Fully general Chao and Zelterman estimators with application to a Whale Shark population

Alessio Farcomeni†
Sapienza – University of Rome, Rome, Italy

Summary. We introduce generalized Chao (GC) and Zelterman (GZ) estimators which include individual, time-varying and behavioural effects. Under mild assumptions in the presence of unobserved heterogeneity, the GC estimator asymptotically provides a lower bound for the population size, and is unbiased otherwise. Corrected versions guarantee bounded estimates. In order to include the best set of predictors we propose a 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, 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 behavioral research. See for instance Pollock (2000); Chao (2001); 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 modeled, in order to obtain an estimate of the number of subjects never observed, or equivalently of the size of the catchable population. Time-homogeneity and absence of behavioral 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 (while 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 untenable due to 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 MSE of Chao and Zelterman estimators can be reduced, possibly substantially, if one explicitly models some relevant sources of heterogeneity. The Zelterman and Chao estimators have already been extended to take into account

†Address for correspondence: Department of Public Health and Infectious Diseases, Sapienza - University of Rome, Piazzale Aldo Moro, 5; 00186 Rome, Italy
E-mail: alessio.farcomeni@uniroma1.it
subject-specific covariates by Böning and van der Heijden (2009) and Böning et al. (2013), respectively. In this paper we further generalize to include also time-varying covariates and certain forms of behavioral 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 number of trials and small success probabilities. A simple correction guarantees bounded estimators. The behavioral effect is modeled, 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 Böning et al. (2013), use of covariates does not necessarily decrease the Mean Squared Error (MSE). The reason is that covariates decrease the bias, but increase the standard error. We propose to choose, among all possible models, the one minimizing a biased empirical Focused Information Criterion (bFIC), which targets the minimal MSE. For background on focused information criteria 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, 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 estimators (Böning et al., 2013) can not be used because there are occasion-specific covariates available, such as cloud cover during the exploration. We will compare all possible generalized Chao and Zelterman 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 general Chao and Zelterman 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 analyze the motivating data in Section 5, and state conclusions in Section 6.

2. Background on Chao and Zelterman estimators

Let \( Y_{ij} \) be a binary indicator of detection of the \( i \)-th individual at the \( j \)-th sampling occasion, for \( j = 1, \ldots, S \). We observe the \( n \) individuals for which \( \sum_j Y_{ij} > 0 \), with a total population size of \( N \geq n \). Let \( p_l = \Pr(\sum_j Y_{ij} = l) \), and \( n_l \) denote the number of subjects observed \( l \) times, \( l \geq 0 \). We assume also that \( \sum_j Y_{ij} \) follows a Poisson distribution with parameter \( \lambda_i \). This assumption in some cases can be seen only as an approximation to the (exact) binomial distribution for \( \sum_j Y_{ij} \).

This will be more precisely formalized below. Assume further that \( \lambda_i \) 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 \( p_l = \int_0^\infty \frac{e^{-\lambda} \lambda^l}{l!} \, dF(\lambda) \). Use of Cauchy-Schwartz inequality yields \( p_l^2 \leq 2p_0p_2 \). Consequently, at least asymptotically, \( n_0 \geq n_2^2/2n_2 \) and \( N \geq n + n_1^2/2n_2 \). 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 an homogeneous Poisson distribution with parameter \( \lambda \) for \( n_1 \) and \( n_2 \), and ignoring the other frequencies. This assumption and
the properties of the Poisson distribution lead to \( \hat{\lambda} = \frac{2}{n_2/n_1} \). 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) note 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 \( i \)-th subject being observed twice. After some algebra with Poisson probabilities, the truncated likelihood can be written as

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

(1)

A log link can be used to specify a model on \( \lambda_i \), thereby including subject-specific covariates. The Zelterman estimator is recovered as the maximum likelihood solution, while 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 Böhning et al. (2013) generalized Chao’s (GC) estimator with subject-specific covariates corresponds to \( n_1 + \sum_{i=1}^{n_1+n_2} 2 \hat{\lambda}_i \), and Böhning and van der Heijden (2009) generalized Zelterman’s (GZ) estimator to \( \sum_{i=1}^{n} \left( 1 - \exp(-\hat{\lambda}_i) \right)^{-1} \), where \( \hat{\lambda}_i \) is the MLE of (1).

3. Fully general Chao and Zelterman estimators

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

We specify the following logistic regression model (see also, e.g., Coull and Agresti (1999, 2000); Farcomeni (2016)):

\[
\log \left( \frac{p_{ij}}{1 - p_{ij}} \right) = X'_{ij} \beta.
\]

(2)

In order to recover Chao and Zelterman type estimators, we base inference on a truncated likelihood, restricting to subjects observed once or twice. Given that we do not consider subjects 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 with parameter \( \sum_{j=1}^{S} p_{ij} \) as soon as \( S \) is large and \( \max_j p_{ij} \) is small.

This allows us to write a convenient approximated truncated likelihood as follows:

\[
L(\theta) = \prod_{i=1}^{n_1+n_2} \frac{\Pr(Y_{i1}, \ldots, Y_{iS})}{\Pr(1 \leq \sum_{j=1}^{S} Y_{ij} \leq 2)} = \prod_{i=1}^{n_1+n_2} \frac{\prod_{j=1}^{S} p_{ij}^{Y_{ij}} (1 - p_{ij})^{1 - Y_{ij}}}{e^{-\sum_{j=1}^{S} p_{ij}} \left( \frac{1}{\sum_{j=1}^{S} p_{ij}} \right)^{\sum_{j=1}^{S} p_{ij}} + 0.5(\sum_{j=1}^{S} p_{ij})^2},
\]

where \( p_{ij} \) is a function of model parameters as in (2). The Poisson approximation theorems have been used to approximate \( \Pr(1 \leq \sum_{j=1}^{S} Y_{ij} \leq 2) \) at the denominator of (3). To maximize (3), we obtain the score in closed form (reported in Appendix) 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 directly obtain a generalized Zelterman’s estimator as

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

Chao’s lower bound estimator is seen to correspond to

$$n + \frac{n_{1} + n_{2}}{\sum_{j=1}^{S} \hat{p}_{ij} + 0.5(\sum_{j=1}^{S} \hat{p}_{ij})^2}. \quad (5)$$

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

$$n + \frac{n_{1} + n_{2}}{\Pr(\sum_{j} Y_{ij} = 0)} \Pr(\sum_{j} Y_{ij} = 1) + \Pr(\sum_{j} Y_{ij} = 2). \quad (6)$$

Correspondingly, due to the Poisson approximation we have that

$$\Pr\left(\sum_{j} Y_{ij} = x\right) = \left(\sum_{j=1}^{S} \hat{p}_{ij}\right)^x \exp\left\{-\sum_{j=1}^{S} \hat{p}_{ij}\right\} / x!. \quad (7)$$

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

There are various intuitive reasons why (5) and (4) can be referred to as generalized Chao and Zelterman estimators. First of all, 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 based on the MLE for Zelterman, $n + E[n_0|n_1, n_2]$ for Chao). Finally (5) provides a lower bound estimate for $N$, and in absence of unobserved heterogeneity it becomes unbiased (as implied by the properties of the MLE).

A bias corrected version of Chao estimator is given by $n + n_{1}(n_{1} - 1)/2(n_{2} + 1)$ (Chao, 1989; Wilson and Collins, 1992). This (i) avoids unbounded estimates and (ii) 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 a fictitious sample point with two captures, and whose covariates are set at the column-wise means of the real sample points. We proceed similarly to avoid unboundedness of the generalized Zelterman’s 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 modeled by letting $p_{ij}$ depend on the previous capture history as follows:

$$\log\left(\frac{p_{ij}}{1 - p_{ij}}\right) = X_{ij}'\beta + g_j(Y_{i,j-1}, \ldots, Y_{i1}), \quad (7)$$
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 is 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 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 \( \sum_{l=1}^{j-1} Y_{il} > 0 \), Chao’s estimator (5) is asymptotically unbounded even if \( N \) is finite. Our task is then now 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 \( \lambda_i = \sum_j E[Y_{ij}] \).

First, if \( \{Y_{ij}, j > 0\} \) can be shown to be \( \phi \)-mixing under the assumed model, then the Poisson approximation is still valid as shown by Chen (1975). The idea of \( \phi \)-mixing is rather simple and involves the fact that the dependence between \( Y_{ij} \) and \( Y_{ij+k} \) decreases fast enough with \( k \), and the two indicators are eventually independent when \( k \to \infty \). A simple class of \( \phi \)-mixing processes is given by Harris-recurrent Markov chains. Hence we can fix \( m > 0 \) (with \( m << S \)) and define \( g_j(Y_{i,j-1}, \ldots, Y_{i1}) = \sum_{l=1}^{m} \eta Y_{i,j-l} \). These correspond to transient behavioural response, where 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 (\( m = 1 \)) are 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
\[ \Pr(Y_{ij} = 1, Y_{ij'} = 1) \geq (\leq) \Pr(Y_{ij} = 1) \Pr(Y_{ij'} = 1) \tag{9} \]
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 (5) and (4). To see this, note that positive association is equivalent to
\[ \Pr(Y_{ij} = 1|Y_{ij'} = 1) - \Pr(Y_{ij} = 1|Y_{ij'} = 0) \geq 0. \tag{10} \]
Assume without loss of generality that \( j > j' \). Suppose \( \eta \geq 0 \). We have that under (8)
\[ \log \left( \frac{\Pr(Y_{ij} = 1|Y_{ij'} = 1)}{\Pr(Y_{ij} = 0|Y_{ij'} = 1)} \right) = \beta_{0j} + X'_{ijj} \beta + \eta \]
and
\[ \log \left( \frac{\Pr(Y_{ij} = 1|Y_{ij'} = 0)}{\Pr(Y_{ij} = 0|Y_{ij'} = 0)} \right) \leq \beta_{0j} + X'_{ijj} \beta + \eta, \]
as it is less likely that \( \sum_{l<j} Y_{il} > 0 \). Hence, (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), that is, for functions of the
kind \( g_j(Y_{i,j-1}, \ldots, Y_{i1}) = \eta I(\sum Y_{i\ell} > k) \) 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 (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 (7) \( E[Y_{ij}] \neq p_{ij} \) in general. The parameter of interest \( \lambda_i = \sum_j E[Y_{ij}] \) 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 (Böning, 2008). First, note that population size estimators are expressed as \( \hat{N} = h(\theta) = \sum_{i=1}^N \Delta_i h_i(\theta) \), where \( \Delta_i = 1 \) if the \( i \)-th subject is observed once or twice (and zero otherwise), \( \theta \) is a short-hand notation for parameters involved in (7), and \( h(\theta) \) is the mapping from \( \theta \) to \( \hat{N} \). By conditioning,

\[
\text{Var}(\hat{N}) = \text{Var}(E[\hat{N}|\Delta_i]) + E[\text{Var}(\hat{N}|\Delta_i)].
\]

(11)

For the generalized Chao’s estimator, after some algebra it can be seen that the first term is unbiasedly estimated by

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

where \( \hat{p}_i = \exp(-\sum_{j=1}^S \hat{p}_{ij})(1 + \sum_{j=1}^S \hat{p}_{ij}/2) \sum_{j=1}^S \hat{p}_{ij} \). For the generalized Zelterman’s estimator, after some algebra we obtain

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

The second term in (11) can be approximated as

\[\triangledown h(\hat{\theta})J(\hat{\theta})^{-1} \triangledown h(\hat{\theta}),\]

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

We now discuss model selection. As intuitive, the (asymptotic) bias of the generalized Chao estimator is non-positive, and is closer to zero than the bias of the original Chao estimator. On the other hand, the variance of the estimate is often larger due to the uncertainty linked with the use of additional parameters to model covariate effects. The trade-off between bias and variance is often in favor 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 considered, there are \( q2^p \) possible models. It is likely that the one minimizing the MSE is not the full or the empty one. A careful tuning of
the covariates included is therefore very important. To date there are no direct approaches for performing this kind of calibration.

We here propose an empirical Focused Information Criterion (FIC) which can be used to select the estimator with least (asymptotically) expected Mean Squared Error (MSE) for the population size estimator. See Bartolucci and Lupparelli (2008) for an example of use of FIC in the context of population size estimation. Unlike other information criteria FIC focuses on the parameter of primary interest. In our context, this is $N$. 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 (hence, FIC) for $N$. The expected MSE is often obtained analytically in other contexts, and data is 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 $q2^p$ possible models, or a subset of them) and $\hat{se}_1, \ldots, \hat{se}_v$ their estimated standard errors. We propose the following information criterion:

$$bFIC_j = (\hat{N}_j - \max_j \hat{N}_j)^2 + \hat{se}_j^2,$$

for $j = 1, \ldots, v$. Note that $bFIC$ is a biased estimator of the MSE of $\hat{N}_j$ as the first addend is a biased estimator of bias. The model minimizing bFIC is selected. The idea behind bFIC is that, given all candidate estimators are negatively biased, the largest one is the less biased. Hence the closest the estimate to $\max_j \hat{N}_j$, the less biased it is. Consequently, the first addend in (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 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 will compare bFIC with 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 $\chi^2$ 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 reasons of space we focus only on Chao and GC estimators. We evaluate bias and MSE of Chao, oracle GC estimator (when the true model is known), and 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 break down of each estimator and the performance of bias-corrected-Chao and bias-corrected-GC as described above.

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 are four in the first scenario and sixteen in the second. We report bias and square root of MSE (RMSE) of each
Table 1. Bias, RMSE, and probability of failure ($p_f$) for Chao, Generalized Chao (GC) using all covariates (oracle), GC with covariates chosen using AIC, GC with covariates chosen using bFIC, and their bias-corrected (BC) versions. 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.

<table>
<thead>
<tr>
<th>N</th>
<th>$S$</th>
<th>Bias</th>
<th>RMSE</th>
<th>$p_f$</th>
<th>Bias</th>
<th>RMSE</th>
<th>$p_f$</th>
<th>Bias</th>
<th>RMSE</th>
<th>$p_f$</th>
<th>Bias</th>
<th>RMSE</th>
<th>$p_f$</th>
</tr>
</thead>
<tbody>
<tr>
<td>250</td>
<td>8</td>
<td>-136.0</td>
<td>136.9</td>
<td>.05</td>
<td>-20.1</td>
<td>201.7</td>
<td>.20</td>
<td>-167.8</td>
<td>166.8</td>
<td>.13</td>
<td>-137.4</td>
<td>142.0</td>
<td>.06</td>
</tr>
<tr>
<td>250</td>
<td>30</td>
<td>-93.7</td>
<td>86.3</td>
<td>.01</td>
<td>-77.9</td>
<td>64.8</td>
<td>.18</td>
<td>-79.8</td>
<td>65.1</td>
<td>.24</td>
<td>-92.2</td>
<td>81.4</td>
<td>.02</td>
</tr>
<tr>
<td>1000</td>
<td>8</td>
<td>-585.0</td>
<td>586.7</td>
<td>.21</td>
<td>-640.8</td>
<td>675.7</td>
<td>.18</td>
<td>-567.1</td>
<td>550.8</td>
<td>.03</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>30</td>
<td>-559.8</td>
<td>566.4</td>
<td>.04</td>
<td>-325.0</td>
<td>324.4</td>
<td>.04</td>
<td>-286.2</td>
<td>263.8</td>
<td>.01</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>100</td>
<td>-208.2</td>
<td>207.4</td>
<td>.02</td>
<td>-164.3</td>
<td>163.0</td>
<td>.02</td>
<td>-151.3</td>
<td>146.9</td>
<td>.01</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>N</th>
<th>$S$</th>
<th>Bias</th>
<th>RMSE</th>
<th>$p_f$</th>
<th>Bias</th>
<th>RMSE</th>
<th>$p_f$</th>
<th>Bias</th>
<th>RMSE</th>
<th>$p_f$</th>
<th>Bias</th>
<th>RMSE</th>
<th>$p_f$</th>
</tr>
</thead>
<tbody>
<tr>
<td>250</td>
<td>8</td>
<td>-148.4</td>
<td>147.2</td>
<td>.00</td>
<td>-164.5</td>
<td>162.4</td>
<td>.00</td>
<td>-177.3</td>
<td>172.5</td>
<td>.00</td>
<td>-160.7</td>
<td>158.9</td>
<td>.00</td>
</tr>
<tr>
<td>250</td>
<td>30</td>
<td>-98.4</td>
<td>97.0</td>
<td>.00</td>
<td>-79.1</td>
<td>82.8</td>
<td>.00</td>
<td>-80.4</td>
<td>82.8</td>
<td>.00</td>
<td>-95.0</td>
<td>95.9</td>
<td>.00</td>
</tr>
<tr>
<td>1000</td>
<td>8</td>
<td>-575.3</td>
<td>559.8</td>
<td>.00</td>
<td>-609.8</td>
<td>601.6</td>
<td>.00</td>
<td>-634.5</td>
<td>621.6</td>
<td>.00</td>
<td>-583.2</td>
<td>560.1</td>
<td>.00</td>
</tr>
<tr>
<td>1000</td>
<td>30</td>
<td>-402.8</td>
<td>396.7</td>
<td>.00</td>
<td>-329.6</td>
<td>331.9</td>
<td>.00</td>
<td>-329.6</td>
<td>331.9</td>
<td>.00</td>
<td>-296.4</td>
<td>280.6</td>
<td>.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>N</th>
<th>$S$</th>
<th>Bias</th>
<th>RMSE</th>
<th>$p_f$</th>
<th>Bias</th>
<th>RMSE</th>
<th>$p_f$</th>
<th>Bias</th>
<th>RMSE</th>
<th>$p_f$</th>
<th>Bias</th>
<th>RMSE</th>
<th>$p_f$</th>
</tr>
</thead>
<tbody>
<tr>
<td>250</td>
<td>8</td>
<td>-149.6</td>
<td>147.6</td>
<td>.03</td>
<td>-137.6</td>
<td>145.4</td>
<td>.14</td>
<td>-162.9</td>
<td>163.1</td>
<td>.08</td>
<td>-121.0</td>
<td>140.0</td>
<td>.22</td>
</tr>
<tr>
<td>250</td>
<td>30</td>
<td>-89.2</td>
<td>84.3</td>
<td>.01</td>
<td>-82.7</td>
<td>80.7</td>
<td>.02</td>
<td>-83.9</td>
<td>81.7</td>
<td>.04</td>
<td>-74.0</td>
<td>69.9</td>
<td>.01</td>
</tr>
<tr>
<td>1000</td>
<td>8</td>
<td>-612.1</td>
<td>636.1</td>
<td>.11</td>
<td>-591.2</td>
<td>614.5</td>
<td>.08</td>
<td>-427.2</td>
<td>390.9</td>
<td>.14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>30</td>
<td>-356.5</td>
<td>342.2</td>
<td>.00</td>
<td>-348.5</td>
<td>344.6</td>
<td>.00</td>
<td>-348.5</td>
<td>344.6</td>
<td>.00</td>
<td>-254.9</td>
<td>246.4</td>
<td>.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>N</th>
<th>$S$</th>
<th>Bias</th>
<th>RMSE</th>
<th>$p_f$</th>
<th>Bias</th>
<th>RMSE</th>
<th>$p_f$</th>
<th>Bias</th>
<th>RMSE</th>
<th>$p_f$</th>
<th>Bias</th>
<th>RMSE</th>
<th>$p_f$</th>
</tr>
</thead>
<tbody>
<tr>
<td>250</td>
<td>8</td>
<td>-159.0</td>
<td>157.5</td>
<td>.03</td>
<td>-154.9</td>
<td>133.4</td>
<td>.00</td>
<td>-169.0</td>
<td>191.4</td>
<td>.00</td>
<td>-159.4</td>
<td>156.5</td>
<td>.00</td>
</tr>
<tr>
<td>250</td>
<td>30</td>
<td>-87.4</td>
<td>80.3</td>
<td>.00</td>
<td>-80.3</td>
<td>77.8</td>
<td>.00</td>
<td>-81.2</td>
<td>79.4</td>
<td>.00</td>
<td>-76.3</td>
<td>71.6</td>
<td>.00</td>
</tr>
<tr>
<td>1000</td>
<td>8</td>
<td>-44.6</td>
<td>43.7</td>
<td>.01</td>
<td>-37.1</td>
<td>35.0</td>
<td>.00</td>
<td>-39.7</td>
<td>37.1</td>
<td>.00</td>
<td>-39.4</td>
<td>37.0</td>
<td>.00</td>
</tr>
<tr>
<td>1000</td>
<td>30</td>
<td>-969.6</td>
<td>966.4</td>
<td>.03</td>
<td>-623.1</td>
<td>624.1</td>
<td>.00</td>
<td>-623.0</td>
<td>623.5</td>
<td>.04</td>
<td>-566.2</td>
<td>395.5</td>
<td>.00</td>
</tr>
<tr>
<td>1000</td>
<td>100</td>
<td>-182.2</td>
<td>166.6</td>
<td>.00</td>
<td>-160.8</td>
<td>107.1</td>
<td>.00</td>
<td>-161.1</td>
<td>107.4</td>
<td>.00</td>
<td>-142.9</td>
<td>93.2</td>
<td>.00</td>
</tr>
</tbody>
</table>

procedure, and the failure probabilities. We estimate the probability of failure by evaluating outlying MSE values, and report the Reweighted Minimum Covariance Determinant (RMCD) over the replicates for bias and MSE. The RMCD is a robust measure of location which can adapt to differing numbers of outliers (see e.g. 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 GC and Chao estimators are negatively biased as expected. GC is biased as it always ignores the relevant covariate $X_0$. Additionally, the Poisson approximation underlying GC is good enough even with $S = 8$, as GC results are satisfactory even in this case. GC is almost always less biased than Chao estimator. This happens at the price of a slightly larger standard error, and in some cases it can even lead the oracle GC to have a larger MSE than Chao. Both estimators have a tendency to break down, with of course a larger breakdown probability for GC than for Chao estimator. Failure is more likely when $S = 8$, as in this case the sampling fraction is very small. Bias-corrected versions of Chao and GC never fail, and in most cases result in MSE which is comparable to 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 bFIC GC 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 covari-
at risk. 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 for reasons of space) with 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 Maldives is one of these aggregation points. Several characteristics of this fascinating species are still unknown, including how they interact, where and when they breed. The subpopulation of the South Ari atoll in Maldives is also peculiar in that surfacing sharks are almost always juveniles.

Every day for a period of six months the area surrounding South Ari atoll was explored by the MWSRP boat. The minimal length observed was three meters, indicating birth well before the beginning of the study period. Mean was 5.73 ± 1.24 meters. 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 shall be also mentioned that estimates reported below are remarkably stable at a sensitivity analysis which removes 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 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/no), sea state (calm/rough), season (high/low). It shall 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, being based on the same likelihood, for the same model specification the AIC of GC and GZ correspond. Of course, their bFIC might be substantially different (see below). In the table, GC\(_o\) denotes a GC estimator based on the subject-specific covariates gender and length, GC\(_t\) a GC estimator based on the occasion-specific covariates, GC\(_b\) and GC\(_b;\text{Markovian}\) 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 described in Huggins (1989), and available in the mra R package, but unfortunately due to the very large number of occasions it was not possible to obtain any estimator. It can be noted that similar computational issues might arise with GC and GZ estimators only with a much larger number of occasions, also due to the fact that likelihood is restricted to singletons and doubletons.

Given that we consider two possible behavioural effects (classical/Markovian) there are 32 possible models. We do not show all of them, but 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 ones, according to bFIC corresponds to a GC including gender and length, cloud coverage and

| No. captures | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 14 | 16 | 17 | 18 | 21 | ≥ 26 |
|-------------|---|---|---|---|---|---|---|---|---|----|----|----|----|----|----|----|----|----|     |
| No. sights  | 45| 15| 9 | 7 | 3 | 3 | 4 | 3 | 3 | 2  | 4  | 4  | 3  | 1  | 2  | 1  | 1  | 1  | 2   |     |
Table 3. Population size estimates, sampling fraction $n/\hat{N}$, standard errors for $\hat{N}$ and AIC for the Whale Shark data set

<table>
<thead>
<tr>
<th>Estimator</th>
<th>$\hat{N}$</th>
<th>$n/\hat{N}$</th>
<th>S.E.</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chao</td>
<td>179</td>
<td>62%</td>
<td>24.70</td>
<td>828.05</td>
</tr>
<tr>
<td>GC</td>
<td>195</td>
<td>57%</td>
<td>44.14</td>
<td>830.13</td>
</tr>
<tr>
<td>GC_t</td>
<td>179</td>
<td>62%</td>
<td>27.46</td>
<td>818.41</td>
</tr>
<tr>
<td>GC_b</td>
<td>204</td>
<td>55%</td>
<td>39.30</td>
<td>804.56</td>
</tr>
<tr>
<td>GC_b, Markovian</td>
<td>179</td>
<td>62%</td>
<td>27.50</td>
<td>829.01</td>
</tr>
<tr>
<td>GC_ot</td>
<td>187</td>
<td>60%</td>
<td>38.43</td>
<td>819.53</td>
</tr>
<tr>
<td>GC_oth</td>
<td>219</td>
<td>51%</td>
<td>60.18</td>
<td>802.88</td>
</tr>
<tr>
<td>GC_tb</td>
<td>204</td>
<td>55%</td>
<td>39.17</td>
<td>794.20</td>
</tr>
<tr>
<td>GC_othb</td>
<td>166</td>
<td>67%</td>
<td>24.18</td>
<td>803.79</td>
</tr>
<tr>
<td>Zelterman</td>
<td>230</td>
<td>49%</td>
<td>48.74</td>
<td>828.05</td>
</tr>
<tr>
<td>GZ</td>
<td>241</td>
<td>46%</td>
<td>61.58</td>
<td>830.13</td>
</tr>
<tr>
<td>GZ_t</td>
<td>228</td>
<td>49%</td>
<td>48.73</td>
<td>818.41</td>
</tr>
<tr>
<td>GZ_b</td>
<td>342</td>
<td>33%</td>
<td>97.99</td>
<td>804.56</td>
</tr>
<tr>
<td>GZ_b, Markovian</td>
<td>229</td>
<td>49%</td>
<td>48.83</td>
<td>829.01</td>
</tr>
<tr>
<td>GZ_ot</td>
<td>231</td>
<td>48%</td>
<td>54.45</td>
<td>819.53</td>
</tr>
<tr>
<td>GZ_oth</td>
<td>343</td>
<td>33%</td>
<td>105.71</td>
<td>802.88</td>
</tr>
<tr>
<td>GZ_othb</td>
<td>342</td>
<td>33%</td>
<td>95.77</td>
<td>794.20</td>
</tr>
<tr>
<td>GZ_othb</td>
<td>253</td>
<td>44%</td>
<td>63.29</td>
<td>803.79</td>
</tr>
</tbody>
</table>

Table 4. Coefficient estimates, standard errors and p-value for the Whale Shark data set. Model selected through bFIC.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>$\beta$</th>
<th>S.E.</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-7.60</td>
<td>1.59</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Gender</td>
<td>-0.86</td>
<td>1.16</td>
<td>0.458</td>
</tr>
<tr>
<td>Length</td>
<td>0.52</td>
<td>0.26</td>
<td>0.050</td>
</tr>
<tr>
<td>Cloud</td>
<td>-0.41</td>
<td>0.05</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Behav. Eff.</td>
<td>1.87</td>
<td>0.29</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

(persistent) behavioural effect. 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 chosen by bFIC more credible than the one selected with AIC. In order to evaluate the goodness of fit of this model, we compute the predicted cell counts with the reported parameter estimates, that correspond to $\hat{n}_1 = 46.12$ and $\hat{n}_2 = 13.88$. The resulting $\chi^2$ 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 9 females were encountered. The extremely low number of females does not allow the model to discriminate 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 approach the surface. When the day is cloudy, the shadow of surfacing sharks is dim, and therefore harder to detect. Additionally, for obvious reasons longer sharks are more easily detected than shorter ones. The behavioural effect on the other hand is explained by a persistence behaviour of the researchers, who tended to return to
the same spots were a shark was observed.

6. Discussion

We have presented two new population size estimators, the generalized Chao and generalized Zelterman’s 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 an homogeneous Poisson only for subjects observed at most twice.

We have provided a new information criterion, the bFIC, to select the model which is more likely to minimize the MSE of the population size estimate. We have seen in simulation that bFIC can substantially improve over AIC in certain cases. We speculate the bFIC criterion compares so well with AIC due to the strong bias-variance trade-off associated with model choice for generalized Chao and Zelterman 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 rather fast and very precise (while 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 fully generalize also the extended Chao’s estimator 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, in order to balance study length and precision of the estimates. For more classical models this was considered for instance in Alumni Fegatelli and Farcomeni (2016).

Acknowledgements

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

A. Score of conditional likelihood

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

$$l(\theta) = \sum_{i=1}^{n_1+n_2} \sum_{j=1}^{S} Y_{ij} \log(p_{ij}) + \sum_{j=1}^{S} (1 - Y_{ij}) \log(1 - p_{ij}) + \sum_{j=1}^{S} p_{ij} - \log \left( \sum_{j=1}^{S} p_{ij} + \frac{(\sum_{j=1}^{S} p_{ij})^2}{2} \right)$$
The score can then be obtained as
\[
\frac{\partial l(\theta)}{\partial \beta_h} = \sum_{i=1}^{n_1+n_2} \sum_{j=1}^S x_{ijh}(Y_{ij} - p_{ij}) + \sum_{j=1}^S \frac{p_{ij}x_{ijh}}{1 + \exp(\mathbf{X}_{ij}'\beta + g_j(Y_{i,j-1}, \ldots, Y_{i1}))} \\
- \frac{1 + \sum_j p_{ij}}{\sum_j p_{ij} + 0.5(\sum_j p_{ij})^2} \sum_{j=1}^S \frac{p_{ij}x_{ijh}}{1 + \exp(\mathbf{X}_{ij}'\beta + g_j(Y_{i,j-1}, \ldots, Y_{i1}))} \\
= \sum_{i=1}^{n_1+n_2} \sum_{j=1}^S x_{ijh}(Y_{ij} - p_{ij}) \\
+ \frac{0.5(\sum_j p_{ij})^2 - 1}{\sum_j p_{ij} + 0.5(\sum_j p_{ij})^2} \sum_{j=1}^S \frac{p_{ij}x_{ijh}}{1 + \exp(\mathbf{X}_{ij}'\beta + g_j(Y_{i,j-1}, \ldots, Y_{i1}))}
\]
and similarly for any parameter involved in \(g_j(\cdot)\).

References


