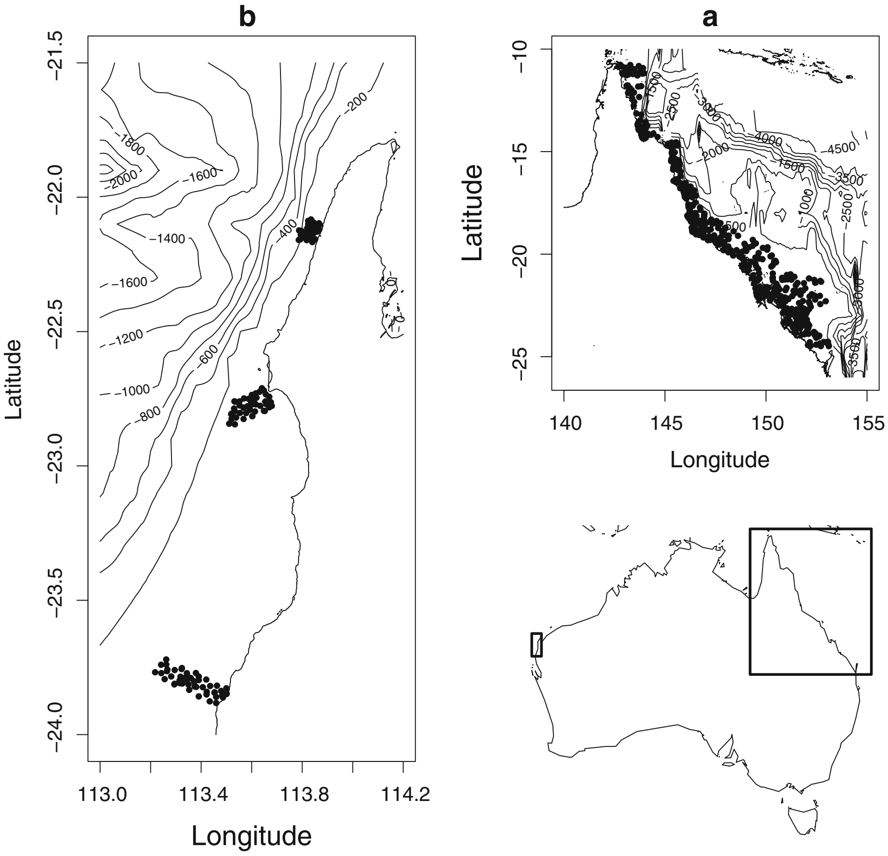

Fig. 1 Survey locations. **a** Great Barrier Reef; **b** Carnarvon Shelf

sieve and then a sample was taken if necessary. The proportion sampled was recorded as the sampling ratio r. The samples were classified into food guild groups and species.

As an example, for illustration, we present the results for two food-guild groups: non-selective and selective feeders. Both guilds ingest sediment and derive nutrients from the microorganisms living on the particles but selective feeders often have a physical structure which enables them to select optimally-sized particles for ingestion (e.g. tentacle with a ciliated groove). Such a physical difference may let them have different preference in their ambient sediment conditions. In our modelling, we assume that a species' preference in sediment conditions is common over all survey sites. That is, there is no interaction between preference and spatial location.

4.1.2 Modelling

We develop a model to describe how the presence/absence of each food guild group responds to the seabed grain size. As each species obviously has different abundances, we fit a model to each species separately and then combine the models for one food guild group.

Let Y_{1ks} be the presence/absence variable of the k-th species being 1 for presence or 0 for absence in the sample at the s-th site; Y_{1ks} is therefore drawn from the population of the interest Y_{0ks}, the presence/absence of the k-th species in the grab at the s-th site. The subscript 1 indicates that it is a sample and follows from Sect. 2. From Eq. (7) and using the complementary log-log link, $\mu_{ks} = \mathrm{E}[Y_{1ks}]$ can be modelled as

$$\log(-\log(1 - \mu_{ks})) = \log(r_s) + \beta_0 + \beta_1 x_s,$$

where x_s is the log-scaled seabed grain size. Here r_s is the sampling fraction given at site s.

The probability of presence of a group is calculated as multiplication of the probabilities of presence of the group's constituent species. Given the two food guild groups, $\mathcal{G}_j$, $j = 1, 2$ say, the probability of presence of the j-th species-group is

$$\mu_{js} = 1 - \prod_{k \in \mathcal{G}_j} (1 - \mu_{ks}), \tag{11}$$

where $\mu_{ks} = \mathrm{E}[Y_{1ks}] = \Pr(Y_{1ks} = 1)$ which is the probability of the k-th species presence.

4.1.3 Result

Different inferences are obtained when the sub-sampling effect is taken into account or ignored (equivalent to assuming that $r_s = 1$). Figure 2 illustrates the probability of presence of each food guild group (Eq. 11). The model that accounts for the sub-sampling effect (the left panel, Fig. 2) suggests that the probability of presence of the non-selective group species decreases as seabed sediments become coarse but the selective group species have little influence of the sediment size. In contrast, the model that ignores the effect (the right panel, Fig. 2) shows that both groups respond to the sediment size and the probability of presence decreases as the sediment size increases. These contradictory results highlight the risk of misinterpretation when the sub-sampling effect is mis-specified in the model. To us, it seems more plausible that the selective feeders should not have much dependence on grain size as this is their evolutionary advantage.

4.2 Great Barrier Reef data

4.2.1 Data and sampling method

Data were collected by trawl nets from 437 sites during a survey of the Great Barrier Reef (GBR) lagoon (Fig. 1a) off the north eastern coast of Australia (Pitcher et al. 2007). The purpose of data collection was to map habitats, assemblages and species throughout the Great Barrier Reef Marine Park.

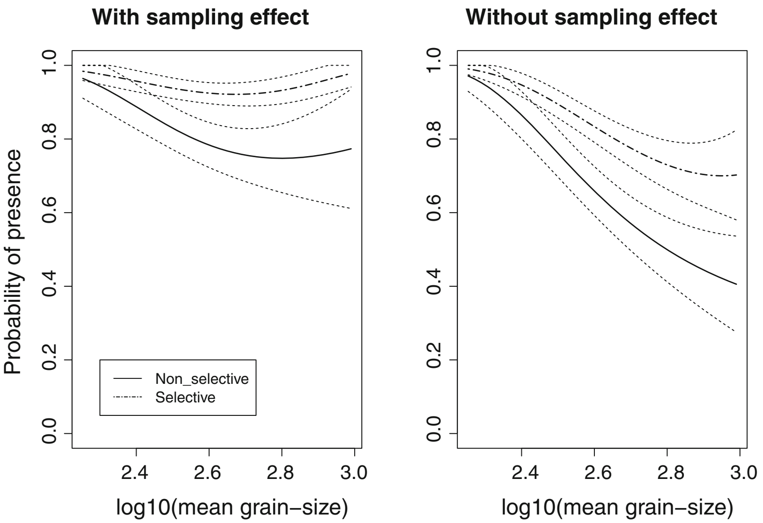

Fig. 2 Comparison of the probabilities of each food guild group presence: when the sub-sampling effect is taken into account (*left*); when the sub-sampling effect is ignored (*right*). The *solid* and *two-dashed lines* respectively represents the result of non-selective (*solid*) and selective (*two-dashed*) species. The *dashed lines* represent the upper and lower bounds of the 95 % confidence interval

The biological samples were collected by a scientific trawl net towed behind a survey vessel. After each tow, the samples were processed entirely or sub-sampled for enumeration, weighing and identification. On the deck, the samples were sorted into rough phylogenetic groups (strata $\mathcal{U}_j$, $j = 1, 2, \ldots, k$) and then a sub-sample was taken from each stratum (group) if necessary. The proportion of sub-sample was recorded as sub-sampling ratio r_j for the j-th stratum. On board taxonomic stratification is a difficult task, and some mis-specification is inevitable—some species were observed in an unexpected stratum some times. The taxonomic sorting suggested some heterogeneity ($p_j \neq 1/k$) was induced, and its mis-specification meant $p_j \neq 1$ for the j-th stratum, $\mathcal{U}_j$, given an organism belonging to the j-th group (stratum). This required extra consideration for p_j, $j = 1, 2, \ldots, k$, when modelling. We consider here four environment variables: depth, % carbonates, % gravel and % sand. Note that none of these percentages sum to 100 %.

4.2.2 Modelling

Let M_{1js} be observed abundance of a species of the j-th stratum at the s-th site, which is a sample drawn from the population of the interest M_{0js}, the species abundance in the j-th stratum actually caught at the s-th site. From Eq. (6) with a log link, $\mu_{js} = \mathrm{E}[M_{1js}]$ can be modelled as

$$\log(\mu_{js}) = \log(r_{js}) + \log(p_{js}) + \eta_0 + \eta(\boldsymbol{x}_s),$$

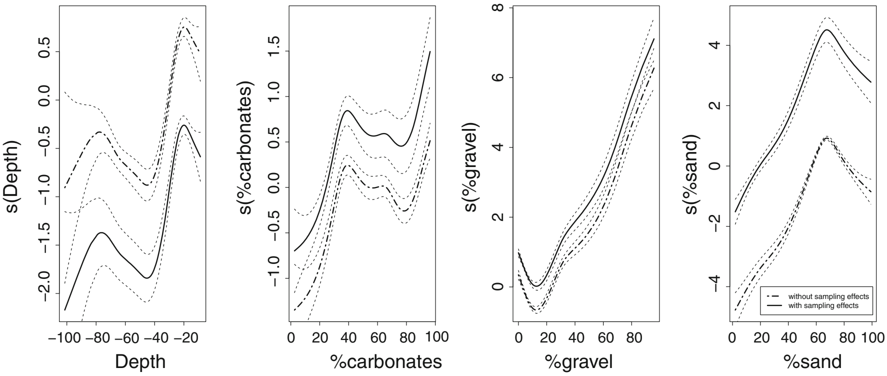

Fig. 3 The relationship of the abundance and each environment variable. *Each panel* shows the estimated linear predictor. The *solid* and *two-dashed lines* respectively represents the result of the model with (*solid*) or without (*two-dashed*) the sub-sampling effect. The *dashed lines* represent the upper and lower bounds of the 95 % confidence interval

where $\boldsymbol{x}_s$ is the vector of the environment variables: depth, % carbonates, % gravel and % sand, and $\eta(\boldsymbol{x}_s)$ is non-linear and smooth function. This model is a generalised additive model (GAM) (Wood 2006). Here r_{js} and p_{js} are respectively the sampling fraction induced by the subsampling and the taxonomic sorting. However, as noted before, p_{js} is not observed so it needs to be estimated as a parameter. We therefore assume that p_j's are common over the sites and fit the model

$$\log(\mu_{js}) = \log(r_{js}) + \xi_j + \eta(\boldsymbol{x}_s), \tag{12}$$

where $\xi_j = \log(p_j) + \eta_0$. Note that ξ_j represents the probability of classification *and* the species intercept.

4.2.3 Result

The model is fitted to the abundance data of a squid species (*Photololligo chinensis*). Each panel in Fig. 3 represents the response of its abundance to the environment variable. The solid and two-dashed line respectively represents the predictor, $\eta(\boldsymbol{x}_s)$, of the model with or without the sub-sampling effect taken into account. The sub-sampling effect is now easily observed as a constant shift in the linear predictor for all covariates, except depth. This does not translate to a constant difference on the response scale though. The response of the model that ignores the sub-sampling effect underestimates the abundance. It also shows that the confidence interval of the model that accounts for the effect tends to be wider than the one ignoring the effect. These observations concur with the theoretical results in Sect. 5.

The two types of the species abundance predictions are illustrated in Fig. 4. The marginal predictions are more variable than the conditional predictions and deviate further from the observations. The conditional predictions are always greater than the observations that they are predicting. This is a consequence of the form for the

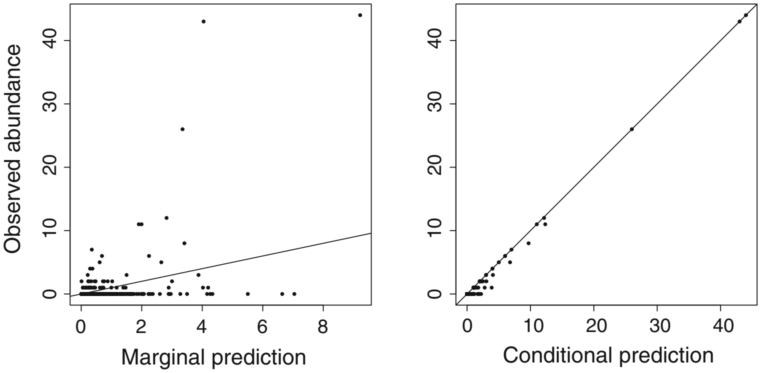

Fig. 4 The marginal (*left*) and conditional (*right*) predictions against the observed abundance of *Photololigo chinensis*. When the predictions match the observations, all points lie on the *solid line*. Departure from this *line* signifies prediction inaccuracy

conditional prediction, which shows that the predictive distribution is the observation with an adjustment term added on.

4.2.4 A technical matter

This stratified sampling used in the GBR data is a two step sampling technique. However, biological data are provided in a form that all biological quantities are aggregated at each site regardless of the stratification that is actually employed. For instance, the stratified sampling is applied, like the GBR data, then the aggregation over the strata, $\sum_j M_{1j}$, may provide insufficient information unless the sampling ratio r_j's are common over the strata ($r_j = r_{j'}, j \neq j'$), which becomes exactly the case of simple random sampling. This can be seen as

$$\mu_s = \sum_j \mathrm{E}[M_{1j}] = \sum_j \mu_{js} = e^{\eta_0 + \eta(\boldsymbol{x}_s)} \sum_j r_{js} p_{js}.$$

The aggregation induces substantial difficulty in dealing with the sampling effect properly since the sampling effect of each stratum has now been aggregated, and becomes unknown. An alternative approach for this kind of already-aggregated data may be random effect models that assumes sampling effects to vary randomly over the sites.

5 Implications for survey design

Up until now, we have discussed a modelling framework that accounts for imperfect observations. Here, we view the framework from a slightly different perspective—designing surveys. Consider the Poisson model for imperfect observations (Eq. 5); the parameter of interest is λ, the expected abundance of a species at a site, and the other notation remains the same as before. Yet, the sampling fraction is unspecified at the design phase so that the sampling fraction α needs to be chosen. A common question here may be: "what is a sufficient number of sites to survey whilst retaining a predetermined precision for estimates of the models parameters?" The variation is a function of the number of observations, n, and the sampling fraction, α; the variance of likelihood estimators is given by the Fisher information,

$$\begin{aligned}\mathrm{Var}[\hat{\lambda}\,(\boldsymbol{M}_1,\boldsymbol{\alpha})] &= -\mathrm{E}_{\boldsymbol{M}_1,\boldsymbol{\alpha}}\left[\frac{\partial u\,(\lambda_0;\boldsymbol{M}_1,\boldsymbol{\alpha})}{\partial\lambda}\right]^{-1}\\ &= -n^{-1}\mathrm{E}_{M_1,\alpha}\left[\frac{\partial u\,(\lambda_0;M_1,\alpha)}{\partial\lambda}\right]^{-1},\end{aligned}\tag{13}$$

for the observable abundance M_{1s} and the sampling fraction α_s at a site $s, s = 1, 2, \ldots, n$. Note that the vector notation here is about the sites so that $\boldsymbol{M}_1 = (M_{11}, M_{12}, \ldots, M_{1n})^\top$. The function $u(\cdot)$ is score function, the first derivative of the log-likelihood function, and λ_0 is the true value of the parameter. The estimated variance can be calculated, substituting λ_0 by the maximum likelihood estimate $\hat{\lambda}$. Clearly, the variance decreases as the number of sites n increases (Eq. 13). However, the expectation term is also a function of the sampling fraction, α. The rate of decrease with n is slowed from $O(n^{-1})$ unless $\mathrm{E}[\alpha] = 1$ (note that $0 < \alpha \leq 1$). This leads us to the idea that the accuracy of the estimator can be designed, taking the balance between the number of observations n and sampling fraction α.

From Eqs (5) and (13), the variance of the maximum likelihood estimator $\hat{\lambda}(\boldsymbol{M}_1, \boldsymbol{\alpha})$ is given as

$$\mathrm{Var}[\hat{\lambda}\,(\boldsymbol{M}_1,\boldsymbol{\alpha})] = \frac{\hat{\lambda}}{n\mathrm{E}[\alpha]}$$

where $0 < \mathrm{E}[\alpha] \leq 1$. This suggests a wider standard error for the same value of $\hat{\lambda}$ when a sample is taken. If it is fully sampled, then $\mathrm{E}[\alpha] = 1$ and $\mathrm{Var}[\hat{\lambda}(\boldsymbol{M}_1,\boldsymbol{\alpha})] = \hat{\lambda}/n$. Let z be the number of survey sites required to obtain the same standard error range such that $\hat{\lambda}/(z\mathrm{E}[\alpha]) = \hat{\lambda}/n$ then

$$z = \frac{n}{\mathrm{E}[\alpha]}.$$

This means that, for example, if a survey is undertaken to be $\mathrm{E}[\alpha] = 0.5$ then it requires two times more observations, $2n$, for keeping the same error range as no sub-sampling is taken. This gives some concrete guidance as to whether to choose between sampling fewer sites with great detail (α near 1), or to sample more sites with less detail (small α). The actual choice will depend on the relative cost of performing more sites or taking a less imperfect sample at each site. Consideration should also be given to the range of covariates that competing sampling strategies will cover. If stratification of sampling locations is to occur, based on a covariate, then it is likely to be that more sites will aid the estimation of that covariate's effect.

6 Summary

We have discussed how the imperfect observation effect due to sampling should be treated in ecological modelling, and presented how a general framework, the compound distribution, can accommodate individual detectability. The model is general and can handle many different types of sampling, including the two examples used in this paper: the commonly used simple random sampling and stratified random sampling. The method of implementing the sampling effect is straight-forward; the sampling effect enters the regression-type model as an offset term by using an appropriate link function. Other types of sampling mechanisms, such as cluster sampling, will require slightly more complex models that allow for the between individual correlation.

Our examples are typical of a sampling technique called ‘sub-sampling’, which is widely used in marine surveys. This is completely an anthropogenic effect induced during the survey process that should be taken into account for modelling. To the authors’ knowledge, anthropogenic sub-sampling is under studied; only two of articles can be found in the literature (Heales et al. 2000, 2003). Another sampling effect in marine surveys is the issue of catchability (also called detectability), which describes how likely the individuals will get caught given the sampling gear employed. We have not discussed this as our data consists of a single observation at each site, so the probability of catching an individual, given presence, is completely confounded with the probability of presence. If a site was visited multiple times then this could be incorporated into the compound distribution framework, and the catchability could be estimated. Commonly, this has been done using *species* occupancy models (MacKenzie et al. 2002). These types of models are an approximation to our framework, see Eq. (9).

Fisher (1934) clearly emphasises the importance of understanding the data collection procedure employed as a statistical commonplace. Accordingly, Rao (1965) generalises Fisher’s idea and proposes a general modelling framework that is able to accommodate a wide class of sampling mechanisms, such as non-observability of events by dealing with individual detectability. The compound model presented (Sect. 2 and Eq. 3) exhibits strong similarities with one of the models described in Rao (1965) and also in Patil and Rao (1978). Ecological studies will always have a limited number of observations from the population of interest, and so the statistical challenge has historically been centred around how to make effective inferences dealing properly with the sampling effect. This challenge will remain into the future. The compound distribution seems to have received little attention in ecological modelling and we show that it naturally underpins an effective modelling framework to account for imperfect observations.

Acknowledgments Many thanks go to Carolyn Huston and Scott Nichol for their constructive comments and suggestions on an early version of this manuscript. HS acknowledges the support by the European Research Council (Project BioTIME 250189). SDF was supported by the Marine Biodiversity Hub, a collaborative partnership supported through funding from the Australian Government’s National Environmental Research Program (NERP). NERP Marine Biodiversity Hub partners include the Institute for Marine and Antarctic Studies, University of Tasmania; CSIRO Wealth from Oceans National Flagship, Geoscience

Australia, Australian Institute of Marine Science, Museum Victoria, Charles Darwin University and the University of Western Australia.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Appendix: Some derivations

Equation (5): Marginal distribution for Poisson population size and simple random sampling

$$\begin{aligned}
f(m_1) &= \sum_{m_0=m_1}^{\infty} \mathsf{Bi}(m_1; m_0, \alpha) \times \mathsf{Po}(m_0; \lambda(\boldsymbol{x})) \\
&= \sum_{m_0=m_1}^{\infty} \frac{m_0!}{m_1!(m_0-m_1)!}\alpha^{m_1}(1-\alpha)^{m_0-m_1}\frac{\lambda^{m_0}}{m_0!}e^{-\lambda} \\
&= \frac{1}{m_1!}e^{-\lambda}\left(\frac{\alpha}{1-\alpha}\right)^{m_1}\sum_{m_0=m_1}^{\infty}\frac{\{\lambda(1-\alpha)\}^{m_0}}{(m_0-m_1)!} \\
&= \frac{(\alpha\lambda)^{m_1}}{m_1!}e^{-\lambda}\sum_{z=0}^{\infty}\frac{\{\lambda(1-\alpha)\}^{z}}{z!}, \qquad (z = m_0 - m_1) \\
&= \frac{(\alpha\lambda)^{m_1}}{m_1!}e^{-\alpha\lambda}
\end{aligned}$$

Equation (6): Marginal distribution for Poisson population size and stratified random sampling

$$\begin{aligned}
f(m_1) &= \sum_{m_0=0}^{\infty}\sum_{\mathcal{M}_0}\prod_{j=1}^{k} \mathsf{Bi}(m_{1j}; m_{0j}, r_j) \times \mathsf{Mn}(\boldsymbol{m}_0; m_0, \boldsymbol{p}) \times \mathsf{Po}(m_0; \lambda(\boldsymbol{x})) \\
&= \sum_{m_0=0}^{\infty}\sum_{\mathcal{M}_0}\prod_{j=1}^{k} \frac{m_{0j}!}{m_{1j}!(m_{0j}-m_{1j})!}r_j^{m_{1j}}(1-r_j)^{m_{0j}-m_{1j}} \\
&\quad \times \frac{m_0!}{\prod_{j=1}^{k} m_{0j}!}\prod_{j=1}^{k} p_j^{m_{0j}}\frac{\lambda^{m_0}}{m_0!}e^{-\lambda} \\
&= e^{-\lambda}\prod_{j=1}^{k}\frac{1}{m_{1j}!}\left(\frac{r_j}{1-r_j}\right)^{m_{1j}}\sum_{m_0=0}^{\infty}\sum_{\mathcal{M}_0}\prod_{j=1}^{k}\frac{\{\lambda p_j(1-r_j)\}^{m_{0j}}}{(m_{0j}-m_{1j})!}
\end{aligned}$$

$$
\begin{aligned}
&= e^{-\lambda}\prod_{j=1}^{k}\frac{1}{m_{1j}!}\left(\frac{r_j}{1-r_j}\right)^{m_{1j}}\prod_{j=1}^{k-1}\sum_{m_{0j}=0}^{\infty}\frac{\left\{\lambda p_j(1-r_j)\right\}^{m_{0j}}}{(m_{0j}-m_{1j})!}\\
&\quad\times\sum_{m_0=0}^{\infty}\frac{\left\{\lambda p_j(1-r_k)\right\}^{m_0-m_{0,-k}}}{(m_0-m_{0,-k}-m_{1k})!}\\
&= e^{-\lambda}\prod_{j=1}^{k}\frac{(r_jp_j\lambda)^{m_{1j}}}{m_{1j}!}e^{(1-r_j)p_j\lambda}, \qquad \left(\sum_{j=1}^{k}p_j=1\right)\\
&= \prod_{j=1}^{k}\frac{(r_jp_j\lambda)^{m_{1j}}}{m_{1j}!}e^{-r_jp_j\lambda}
\end{aligned}
$$

where $\boldsymbol{\mathcal{M}}_0 = \left\{\boldsymbol{m}_0 : \sum_{j=1}^{k} m_{0j} = m_0\right\}$ and $m_{0,-k} = \sum_{j\neq k} m_{0j}$.
Compound negative binomial distribution for simple random sampling

$$
\begin{aligned}
f(m_1) &= \sum_{m_0=m_1}^{\infty}\mathsf{Bi}(m_1; m_0, \alpha)\times\mathsf{NB}(m_0; s, t)\\
&= \sum_{m_0=m_1}^{\infty}\binom{m_0}{m_1}\alpha^{m_1}(1-\alpha)^{m_0-m_1}\times\binom{-s}{m_0}t^s\{-(1-t)\}^{m_0}\\
&= \sum_{m_0=m_1}^{\infty}\frac{m_0!}{m_1!(m_0-m_1)!}\alpha^{m_1}\beta^{m_0-m_1}\frac{(-s)_{m_0}}{m_0!}t^s(-u)^{m_0},\\
&\qquad(\beta = 1-\alpha;\ u = 1-t)\\
&= \frac{t^s\alpha^{m_1}}{m_1!}\sum_{m_0=m_1}^{\infty}\frac{(-s)_{m_1}(-s-m_1)_{m_0}}{(m_0-m_1)!}\beta^{m_0-m_1}(-u)^{m_0}\\
&= \frac{(-s)_{m_1}}{m_1!}\frac{t^s(-u\alpha)^{m_1}}{(u\beta)^{s+m_1}}\sum_{z=0}^{\infty}\frac{(-s-m_1)_z}{z!}(u\beta)^{s+m_1}(-u\beta)^z,\ (z = m_0-m_1)\\
&= \binom{-s}{m_1}\frac{t^s(-u\alpha)^{m_1}}{(u\beta)^{s+m_1}}(u\beta)^{s+m_1}(1-u\beta)^{-(s+m_1)}\\
&= \binom{-s}{m_1}\left(\frac{t}{1-u\beta}\right)^s\left\{-\left(1-\frac{t}{1-u\beta}\right)\right\}^{m_1},
\end{aligned}
$$

where $(d)_k$, $0\le k\le d$ represents the descending factorial moment $d(d-1)\cdots(d-k+1)$.

References

Bax NJ (ed) (2011) Marine Biodiversity Hub, Commonwealth Environment Research Facilities. Final report 2007–2010. Report to Department of Sustainability, Environment, Water, Population and Communities. Canberra, Australia

Borchers DL, Buckland ST, Zucchini W (2002) Estimating animal abundance. Springer, London

Brooke B, Nichol S, Hughes M, McArthur M, Anderson T, Przeslawski R, Siwabessy J, Heyward A, Battershill C, Colquhoun J, Doherty P (2009) Carnarvon Shelf survey pos-survey report. Record 2009/02, Geoscience Australia

Buckland ST, Anderson DR, Burnham KP, Laake JL, Borchers DL, Thomas L (eds) (2004) Advanced distance sampling. Oxford University Press, Oxford

Cochran WG (1977) Sampling techniques, 3rd edn. Wiley, New York

Dennis EB, Morgan BJ, Ridout MS (2015) Computational aspects of n-mixture models. Biometrics 71:237–246

Dunn PK, Smyth GK (2005) Series evaluation of Tweedie exponential dispersion model densities. Stat Comput 15(4):267–280

Elith J, Leathwick JR (2009) Species distribution models: ecological explanation and prediction across space and time. Annu Rev Ecol Evol Syst 40:677–697

Feller W (1968) An introduction to probability theory and its applications, 3rd edn. Wiley, New York

Ferrier S, Guisan A (2006) Spatial modelling of biodiversity at the community level. J Appl Ecol 43:393–404

Fisher R (1922) On the mathematical foundations of theoretical statistics. Philos Trans R Soc Lond Ser A 222:309–368

Fisher RA (1934) The effect of methods of ascertainment upon the estimation of frequencies. Ann Eugen 6:13–25

Fisher RA, Corbet AS, Williams CB (1943) The relation between the number of species and the number of individuals in a random sample of animal population. J Anim Ecol 12(1):42–58

Foster SD, Bravington MV (2013) A poisson-gamma model for analysis of ecological non-negative continuous data. Environ Ecol Stat 20:533–552

Foster SD, Shimadzu H, Darnell R (2012) Uncertainty in spatially predicted covariates: Is it ignorable? J R Stat Soc Ser C 61(4):637–652

Franklin J (2010) Mapping species distributions. Cambridge University Press, Cambridge

Gattone SA, Battista TD (2009) A functional approach to diversity profiles. J R Stat Soc Ser C 58(2):267–284

Guisan A, Zimmermann NE (2000) Predictive habitat distribution models in ecology. Ecol Model 135(2–3):147–186

Gurland J (1957) Some interrelations among compound and generalized distributions. Biometrika 44:265–268

Hastie TJ, Tibshirani RJ (1990) Generalized additive models. Chapman & Hall, Florida

Heales DS, Brewer DT, Wang YG (2000) Subsampling multi-species trawl catches from tropical northern Australia: Does it matter which part of the catch is sampled? Fish Res 48:117–126

Heales DS, Brewer DT, Wang YG, Jones PN (2003) Does the size of subsamples taken from multispecies trawl catches affect estimates of catch composition and abundance? Fish Bull 101:790–799

Heck KLJ, van Belle G, Simberloff D (1975) Explicit calculation of the rarefaction diversity measurement and the determination of sufficient sample size. Ecology 56(6):1459–1461

Howard C, Stephens PA, Pearce-Higgins JW, Gregory RD, Willis SG (2014) Improving species distribution models: the value of data on abundance. Methods Ecol Evol 5(6):506–513

Hurlbert SH (1971) The nonconcept of species diversity: a critique and alternative parameters. Ecology 52(4):577–586

Jørgenson B (1997) The theory of dispersion models. Chapman and Hall, London

Johnson NL, Kotz S, Kemp AW (1992) Univariate discrete distributions, 2nd edn. Wiley, New Jersey

Lozier JD, Aniello P, Hickerson MJ (2009) Predicting the distribution of Sasquatch in western North America: anything goes with ecological niche modelling. J Biogeogr 36:1623–1627

MacKenzie DI, Nichols JD, Lachman GB, Droege S, Royle JA, Langtimm CA (2002) Estimating site occupancy rates when detection probabilities are less than one. Ecology 83(8):2248–2255

MacKenzie DI, Nichols JD, Royle JA, Pollock KH, Bailey LL, Hines JE (2006) Occupancy estimation and modeling. Academic Press, Cambridge

McCullagh P, Nelder J (1989) Generalized linear models, 2nd edn. Chapman and Hall, Florida

Patil G, Rao C (1978) Weighted distributions and size-biased sampling with applications to wildlife populations and human families. Biometrics 34:179–189

Pitcher C, Doherty P, Arnold P, Hooper J, Gribble N, Bartlett C, Browne M, Campbell N, Cannard T, Cappo M, Carini G, Chalmers S, Cheers S, Chetwynd D, Colefax A, Coles R, Cook S, Davie P, De'ath G, Devereux D, Done B, Donovan T, Ehrke B, Ellis N, Ericson G, Fellegara I, Forcey K, Furey M, Gledhill D, Good N, Gordon S, Haywood M, Jacobsen I, Johnson J, Jones M, Kinninmoth S, Kistle

S, Last P, Leite A, Marks S, McLeod I, Oczkowicz S, Rose C, Seabright D, Sheils J, Sherlock M, Skelton P, Smith D, Smith G, Speare P, Stowar M, Strickland C, Sutcliffe P, Van der Geest C, Venables W, Walsh C, Wassenberg T, Welna A, Yearsley G (2007) Seabed biodiversity on the continental shelf of the Great Barrier Reef world heritage area. Technical report of CSIRO marine and atmospheric research

Rao C (1965) On discrete distributions arising out of methods of ascertainment. Sankhya 27:311–324

Royle JA (2004) N-mixture models for estimating population size from spatially replicated counts. Biometrics 60:108–115

Royle JA, Nichols JD (2003) Estimating abundance from repeated presence-absence data or point counts. Ecology 84(3):777–790

Sanders HL (1968) Marine benthic diversity: a comparative study. Am Nat 102(925):243–282

Shimadzu H, Darnell R (2015) Attenuation of species abundance distributions by sampling. R Soc Open Sci 2(140):219. doi:10.1098/rsos.140219

Simberloff D (1972) Properties of the rarefaction diversity measurement. Am Nat 106(949):414–418

Sprott D (1965) Some comments on the question of identifiability of parameters raised by Rao. Sankhya Indian J Stat Ser A 27(2/4):365–368

Stoklosa J, Daly C, Foster SD, Ashcroft MB, Warton DI (2015) A climate of uncertainty: accounting for error in climate variables for species distribution models. Methods Ecol Evol 6(4):412–423

Wood S (2006) Generalized additive models: an introduction with R. Chapman and Hall, Florida

Hideyasu Shimadzu is a Lecturer in Statistics with the Department of Mathematical Sciences, Loughborough University and a visiting scholar at the Centre for Biological Diversity and Scottish Oceans Institute of the University of St Andrews, UK. His general research interests are in the science of data, Data Science—how to extract useful information from data for answering scientific questions—with particular emphasis on the intersection of statistics and environmental/ecological sciences.

Scott D. Foster is a statistician with the Commonwealth Science and Industrial Research Organisation and is located in Hobart, Australia. His interests lay in statistical methods for solving problems arising from marine ecology. This encompasses data analysis and survey design issues stemming from fisheries, community ecology and biodiversity research.

Ross Darnell is a scientist with the Commonwealth Science and Industrial Research Organisation based in Brisbane, Australia. His focus is on survey and experimental design and statistical modelling of aquatic and agricultural systems.