ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Applied Statistics
Series C

Appl. Statist. (2017)

Fully general Chao and Zelterman estimators with application to a whale shark population

Alessio Farcomeni

Sapienza—University of Rome, Italy

[Received May 2016. Final revision January 2017]

Summary. We introduce generalized Chao and generalized Zelterman estimators which include individual, time varying and behavioural effects. Under mild assumptions in the presence of unobserved heterogeneity, the generalized Chao estimator asymptotically provides a lower bound for the population size and is unbiased otherwise. Corrected versions guarantee bounded estimates. To include the best set of predictors we propose the biased empirical focused information criterion bFIC. Simulations indicate that bFIC might give considerable improvements over other selection criteria in our context. We illustrate with an original application to size estimation of a whale shark (*Rhincodon typus*) population in South Ari Atoll, in the Maldives.

Keywords: Capture–recapture; Focused information criterion; Heterogeneity; Lower bound estimator; Whale sharks

1. Introduction

Estimating the size of a hidden or elusive population is a primary concern in a wide range of problems in ecology, agriculture, veterinary science, public health, medical studies, software engineering and behavioural research. See for instance Pollock (2000), Chao (2001) and McCrea and Morgan (2014) for detailed reviews of rationale and methods. In these experiments, each subject might be repeatedly observed. The counting process of sightings is then modelled, to obtain an estimate of the number of subjects that are never observed, or equivalently of the size of the catchable population. Time homogeneity and absence of behavioural response to capture are often assumed.

The Chao and Zelterman estimators (Chao, 1987, 1989; Zelterman, 1988) are very popular and have very simple expressions. Zelterman's estimator is popular in socio-economical applications and is based on the assumption that singletons and doubletons follow a homogeneous Poisson distribution (whereas other counts might be arbitrarily distributed). Chao's estimator is popular in enviromental applications and provides a sensible lower bound for the population size, taking into account unspecified unobserved heterogeneity. It is based on the assumption that counts are obtained from a mixture of Poisson distributions, with unknown mixing distribution.

Unobserved heterogeneity is a challenging problem in population size estimation. Parametric assumptions are untestable because of the presence of an unobserved fraction of the population. Non-parametric assumptions often lead to non-identifiable or at least inconsistent estimates (Link, 2003; Holzmann *et al*., 2006; Mao, 2008; Farcomeni and Tardella, 2012).

The goal of this work is to demonstrate that the mean-squared error (MSE) of Chao and

Address for correspondence: Alessio Farcomeni, Department of Public Health and Infectious Diseases, Sapienza—University of Rome, Piazzale Aldo Moro 5, 00186 Rome, Italy.
E-mail: alessio.farcomeni@uniroma1.it

 0035–9254/17/67000

Zelterman estimators can be reduced, possibly substantially, if we explicitly model some relevant sources of heterogeneity. The Zelterman and Chao estimators have already been extended to take into account subject-specific covariates by Böhning and van der Heijden (2009) and Böhning *et al*. (2013). In this paper we further generalize to include also time varying covariates and certain forms of behavioural effects.

The generalization will be achieved by exploiting a logistic parameterization of conditional detection probabilities, and the Poisson approximating properties of the sum of Bernoulli trials for large numbers of trials and small success probabilities. A simple correction guarantees bounded estimators. The behavioural effect is modelled, as in Farcomeni (2011, 2016), through detection probabilities conditioned on the previous observation history. The resulting estimators are simple and very fast to obtain. Some more efforts are needed for computing the standard errors.

An additional problem is given by the fact that, as testified also in the simulations in Böhning *et al*. (2013), the use of covariates does not necessarily decrease the MSE. The reason is that covariates decrease the bias but increase the standard error. We propose to choose, among all possible models, the model minimizing a biased empirical focused information criterion (FIC) bFIC, which targets the minimal MSE. For background on FICs see Claeskens and Hjort (2003). Our proposal compares well with other information criteria, and provides substantial MSE reduction especially when the true population size is large.

Our work is motivated by an original data set regarding a survey of whale sharks (*Rhincodon typus*) of South Air Atoll, in the Maldives. The area surrounding the atoll was explored every day by boat, in search of whale sharks. A total of $n = 112$ individuals were repeatedly identified and measured by laser photogrammetry (Rohner *et al*., 2011) whenever possible. Generalized Chao (GC) estimators (Böhning *et al.*, 2013) cannot be used because occasion-specific covariates are available, such as cloud cover during the exploration. We shall compare all possible GC and generalized Zelterman (GZ) estimators through the biased empirical FIC.

The remainder of the paper is organized as follows. In the next section we give background on Chao and Zelterman estimators. In Section 3 we describe our GC and GZ estimators incorporating subject-specific covariates, time-specific covariates and behavioural effects. In Section 3.2 we describe how to estimate standard errors, and our model selection strategy. Simulations are reported in Section 4. We describe and analyse the motivating data in Section 5 and state conclusions in Section 6.

The programs that were used to analyse the data can be obtained from

`http://wileyonlinelibrary.com/journal/rss-datasets`

2. Background on Chao and Zelterman estimators

Let Y_{ij} be a binary indicator of detection of the ith individual at the jth sampling occasion, for $j = 1, \ldots, S$. We observe the n individuals for which $\Sigma_j Y_{ij} > 0$, with a total population size of $N \geqslant n$. Let $p_{il} = \Pr(\Sigma_j Y_{ij} = l)$, and n_l denote the number of subjects observed l times, $l \geqslant 0$. We assume also that $\Sigma_j Y_{ij}$ follows a Poisson distribution with parameter λ_i. This assumption in some cases can be seen only as an approximation to the (exact) binomial distribution for $\Sigma_j Y_{ij}$. This will be more precisely formalized below. Assume further that λ_i is a random effect that is sampled from $F(\cdot)$, where $F(\cdot)$ is a mixing distribution summarizing unobserved heterogeneity. Leave unspecified $F(\lambda)$, and denote

$$p_l = \int_0^\infty \frac{\exp(-\lambda)\lambda^l}{l!} \mathrm{d}F(\lambda).$$

Use of the Cauchy–Schwarz inequality yields $p_1^2 \leqslant 2p_0p_2$. Consequently, at least asymptotically, $n_0 \geqslant n_1^2/(2n_2)$ and $N \geqslant n + n_1^2/(2n_2)$. Chao's estimate is defined as $n + n_1^2/(2n_2)$, which is guaranteed to provide a lower bound for N as n increases.

The Zelterman estimator is obtained instead by assuming a homogeneous Poisson distribution with parameter λ for n_1 and n_2, and ignoring the other frequencies. This assumption and the properties of the Poisson distribution lead to $\hat{\lambda} = 2n_2/n_1$. The Horvitz–Thompson approach yields the population size estimator $n/\{1 - \exp(-\hat{\lambda})\}$.

Böhning and van der Heijden (2009) and Böhning *et al.* (2013) noted that both estimators arise from a truncated likelihood restricted to subjects observed at most twice. To see this, let z_i be a binary indicator of the ith subject being observed twice. After some algebra with Poisson probabilities, the truncated likelihood can be written as

$$\sum_{i=1}^{n_1+n_2} \left(\frac{2}{2+\lambda_i}\right)^{1-z_i} \left(\frac{\lambda_i}{2+\lambda_i}\right)^{z_i}. \tag{1}$$

A log-link can be used to specify a model on λ_i, thereby including subject-specific covariates. The Zelterman estimator is recovered as the maximum likelihood solution, whereas Chao's estimator is recovered from the predicted value for n_0 (e.g. Böhning *et al.* (2013), theorem 1). More formally, it can be seen that the Böhning *et al.* (2013) GC estimator with subject-specific covariates corresponds to $n + \Sigma_{i=1}^{n_1+n_2} 2/(2\hat{\lambda}_i + \hat{\lambda}_i^2)$, and the Böhning and van der Heijden (2009) GZ estimator to $\Sigma_{i=1}^{n} \{1 - \exp(-\hat{\lambda}_i)\}^{-1}$, where $\hat{\lambda}_i$ is the MLE of likelihood (1).

3. Fully general Chao and Zelterman estimators

Let $p_{ij} = \Pr(Y_{ij} = 1)$, and $\mathbf{X}_{ij}$ denote a time varying subject-specific vector of covariates which are always observed. Here $\mathbf{X}_{ij}$ includes time-fixed subject-specific covariates (e.g. gender), occasion-specific covariates (e.g. weather and sea conditions, time trends and occasion dummies) and interactions between them.

We specify the following logistic regression model (see also, for example, Coull and Agresti (1999, 2000) and Farcomeni (2016)):

$$\log\left(\frac{p_{ij}}{1 - p_{ij}}\right) = \mathbf{X}_{ij}'\boldsymbol{\beta}. \tag{2}$$

To recover Chao- and Zelterman-type estimators, we base inference on a truncated likelihood, restricting to subjects that are observed once or twice. Given that we do not consider subjects that are never detected, there are no missing covariates (Huggins, 1989; Alho, 1990).

It has long been known (Hodges and Le Cam, 1960; Le Cam, 1960) that the distribution of a sum of S inhomogeneous Bernoulli trials, each with success probability p_{ij}, $j = 1, \ldots, S$, is well approximated by a Poisson distribution with parameter $\Sigma_j p_{ij}$ as soon as S is large and $\max_j p_{ij}$ is small.

This allows us to write a convenient approximated truncated likelihood as

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n_1+n_2} \frac{\Pr(Y_{i1}, \ldots, Y_{iS})}{\Pr\left(1 \leqslant \sum_{j=1}^{S} Y_{ij} \leqslant 2\right)} \\ &= \prod_{i=1}^{n_1+n_2} \frac{\prod_{j=1}^{S} p_{ij}^{Y_{ij}} (1 - p_{ij})^{1-Y_{ij}}}{\exp\left(-\sum_{j=1}^{S} p_{ij}\right)\left\{\sum_{j=1}^{S} p_{ij} + 0.5\left(\sum_{j=1}^{S} p_{ij}\right)^2\right\}}, \end{aligned} \tag{3}$$

where p_{ij} is a function of model parameters as in equation (2). The Poisson approximation theorems have been used to approximate $\Pr(1 \leqslant \Sigma_{j=1}^{S} Y_{ij} \leqslant 2)$ in the denominator of equation (3). To maximize equation (3), we obtain the score in closed form (reported in Appendix A) and the Hessian through a numerical first derivative of the score. We then set up Newton–Raphson iterations.

Let $\hat{p}_{ij}$ denote the estimated parameters. The approximating properties just mentioned and invariance of the MLE allow us to obtain a GZ estimator directly as

$$\sum_{i=1}^{n}\left\{1-\exp\left(-\sum_{j=1}^{S}\hat{p}_{ij}\right)\right\}^{-1}. \tag{4}$$

Chao's lower bound estimator is seen to correspond to

$$n+\sum_{i=1}^{n_1+n_2}\frac{1}{\sum_{j=1}^{S}\hat{p}_{ij}+0.5\left(\sum_{j=1}^{S}\hat{p}_{ij}\right)^2}. \tag{5}$$

To see this, we rely on theorem 1 in Böhning *et al.* (2013) where it is shown that the Chao estimator can be represented as

$$n+\sum_{i=1}^{n_1+n_2}\frac{\Pr\left(\sum_j Y_{ij}=0\right)}{\Pr\left(\sum_j Y_{ij}=1\right)+\Pr\left(\sum_j Y_{ij}=2\right)}. \tag{6}$$

Correspondingly, because of the Poisson approximation we have that

$$\Pr\left(\sum_j Y_{ij}=x\right)=\left(\sum_{j=1}^{S}\hat{p}_{ij}\right)^{x}\exp\left(-\sum_{j=1}^{S}\hat{p}_{ij}\right)\Big/x!.$$

Substituting the last expression into expression (6) gives expression (5).

There are various intuitive reasons why expressions (5) and (4) can be referred to as GC and GZ estimators. First, they reduce to Chao and Zelterman estimators when no covariates are used. Secondly, they share the same properties and are obtained through the same rationale (Horvitz–Thompson estimators based on the MLE for the Zelterman estimator and $n+E[n_0|n_1,n_2]$ for the Chao estimator). Finally expression (5) provides a lower bound estimate for N, and in the absence of unobserved heterogeneity it becomes unbiased (as implied by the properties of the MLE).

A bias-corrected version of the Chao estimator is given by $n+n_1(n_1-1)/\{2(n_2+1)\}$ (Chao, 1989; Wilson and Collins, 1992). This

(a) avoids unbounded estimates and
(b) has a lower bias than the original Chao estimator.

When n_1 and n_2 are small, bias-corrected estimation is generally recommended. A simplified version of the bias-corrected estimator is given by $n+n_1^2/\{2(n_2+1)\}$. See Böhning (2010) for more details on this point. We here generalize the simplified bias-corrected estimator, which is guaranteed to avoid unbounded estimates, by adding to the data set a fictitious sample point with two captures, and whose covariates are set at the columnwise means of the real sample points. We proceed similarly to avoid unboundedness of the GZ estimator.

3.1. Behavioural effects

We have dealt so far with subject-specific and occasion-specific effects. A further generalization is given by the possibility of including behavioural effects. The classical behavioural model is based on two different capture probabilities depending on whether the animal is at its first sight or has been previously observed. This effect is simply modelled by letting p_{ij} depend on the previous capture history as follows:

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \mathbf{X}'_{ij}\boldsymbol{\beta} + g_j(Y_{i,j-1}, \ldots, Y_{i1}), \tag{7}$$

with

$$g_j(Y_{i,j-1}, \ldots, Y_{i1}) = \eta\, I\left(\sum_{l=1}^{j-1} Y_{il} > 0\right), \tag{8}$$

where $I(\cdot)$ is the indicator function. We adopt the convention that $g_1(\cdot) = 0$, as there are no previous captures. It was shown in Farcomeni (2011, 2016) that all possible behavioural effects are obtained with different specifications of $g_j(\cdot)$. Regardless of $g_j(\cdot)$, parameters can be estimated as before through similar Newton–Raphson routines.

A problem with behavioural effects is that capture events Y_{ij} for each subject i are not independent anymore. Dependence of Y_{ij} on Y_{il}, $l < j$, in general makes the Poisson approximation that was used in the previous section invalid, meaning that we cannot specify any function $g_j(\cdot)$ and still obtain an approximation of the likelihood. For instance, if $g_j(\cdot)$ diverges negatively as soon as $\Sigma_{l=1}^{j-1} Y_{il} > 0$, Chao's estimator (5) is asymptotically unbounded even if N is finite. Our task is then to define classes of functions $g_j(\cdot)$, and consequently classes of behavioural models, leading to binary indicators which can still be well approximated by a Poisson random variable with parameter $\lambda_i = \Sigma_j E[Y_{ij}]$.

First, if $\{Y_{ij}, j > 0\}$ can be shown to be ϕ mixing under the model assumed, then the Poisson approximation is still valid as shown by Chen (1975). The idea of ϕ mixing is quite simple and involves the fact that the dependence between Y_{ij} and $Y_{i,j+k}$ decreases sufficiently fast with k, and the two indicators are eventually independent when $k \to \infty$. A simple class of ϕ-mixing processes is given by Harris recurrent Markov chains. Hence we can fix $m > 0$ (with $m \ll S$) and define $g_j(Y_{i,j-1}, \ldots, Y_{i1}) = \Sigma_{l=1}^{m} \eta_l Y_{i,j-l}$. These correspond to a transient behavioural response, where the capture probability is changed by a capture event within the m previous occasions. This class of models satisfies the mixing assumptions above, leading to satisfactory Poisson approximation of the likelihood. First-order Markov chains ($m = 1$) were introduced in Yang and Chao (2005) and generalized to m-order chains in Farcomeni (2011).

Secondly, if $\{Y_{ij}, j > 0\}$ is either associated or negatively associated, the Poisson approximation is once again valid as shown in Boutsikas and Koutras (2000). Two binary random variables are (negatively) associated as soon as

$$\Pr(Y_{ij} = 1, Y_{ij'} = 1) \geqslant (\leqslant) \Pr(Y_{ij} = 1) \Pr(Y_{ij'} = 1) \tag{9}$$

for any $j \neq j'$. See Esary *et al.* (1967) for a general discussion on association. This result can be used to show that also the classical 'permanent memory' behavioural model (8) leads to valid inference with expressions (5) and (4). To see this, note that positive association is equivalent to

$$\Pr(Y_{ij} = 1 | Y_{ij'} = 1) - \Pr(Y_{ij} = 1 | Y_{ij'} = 0) \geqslant 0. \tag{10}$$

Assume without loss of generality that $j > j'$. Suppose that $\eta \geqslant 0$. We have that under equation (8)

$$\log\left\{\frac{\Pr(Y_{ij}=1|Y_{ij'}=1)}{\Pr(Y_{ij}=0|Y_{ij'}=1)}\right\}=\beta_{0j}+\mathbf{X}'_{ij}\boldsymbol{\beta}+\eta$$

and

$$\log\left\{\frac{\Pr(Y_{ij}=1|Y_{ij'}=0)}{\Pr(Y_{ij}=0|Y_{ij'}=0)}\right\}\leqslant\beta_{0j}+\mathbf{X}'_{ij}\boldsymbol{\beta}+\eta,$$

as it is less likely that $\Sigma_{l<j}Y_{il}>0$. Hence, inequality (10) holds. A similar reasoning can be used to show negative association when $\eta<0$. The same results can also be shown to hold for the class of 'delayed onset' models (Farcomeni and Scacciatelli, 2013), i.e. for functions of the kind $g_j(Y_{i,j-1},\ldots,Y_{i1})=\eta\, I(\Sigma_l Y_{il}>k)$ for any fixed $k>0$. In delayed onset models the behavioural response occurs only after k repeated sightings.

In summary, short memory Markovian models and long memory delayed onset models (including the classical behavioural response model) can be used in model (7) in association with expression (5) or (4). In general the user can specify any $g_j(\cdot)$, but this needs to be checked for either mixing or (negative) association.

We conclude noting that under model (7) $E[Y_{ij}]\neq p_{ij}$ in general. The parameter of interest $\lambda_i=\Sigma_j E[Y_{ij}]$ is nevertheless a function of p_{ij}, involving a simple marginalization over the previous detection probabilities.

3.2. Standard errors and model selection

Standard errors for the population size estimators must be obtained taking into account the fact that the sample is biased and some subjects were never observed, and the fact that the likelihood is truncated (Böhning, 2008). First, note that population size estimators are expressed as $\hat{N}=h(\boldsymbol{\theta})=\Sigma_{i=1}^{N}\Delta_i h_i(\boldsymbol{\theta})$, where $\Delta_i=1$ if the ith subject is observed once or twice (and $\Delta_i=0$ otherwise), $\boldsymbol{\theta}$ is a shorthand notation for parameters involved in model (7), and $h(\boldsymbol{\theta})$ is the mapping from $\boldsymbol{\theta}$ to $\hat{N}$. By conditioning,

$$\operatorname{var}(\hat{N})=\operatorname{var}(E[\hat{N}|\Delta_i])+E[\operatorname{var}(\hat{N}|\Delta_i)]. \tag{11}$$

For the GC estimator, after some algebra it can be seen that the first term is unbiasedly estimated by

$$\sum_{i=1}^{n_1+n_2}(1-\hat{p}_i)\left\{1+\exp\left(-\sum_{j=1}^{S}\hat{p}_{ij}\right)\Big/\hat{p}_i\right\}^2,$$

where $\hat{p}_i=\exp(-\Sigma_{j=1}^{S}\hat{p}_{ij})(1+\Sigma_{j=1}^{S}\hat{p}_{ij}/2)\Sigma_{j=1}^{S}\hat{p}_{ij}$. For the GZ estimator, after some algebra we obtain

$$\sum_{i=1}^{n_1+n_2}\exp\left(-\sum_{j=1}^{S}\hat{p}_{ij}\right)\Big/\left\{1-\exp\left(-\sum_{j=1}^{S}\hat{p}_{ij}\right)\right\}^2$$

The second term in equation (11) can be approximated as

$$\nabla h(\hat{\theta})\, J(\hat{\theta})^{-1}\nabla h(\hat{\theta}),$$

where $J(\boldsymbol{\theta})$ is the Fisher information matrix obtained as minus the numerical derivative of the score, and evaluated at the parameter estimates. The gradient of $h(\boldsymbol{\theta})$ is readily obtained via numerical differentiation.

We now discuss model selection. As is intuitive, the (asymptotic) bias of the GC estimator is non-positive and is closer to 0 than the bias of the original Chao estimator. In contrast, the

variance of the estimate is often larger because of the uncertainty that is linked with the use of additional parameters to model covariate effects. The trade-off between bias and variance is often in favour of estimators with more covariates, but not necessarily. It is even possible that the MSE (which is the sum of squared bias and variance) of Chao's estimator is smaller than that of any GC estimator. More in general, if p covariates are measured (including for instance time dummies) and q possible behavioural effects are considered, there are $2^p q$ possible models. It is likely that the model minimizing the MSE is not the full or the empty model. A careful tuning of the covariates that are included is therefore very important. To date there are no direct approaches for performing this kind of calibration.

We here propose an empirical FIC which can be used to select the estimator with least (asymptotically) expected MSE for the population size estimator. See Bartolucci and Lupparelli (2008) for an example of use of the FIC in the context of population size estimation. Unlike other information criteria the FIC focuses on the parameter of primary interest. In our context, this is N. The FIC is defined as the expected MSE for the parameter of interest under a model. In our context, we propose to use the model with lowest expected MSE (and, hence, FIC) for N. The expected MSE is often obtained analytically in other contexts, and data are then used to estimate the expected MSE. Unfortunately, for GC and GZ estimators an expression for the asymptotic MSE is not readily available. Additionally, an unbiased MSE estimator is also not readily available.

We propose a *biased* empirical MSE estimator as follows: let $\hat{N}_1, \ldots, \hat{N}_v$ denote a collection of candidate population size estimates (e.g. all of $2^p q$ possible models, or a subset of them) and $\widehat{\text{se}}_1, \ldots, \widehat{\text{se}}_v$ their estimated standard errors. We propose the following information criterion:

$$\text{bFIC}_j = \left(\hat{N}_j - \max_j \hat{N}_j\right)^2 + \widehat{\text{se}}_j^2, \tag{12}$$

for $j = 1, \ldots, v$. Note that bFIC is a biased estimator of the MSE of $\hat{N}$, as the first addend is a biased estimator of bias. The model minimizing bFIC is selected. The idea behind bFIC is that, given that all candidate estimators are negatively biased, the largest is the least biased. Hence the closer the estimate to $\max_j \hat{N}_j$, the less biased it is. Consequently, the first addend in equation (12) is a biased estimate of bias. The key idea is that the rank of $\hat{N}_j$ with respect to bias is (at least asymptotically) the same as we would obtain if we replaced $\max_j \hat{N}_j$ with an unbiased estimate for N. The second addend is the variance of the estimate. Consequently, by definition bFIC is a biased estimator of MSE which ranks $N_1, \ldots, N_v$ in agreement with an ideal, unavailable, unbiased estimate of MSE.

In our simulation study below we shall compare bFIC with the Akaike information criterion AIC, in terms of their ability to pick the model with the smallest MSE.

A related issue is goodness of fit of the model. We propose to proceed simply by computing a χ^2-statistic comparing observed and predicted counts, and performing the related test.

4. Simulations

In this section we run a brief simulation study to evaluate our proposed estimators. For brevity we focus on Chao and GC estimators only. We evaluate the bias and MSE of the Chao, oracle GC estimator (when the true model is known) and the GC estimator of a model chosen with AIC or bFIC. As mentioned before, Chao and GC estimators might break down. We therefore also evaluate the risk of breakdown of each estimator and the performance of bias-corrected Chao and bias-corrected GC estimators as described above.

Table 1. Bias, RMSE and probability of failure, p_f, for the Chao and GC estimators by using all covariates (oracle), GC with covariates chosen by using AIC, the GC estimator with covariates chosen by using bFIC and their bias-corrected versions†

N	S	*Chao estimator*			*Oracle GC estimator*			*AIC GC estimator*			*bFIC GC estimator*		
		Bias	*RMSE*	p_f	*Bias*	*RMSE*	p_f	*Bias*	*RMSE*	p_f	*Bias*	*RMSE*	p_f
2 covariates													
250	8	−136.0	136.9	0.05	−20.1	201.7	0.20	−167.8	166.8	0.13	−137.4	142.0	0.06
250	30	−91.3	87.6	0.01	−77.9	64.8	0.18	−79.8	65.1	0.24	−92.2	81.4	0.02
250	100	−50.5	49.1	0.01	−40.5	37.6	0.03	−41.2	38.1	0.03	−43.2	40.1	0.03
1000	8	−585.0	549.6	0.01	−586.2	587.6	0.21	−640.8	675.7	0.18	−567.1	550.8	0.03
1000	30	−399.9	396.4	0.01	−325.0	324.4	0.04	−325.0	324.5	0.04	−286.2	263.8	0.01
1000	100	−208.2	207.4	0.01	−164.3	163.0	0.02	−164.3	163.0	0.02	−151.3	146.9	0.01
		BC Chao estimator			*BC oracle GC estimator*			*AIC BC GC estimator*			*bFIC BC GC estimator*		
250	8	−148.4	147.2	0.00	−164.5	162.4	0.00	−177.3	172.5	0.00	−160.7	158.9	0.00
250	30	−98.4	97.0	0.00	−79.1	82.8	0.00	−80.4	82.8	0.00	−95.0	95.9	0.00
250	100	−50.5	50.1	0.00	−41.0	39.4	0.00	−41.3	39.6	0.00	−44.4	41.5	0.00
1000	8	−575.3	559.8	0.00	−609.8	601.6	0.00	−634.5	621.6	0.00	−583.2	560.1	0.00
1000	30	−402.8	396.7	0.00	−329.6	331.9	0.00	−329.6	331.9	0.00	−296.4	280.6	0.00
1000	100	−206.0	205.8	0.00	−167.1	164.7	0.00	−167.1	164.7	0.00	−153.2	146.9	0.00
		Chao estimator			*Oracle GC estimator*			*AIC GC estimator*			*bFIC GC estimator*		
4 covariates													
250	8	−149.6	147.6	0.03	−137.6	145.4	0.14	−162.9	163.1	0.08	−121.0	140.0	0.22
250	30	−89.2	84.3	0.01	−82.7	80.7	0.02	−83.9	81.7	0.04	−74.0	69.9	0.01
250	100	−43.5	41.7	0.01	−36.1	33.5	0.01	−38.0	35.2	0.01	−37.2	35.2	0.00
1000	8	−612.1	636.1	0.01	−566.1	586.2	0.11	−591.2	614.5	0.08	−427.2	390.9	0.14
1000	30	−356.5	342.2	0.00	−348.5	344.6	0.00	−348.5	344.6	0.00	−254.9	246.4	0.00
1000	100	−183.6	154.6	0.00	−16.1	106.8	0.00	−16.3	106.9	0.00	−144.6	94.6	0.00
		BC Chao estimator			*BC oracle GC estimator*			*AIC BC GC estimator*			*bFIC BC GC estimator*		
250	8	−159.0	157.5	0.00	−154.9	133.4	0.00	−169.0	191.4	0.00	−159.4	156.5	0.00
250	30	−87.4	80.3	0.00	−80.3	77.8	0.00	−81.2	79.4	0.00	−76.3	71.6	0.00
250	100	−44.6	43.7	0.00	−37.1	35.0	0.00	−39.7	37.1	0.00	−39.4	37.0	0.00
1000	8	−639.6	666.8	0.00	−613.1	613.1	0.00	−621.6	627.8	0.00	−561.2	393.5	0.00
1000	30	−346.1	331.1	0.00	−340.7	341.8	0.00	−340.5	341.5	0.00	−266.1	254.5	0.00
1000	100	−182.2	166.6	0.00	−160.8	107.1	0.00	−161.1	107.4	0.00	−142.9	93.2	0.00

†The results refer to two simulation settings (with two measured covariates and one not measured, and with four measured covariates and one not measured). Results are based on $B = 1000$ replicates.

We fix $N = \{250, 1000\}$, $S = \{8, 30, 100\}$ and two scenarios. In the first we generate three covariates: a fair binary subject-specific X_1, a standard Gaussian occasion-specific X_2 and a subject-specific standard Gaussian X_0 which is ignored by all models (therefore leading to unobserved heterogeneity and biased estimators). In the second scenario we generate five covariates: X_1 and X_2 as before, X_3 and X_4 as independent replicates of X_1 and X_2, and X_0 as above. For AIC and bFIC we consider all possible models, which number 4 in the first scenario and 16 in the second. We report the bias and square root of the MSE, RMSE, of each procedure, and the failure probabilities. We estimate the probability of failure by evaluating outlying MSE values, and report the reweighted minimum covariance determinant over the replicates for bias and MSE. The reweighted minimum covariance determinant is a robust measure of location which can adapt to differing numbers of outliers (see for example Farcomeni and Greco (2015)). Results are based on $B = 1000$ replicates for each setting, and are reported in Table 1.

It can be seen that the GC and Chao estimators are negatively biased as expected. The GC estimator is biased as it always ignores the relevant covariate X_0. Additionally, the Poisson approximation underlying it is sufficiently good even with $S = 8$, as GC results are satisfactory even in this case. The GC is almost always less biased than the Chao estimator. This happens at the price of a slightly larger standard error, and in some cases it can even lead the oracle GC estimator to have a larger MSE than does the Chao estimator. Both estimators have a tendency to break down, with of course a larger breakdown probability for the GC than for the Chao estimator. Failure is more likely when $S = 8$, as in this case the sampling fraction is very small. Bias-corrected versions of the Chao and GC estimators never fail, and in most cases result in an MSE which is comparable with that of the uncorrected estimators.

The best performance is almost always obtained with bFIC model selection. The improvement in terms of MSE with respect to the null, full (oracle) and AIC-selected models is in some cases substantial. The reduction in RMSE can be dramatic when $N = 1000$. Additionally, in the settings where the MSE of the bFIC GC estimator is not the lowest, it is still very close to the best. In many cases, bFIC selects an in-between model that is based on a subset of available covariates. These are difficult to identify otherwise (e.g. with tests for significance or other model selection strategies). We therefore recommend our bFIC-strategy for model selection. We also compared (but do not report the results for brevity) with the Bayesian and Takeuchi information criteria, obtaining analogous results.

5. Data description and analysis

Our motivating data set was collected by the Maldives Whale Shark Research Programme (`http://mwsrp.org`). Whale sharks are the largest fish in the world and are usually observed at aggregation locations. South Ari Atoll in the Maldives is one of these aggregation points. Several characteristics of this fascinating species are still unknown, including how they interact and where and when they breed. The subpopulation of the South Ari Atoll in the Maldives is also peculiar in that surfacing sharks are almost always jouveniles.

Every day for a period of 6 months the area surrounding South Ari Atoll was explored by the Maldives Whale Shark Research Programme's boat. The minimal length observed was 3 m, indicating birth well before the beginning of the study period. The mean was 5.73 ± 1.24 m. Additionally, whale sharks have no predators and have a lifespan of about 70 years, indicating a low likelihood of deaths within the study period. It should also be mentioned that estimates reported below are remarkably stable in a sensitivity analysis which removes the youngest sharks or adds fictitious elder sharks. At each encounter sharks were photographed, and their identity was later confirmed by matching the unique spots patterns through registration. The spots on the skin are unique to each individual. Length was assessed at each encounter, and a linear regression model was used to estimate the length at the beginning of the study period, which was used as a subject-specific covariate together with gender. Occasion-specific covariates included cloud cover (complete or no), sea state (calm or rough) and season (high or low). It should be mentioned that South Ari Atoll is one of the few spots in the world where whale sharks surface all year round. January, April, May and June are nevertheless considered as 'high season', as the likelihood of encounters is slightly larger than in other months. Data on sightings for $S = 181$ occasions are reported in Table 2.

In Table 3 we report Chao, Zelterman and several GC and GZ estimates. Note that, as they are based on the same likelihood, for the same model specification AIC for the GC and GZ estimators correspond. Of course, their bFIC might be substantially different (see below). In Table 3, GC_{o} denotes a GC estimator based on the subject-specific covariates gender and length,

Table 2. Counting distribution of sightings for $S = 181$ occasions for the whale shark data set

Number of captures	1	2	3	4	5	6	7	8	9	10	11	12	14	16	17	18	21	$\geqslant 26$
Number of sights	45	15	9	7	3	3	4	3	3	2	4	4	3	1	2	1	1	2

Table 3. Population size estimates, sampling fraction $n/\hat{N}$, standard errors SE for $\hat{N}$ and AIC for the whale shark data set

Estimator	$\hat{N}$	$n/\hat{N}$ (%)	*SE*	*AIC*
Chao	179	62	24.70	828.05
GC_o	195	57	44.14	830.13
GC_t	179	62	27.46	818.41
GC_b	204	55	39.30	804.56
$GC_{b:Markovian}$	179	62	27.50	829.01
GC_{ot}	187	60	38.43	819.53
GC_{ob}	219	51	60.18	802.88
GC_{tb}	204	55	39.17	794.20
GC_{otb}	166	67	24.18	803.79
Zelterman	230	49	48.74	828.05
GZ_o	241	46	61.58	830.13
GZ_t	228	49	48.73	818.41
GZ_b	342	33	97.99	804.56
$GZ_{b:Markovian}$	229	49	48.83	829.01
GZ_{ot}	231	48	54.45	819.53
GZ_{ob}	343	33	105.71	802.88
GZ_{tb}	342	33	95.77	794.20
GZ_{otb}	253	44	63.29	803.79

GC_t a GC estimator based on the occasion-specific covariates and GC_b and $GC_{b:Markovian}$ denote the use of a classical and Markovian behavioural effect respectively. These effects are then combined in GC_{tb}, GC_{ot} models and so on. For comparison we have tried fitting the model that was described in Huggins (1989), and is available in the `mra` R package, but unfortunately because of the very large number of occasions it was not possible to obtain any estimator. It should be noted that similar computational issues might arise with GC and GZ estimators only with a much larger number of occasions, also because the likelihood is restricted to singletons and doubletons.

Given that we consider two possible behavioural effects (classical and Markovian) there are 32 possible models. We do not show all of them, but we estimate AIC and bFIC for all of them. We do so separately for the GC and GZ estimators. The optimal model, among the 32 possible models, according to bFIC corresponds to a GC estimator including gender and length, cloud coverage and (persistent) behavioural effects. This model has $\hat{N} = 222$, standard error 61.84 and AIC 802.10. Even if the standard error is slightly larger, we deem the model that is chosen by bFIC to be more credible than the model that is selected with AIC. To evaluate the goodness of fit of this model, we compute the predicted cell counts with the reported parameter estimates, which correspond to $\hat{n}_1 = 46.12$ and $\hat{n}_2 = 13.88$. The resulting χ^2-test has p-value 0.731, indicating an acceptable fit.

We report parameter estimates for the chosen model in Table 4, together with standard errors and p-values computed on the basis of Wald test statistics.

Males tend to surface remarkably more often than females, and in fact only nine females were encountered. The extremely low number of females does not allow the model to discriminate

Table 4. Coefficient estimates, standard errors SE and *p*-value for the whale shark data set: model selected through bFIC

Predictor	β	*SE*	*p-value*
Intercept	−7.60	1.59	<0.001
Gender	−0.86	1.16	0.458
Length	0.52	0.26	0.050
Cloud	−0.41	0.05	<0.001
Behavioural effect	1.87	0.29	<0.001

between a low prevalence and a low capture probability, therefore making the parameter estimate not significant. A strong sex bias towards male sharks is common to many aggregation points around the world. The importance of the other covariates is easily explained: sharks are detected by spotting their shadow from the boat when they approach the surface. When the day is cloudy, the shadow of surfacing sharks is dim and therefore more difficult to detect. Additionally, for obvious reasons longer sharks are more easily detected than shorter sharks. The behavioural effect in contrast is explained by a persistence behaviour of the researchers, who tended to return to the same spots where a shark was observed.

6. Discussion

We have presented two new population size estimators, the GC and GZ estimator, which can take into account observed heterogeneity (i.e. subject-specific and occasion-specific covariates), time-specific effects and behavioural effects. We have also provided a simple bias correction to guarantee bounded estimates when the number of recaptures is small. The GC estimator is robust with respect to residual unobserved heterogeneity and provides a lower bound for the population size. The only assumption is that the mixing distribution is Poisson. The GZ estimator is based on possibly even milder assumptions, requiring that conditional counts are distributed according to a homogeneous Poisson distribution only for subjects observed at most twice.

We have provided a new information criterion, bFIC, to select the model which is more likely to minimize the MSE of the population size estimate. We have seen in simulations that bFIC can substantially improve over AIC in certain cases. We speculate that bFIC compares so well with AIC because of the strong bias–variance trade-off that is associated with model choice for GC and GZ estimators. The idea has nevertheless wider applicability, which could be explored in further work.

The estimates are obtained through Newton–Raphson iterations, where the Hessian is computed via a numerical first derivative of the score. The numerical first derivative is quite fast and very precise (whereas taking a numerical second derivative of the log-likelihood would not have been as accurate). We had to restrict to certain classes of behavioural effects: we did so to be able to use Poisson approximation results under dependence. Poisson approximation was needed to establish a link between binomial models and Poisson MLEs. A possibility for further work is investigation of a method for working with any possible behavioural response. Additionally, Poisson approximation properties can be used to generalize fully also the extended Chao estimator that was proposed in Lanumteang and Böhning (2011).

Finally, our motivating data set was based on $S = 181$ occasions. An open issue is how to fix S in advance for Chao, GC, Zelterman and GZ estimators, to balance study length and precision

of the estimates. For more classical models this was considered for instance in Alunni Fegatelli and Farcomeni (2016).

Acknowledgements

The author is grateful to the Maldives Whale Shark Research Programme for permission to use their database and for help in understanding the data collection mechanisms, and to two referees for kind suggestions.

Appendix A: Score of conditional likelihood

Let p_{ij} be specified according to model (7). The expression for the approximated truncated likelihood is as in equation (3). Consequently, the log-likelihood can be written as

$$l(\theta) = \sum_{i=1}^{n_1+n_2}\sum_{j=1}^{S} Y_{ij}\log(p_{ij}) + \sum_{j=1}^{S}(1-Y_{ij})\log(1-p_{ij}) + \sum_{j=1}^{S} p_{ij} - \log\left\{\sum_{j=1}^{S} p_{ij} + \frac{1}{2}\left(\sum_{j=1}^{S} p_{ij}\right)^2\right\}.$$

The score can then be obtained as

$$\begin{aligned}\frac{\partial l(\theta)}{\partial \beta_h} &= \sum_{i=1}^{n_1+n_2}\sum_{j=1}^{S} x_{ijh}(Y_{ij}-p_{ij}) + \sum_{j=1}^{S}\frac{p_{ij}x_{ijh}}{1+\exp\{\mathbf{X}'_{ij}\boldsymbol{\beta}+g_j(Y_{i,j-1},\ldots,Y_{i1})\}} \\ &\quad - \frac{1+\sum_j p_{ij}}{\sum_j p_{ij} + 0.5\left(\sum_j p_{ij}\right)^2}\sum_{j=1}^{S}\frac{p_{ij}x_{ijh}}{1+\exp\{\mathbf{X}'_{ij}\boldsymbol{\beta}+g_j(Y_{i,j-1},\ldots,Y_{i1})\}} \\ &= \sum_{i=1}^{n_1+n_2}\sum_{j=1}^{S} x_{ijh}(Y_{ij}-p_{ij}) \\ &\quad + \frac{0.5\left(\sum_j p_{ij}\right)^2 - 1}{\sum_j p_{ij} + 0.5\left(\sum_j p_{ij}\right)^2}\sum_{j=1}^{S}\frac{p_{ij}x_{ijh}}{1+\exp\{\mathbf{X}'_{ij}\boldsymbol{\beta}+g_j(Y_{i,j-1},\ldots,Y_{i1})\}}\end{aligned}$$

and similarly for any parameter involved in $g_j(\cdot)$.

References

Alho, J. M. (1990) Logistic regression in capture-recapture models. *Biometrics*, **46**, 623–635.
Alunni Fegatelli, D. and Farcomeni, A. (2016) On the design of closed recapture experiments. *Biometr. J.*, **58**, 1273–1294.
Bartolucci, F. and Lupparelli, M. (2008) Focused information criterion for capture-recapture models for closed populations. *Scand. J. Statist.*, **35**, 629–649.
Böhning, D. (2008) A simple variance formula for population size estimators by conditioning. *Statist. Methodol.*, **5**, 410–423.
Böhning, D. (2010) Some general comparative points on Chao's and Zelterman's estimators of the population size. *Scand. J. Statist.*, **37**, 221–236.
Böhning, D. and van der Heijden, P. G. M. (2009) A covariate adjustment for zero-truncated approaches to estimating the size of hidden and elusive populations. *Ann. Appl. Statist.*, **3**, 595–610.
Böhning, D., Vidal-Diez, A., Lerdsuwansri, R., Viwatwongkasem, C. and Arnold, M. (2013) A generalization of Chao's estimator for covariate information. *Biometrics*, **69**, 1033–1042.
Boutsikas, M. V. and Koutras, M. V. (2000) A bound for the distribution of the sum of discrete associated or negatively associated random variables. *Ann. Appl. Probab.*, **10**, 1137–1150.
Chao, A. (1987) Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, **43**, 783–791.

Chao, A. (1989) Estimating population size for sparse data in capture-recapture experiments. *Biometrics*, **45**, 427–438.
Chao, A. (2001) An overview of closed capture-recapture models. *J. Agric. Biol. Environ. Statist.*, **6**, 158–175.
Chen, L. H. Y. (1975) Poisson approximation for dependent trials. *Ann. Probab.*, **3**, 534–545.
Claeskens, G. and Hjort, N. (2003) The focused information criterion. *J. Am. Statist. Ass.*, **98**, 900–916.
Coull, B. A. and Agresti, A. (1999) The use of mixed logit models to reflect heterogeneity in capture-recapture studies. *Biometrics*, **55**, 294–301.
Coull, B. A. and Agresti, A. (2000) Random effects modeling of multiple binomial responses using the multivariate binomial logit-normal distribution. *Biometrics*, **56**, 73–80.
Esary, J. D., Proschan, F. and Walkup, D. W. (1967) Association of random variables, with applications. *Ann. Math. Statist.*, **38**, 1466–1474.
Farcomeni, A. (2011) Recapture models under equality constraints for the conditional capture probabilities. *Biometrika*, **98**, 237–242.
Farcomeni, A. (2016) A general class of recapture models based on the conditional capture probabilities. *Biometrics*, **72**, 116–124.
Farcomeni, A. and Greco, L. (2015) *Robust Methods for Data Reduction*. Boca Raton: Chapman and Hall–CRC.
Farcomeni, A. and Scacciatelli, D. (2013) Heterogeneity and behavioral response in continuous time capture-recapture, with application to street cannabis use in Italy. *Ann. Appl. Statist.*, **7**, 2293–2314.
Farcomeni, A. and Tardella, L. (2012) Identifiability and inferential issues in capture-recapture experiments with heterogeneous detection probabilities. *Electron. J. Statist.*, **6**, 2602–2626.
Hodges, J. L. and Le Cam, L. (1960) The Poisson approximation to the Poisson binomial distribution. *Ann. Math. Statist.*, **31**, 737–740.
Holzmann, H., Munk, A. and Zucchini, W. (2006) On identifiability in capture-recapture models. *Biometrics*, **62**, 934–936.
Huggins, R. (1989) On the statistical analysis of capture experiments. *Biometrika*, **76**, 133–140.
Lanumteang, K. and Böhning, D. (2011) An extension of Chao's estimator of population size based on the first three capture frequency counts. *Computnl Statist. Data Anal.*, **55**, 2302–2311.
Le Cam, L. (1960) An approximation theorem for the Poisson binomial distribution. *Pacif. J. Math.*, **10**, 1181–1197.
Link, W. A. (2003) Nonidentifiability of population size from capture-recapture data with heterogeneous detection probabilities. *Biometrics*, **59**, 1123–1130.
Mao, C.-X. (2008) On the nonidentifiability of population sizes. *Biometrics*, **64**, 977–979.
McCrea, R. S. and Morgan, B. J. T. (2014) *Analysis of Capture-recapture Data*. Boca Raton: CRC Press.
Pollock, K. H. (2000) Capture-recapture models. *J. Am. Statist. Ass.*, **95**, 293–296.
Rohner, C. A., Richardson, A. J., Marshall, A. D., Weeks, S. J. and Pierce, S. J. (2011) How large is the world's largest fish?: Measuring whale sharks *rhincodon typus* with laser photogrammetry. *J. Fish Biol.*, **78**, 378–385.
Wilson, R. M. and Collins, M. F. (1992) Capture-recapture estimation with samples of size one using frequency data. *Biometrika*, **79**, 543–553.
Yang, H.-C. and Chao, A. (2005) Modeling animals' behavioral response by Markov chain models for capture-recapture experiments. *Biometrics*, **61**, 1010–1017.
Zelterman, D. (1988) Robust estimation in truncated discrete distributions with application to capture-recapture experiments. *J. Statist. Plann Inf.*, **18**, 225–237.