Environmental and Ecological Statistics manuscript No.
(will be inserted by the editor)

A generalized Chao estimator with measurement error and external information

Francesco Dotto · Alessio Farcomeni

Received: date / Accepted: date

Abstract We present a generalized Chao (GC) estimator based on a subject-occasion-specific design matrix. We then extend the GC estimator to (i) external information, in the form of non-linear constraints on subpopulation sizes and (ii) measurement error. For the first, we propose a reparameterization of the estimating equations. As a result, the constrained MLE can be found with no additional computational efforts. For the second we generalize SIMEX procedure to multiple measurement methods. In simulation we show that (even incorrect) external information can substantially decrease the MSE. We illustrate with an application to a whale shark (*Rhincodon typus*) population, where mostly jouvenile males are observed. We use external information on gender ratio of whale sharks to correct for low catchability of females, and our multivariate SIMEX procedure to correct for measurement error in assessment of shark length. The resulting population size estimates are about 60% larger than the unconstrained-uncorrected counterparts.

Keywords Chao estimator · SIMEX · external information

1 Introduction

The assessment and monitoring of abundance of animal populations is a central issue for conservation, management and in general for comprehension of natural phenomena. It is well acknowledged that exhaustive enumeration of members of a population of interest is very often impossible. Hence population size is almost always estimated. This can be done for instance by means of distance sampling, mark-recapture or other kind of experiments. See for instance McCrea and Morgan (2014) among many possible references. Capture-recapture experiments for closed populations are based on repeated identification of subjects within a population of interest. The information on subjects identified

F. Dotto · A. Farcomeni E-mail: alessio.farcomeni@uniroma1.it
Sapienza - University of Rome

at least once is used to estimate how many were never seen, and consequently the total number of subjects originally available for sampling. Among several possible estimates of population size a very popular one is that of Chao (1987, 1989), which is a very simple function of the number of subjects identified exactly once and the number identified exactly twice. It is unbiased under homogeneity, and has a negative bias (that is, the estimate is expected to be lower than the true population size) under any form of unspecified unobserved heterogeneity. This is a particularly complex setting (Farcomeni and Tardella (2012) and references therein). More in general, it can be seen that Chao estimator is nearly unbiased if the mean capture probability of the undetected subjects is the same as that of the subjects that are detected in only one occasion (Chao and Colwell, 2017). Also if that all seldomly sampled subjects have approximately homogeneous detection probabilities in any occasion there is near unbiasedness. In this sense, frequently sampled subject can have highly heterogeneous capture probabilities without affecting the properties of Chao's estimator.

Chao's estimator was generalized to subject-specific covariates in Böhning et al (2013) and to the completely general situation of subject-occasion specific covariates and behavioural response (Farcomeni, 2011) in Farcomeni (2017). The latter is based on the *capture history* of subjects identified exactly once or twice. Generalized Chao estimators take into account available observed heterogeneity. They are unbiased if all heterogeneity has been included in the model, and have a negative bias under any form of unspecified unobserved heterogeneity. The generalized Chao estimator often has an impressively lower bias than the classical Chao estimator. If well calibrated (see Farcomeni (2017) and Bartolucci and Lupparelli (2008) on this point) its final mean square error can be minimal conditionally on the information available.

In this work we deal with two issues, which are both quite general. On the one hand, external information might be available and could be used to decrease the MSE of the final population size estimate. External information is often available in some form, and usually taken into account through Bayesian methods (e.g., King et al 2005, 2014). Here we take a different route. We focus on knowledge of non-linear functions of subpopulation sizes, and specifically of the gender ratio, but our approach is more general and straightforward to extend. We include external information by expressing probability of capture (also) as a function of subpopulation indicators, instead of the related parameters, we constrain them by solving a system of non-linear equations. Another issue that is often open with real data problems is that of measurement error. One or more covariates might be measured with noise, and treating them as correctly measured can lead to bias (e.g., Carrol et al (2006)). For general reviews on measurement error see Carrol et al (2006); Gustafson (2003), while contributions in capture-recapture are for instance Hwang and Huang (2003); Huggins and Hwang (2010); Stoklosa et al (2016). Here we will use replicates (White et al, 2001) in a SIMulation-EXtrapolation (SIMEX) framework (Gould et al, 1999). See also Keogh and White (2014). The SIMEX framework proceeds by randomly perturbing the mismeasured covariate at different noise

levels (simulation step), and then using an appropriate model to estimate parameters with a denoised covariate (extrapolation step). We found this quite convenient for GC estimators, while other approaches (e.g., regression calibration) might be more difficult to consider. A limitation of the classical SIMEX framework which we overcome in this paper is the fact that usually a single measurement method is used, while (as in our application) the same covariate might be repeatedly measured with different measurement methods of different precision. We will show how to take into account efficiently external information and measurement error (via replicated measurements, possibly made with different techniques) in conjunction with generalized Chao estimators, but our approach is more general and could be extended in principle to any population size estimation procedure.

We are motivated by an application to the enumeration of a whale shark (*Rhincodon typus*) population which is residential in the South Ari atoll, in Maldives. The same data set has been considered in Farcomeni (2017), where the issues of measurement error and external information have been ignored. A substantial difference in terms of final estimates is obtained when taking into account these two issues. The issue of measurement error arises with the shark length, which is an important covariate. Shark length is important as sightings mostly occur when spotting the shark's shadow from a boat, and of course the longer the shark the largest the shadow, and the more likely the sighting. Length was measured at each sighting occasion with one of three methods, one of which is a simple eyeballing from the boat. All methods are prone to measurement error. Let us note that shark length can be assumed to be approximately constant over the short time frame of the study (Rohner et al, 2015). External information can be very useful since a very strong gender bias in observation is common all over the world (e.g., Meekan et al (2006); Rohner et al (2015)), with overwhelming majority of sightings being of male (and jouvenile) animals. This bias is commonly attributed to different surfacing habits of males and females, and not to a truly severely unbalanced sex ratio. The consequence for capture-recapture studies is that, as in Farcomeni (2017), female sightings are often so rare that no gender effects are significantly detected. The consequence is a severe underestimation of the number of females, and consequently of the total population size. In this work we propose including external information about sex ratio in the population size estimator, in the form of a constraint/offset. A crucial contribution to the issue of estimation of sex ratio in whale sharks is given by the seminal work of Chang et al (1997), where embryos of whale sharks were observed and a gender ratio of 1.02:1 (male to females) is reported. The gender ratio of these animals is therefore quite similar to that of other animal populations (and humans). Given that it is often reported that there are no significant survival differences by gender, we can assume that the gender ratio is approximately 1.02:1 at all ages. The assumption that the gender ratio is constant conditionally on age is a simplifying one, and possibly it is not true. It shall be noted however that estimates are fairly stable for several values of the average gender ratio, which

will very likely lie in the interval used at our sensitivity analysis stage. See Section 5 for more details.

For our problem, a direct account of the capture-recapture likelihood might seem more natural, given that a simple explicit formula for the GC estimator and its extension is not available. Unfortunately, a direct account of the capture-recapture likelihood is unfeasible for the data at hand as there are $S = 181$ occasions. If we wanted to consider the likelihood for full data we would only be able to work with individual counts of captures, ignoring time-specific covariates and hence increasing uncertainty. Additionally, GC allows us to take into account residual unobserved heterogeneity non-parametrically, without having to specify a working distribution. This is the recommended route whenever, as in our case, different assumptions lead to different conclusions.

The rest of the paper is as follows: in the next section we give our set up and necessary background on Chao and generalized Chao estimators. In Section 3 we discuss external information and heteroscedastic measurement error. We illustrate the performance of the proposed methodology with a simulation study in Section 4. Analysis of the whale shark data set is reported in Section 5 and some concluding remarks are given in Section 6.

2 Set up: generalized Chao estimators

Our set up is based on a repeated identification experiment, based on S occasions. An underlying population of N subjects is searched for at each occasion, and at the end of experiment $n \leq N$ subjects have been identified at least once. Call Y_{ij} the binary indicator for the i-th subject having been identified at the j-th occasion, with $\Pr(Y_{ij} = 1) = p_{ij}$. The population is (approximately) closed, so that there are no births, deaths, immigration or emigration during the observation period. Let n_j, $j \geq 0$, denote the number of subjects observed exactly j-times. Obviously, $n = \sum_{j>0} n_j$ and $N = n_0 + n$.

Chao (1987, 1989) estimator is obtained through the simple expression $\hat{N} = n + n_1^2/2n_2$. It can be shown with a simple stochastic inequality that $E[\hat{N}] \leq N$ given any unspecified mixing distribution summarizing unobserved heterogeneity, and that $E[\hat{N}] = N$ under homogeneity (that is, if all subjects share the same probability of being observed at every occasion).

Let now X_{ij} denote a vector of subject-occasion specific covariates. It is shown in Farcomeni (2017) that by expressing

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = X_{ij}'\beta$$

and maximizing the approximate truncated likelihood

$$L(\beta) = \prod_{i=1}^{n_1+n_2} \frac{\Pr(Y_{i1},\ldots,Y_{iS})}{\Pr\left(1 \leq \sum_{j=1}^{S} Y_{ij} \leq 2\right)} = \tag{1}$$

$$= \prod_{i=1}^{n_1+n_2} \frac{\prod_{j=1}^{S} p_{ij}^{Y_{ij}} (1-p_{ij})^{1-Y_{ij}}}{e^{-\sum_{j=1}^{S} p_{ij}} \left(\sum_{j=1}^{S} p_{ij} + 0.5 (\sum_{j=1}^{S} p_{ij})^2 \right)},$$

where only subjects having been observed at most twice are included, a generalized Chao estimator is obtained by estimating

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

The properties of the generalized Chao estimator are that bias is non-positive (once again, that $E[\hat{N}] \leq N$ asymptotically), and $E[\hat{N}] = N$ when all sources of heterogeneity have been included in the model. Additionally, when no covariates are used, (2) reduces to Chao's estimator.

3 External information and measurement error

For simplicity we assume only X_{1ij} is measured with error. Generalization to several covariates measurement error is relatively straightforward. Assume furthermore that equality constraints are available on a possibly non-linear function of the total population size N. Once again for simplicity and in parallel with our application we assume the constraints are a function g of a finite number of k sub-population sizes $N_1, \ldots, N_k$.

The general model can then be expressed as

$$\begin{cases} \log\left(\frac{p_{ij}}{1-p_{ij}}\right) &= X'_{ij}\beta \\ X_{1ij} &= W_i + \varepsilon_{ij} \\ g(\hat{N}_1, \ldots, \hat{N}_k) &= c. \end{cases}$$

Here the first row of the model, when we restrict to singletons and doubletons, corresponds to the generalized Chao estimator of Farcomeni (2017). See also Bartolucci and Forcina (2006). The second row clarifies that X_{1ij} is only a noisy version of a time-fixed underlying covariate W_i, and the last row specifies the constraint summarizing external information.

We have heteroscedastic measurement error as we assume $\varepsilon_{ij} \sim N(0, \sigma^2_{ij})$. A further modeling assumption is needed for σ^2_{ij}, for identifiability. In our application, X_{1ij} is shark length as observed at the j-th occasion. The error variance can be assumed to be constant within measurement type, where we have three measurement methods. Therefore, depending on a known covariate (measurement method), we assume that $\sigma^2_{ij} = \sigma^2_{h_j}$, with $h = 1, 2, 3$, respectively.

Additionally, in our application the external information we have regards gender ratio. Letting N_m denote the number of males and N_f denote the number of females in the population, external information is that $N_m/N_f = 1.02$ (a detailed discussion is given below). When we plug-in the estimator for each

subpopulation, we have consequently that in our application the constraint is expressed as

$$\frac{n_m + \sum_{i=1}^{n_{1m}+n_{2m}} \frac{1}{\sum_j p_{ij} + 0.5(\sum_j p_{ij})^2}}{n_f + \sum_{i=1}^{n_{1f}+n_{2f}} \frac{1}{\sum_j p_{ij} + 0.5(\sum_j p_{ij})^2}} = c, \tag{3}$$

where $c = 1.02$.

In our proposed model we use a specific link function (the logistic one) and a specific functional constraint to take into account external information. We recommend routinely checking these assumptions by assessing the goodness of fit. This can be simply done, for instance, by computing G or the chi-squared statistics for the observed cells. Instead of evaluating significance of these statistics, we suggest computing the Root Mean Squared Error of Approximation (RMSEA), which is the square root of the difference between the observed statistic and its degrees of freedom under the null, divided by the square root of the degrees of freedom multiplied by $n_1 + n_2 - 1$. RMSEA$< .1$ indicates acceptable fit. In case the fit is not acceptable other link functions (e.g., complementary-log-log, probit) or other functional forms for the constaint could be explored.

3.1 Inference with external information

Let us ignore for a moment the presence of covariates measured with error, and focus on external information. This is given in the form of a constraint, hence our task is now that of maximizing the likelihood (1) under a constraint. We could do so by basing inference on a penalized likelihood function $L + \lambda(g(\hat{N}_1, \ldots, \hat{N}_k) - c)^2$. Any finite positive value of the penalty parameter would enforce a soft-form of the constraint by penalizing the likelihood function for departures from the known gender ratio. This would allow the user to fine balance disagreement between the MLE and the constrained MLE with small departures from the constraint. It shall be noted though that this trade-off would have to be calibrated heuristically, as the data contain no (additional) information on the matter. Additionally, penalized likelihood maximization would be cumbersome and time-consuming. In order to overcome these difficulties we enforce the constraint *exactly*, by including working covariates and obtaining the corrisponding β parameters in order to satisfy the constrain. An additional advantage is that a solution is found at approximately the same computational expense needed for maximization of the unconstrained likelihood. For the sake of exposition assume that, as in our application, $k = 2$. Let $X_{ig} = 1$ if the i-th subject belongs to the first population and $X_{ig} = -1$ if the i-th subject belongs to the second population. This covariate is associated with an additional parameter β_g. Define

$$\log\left(\frac{p_{ij}^{(g)}}{1 - p_{ij}^{(g)}}\right) = X_{ij}'\beta + \beta_g X_{ig},$$

and $\hat{N}$ as in (2), but as a function of $p_{ij}^{(g)}$. It is straightforward to check that for each $\beta \in \mathcal{R}^p$ there exists $\beta_g \in \mathcal{R}$ such that (3) is satisfied. The sum-to-zero parameterization of X_{ig} indeed guarantees existence, as in fact as β_g diverges the ratio of the probabilities of occurrence can diverge as well. For each β, β_g is found by solving a simple non-linear equation as

$$\beta_g \,:\, g\left(n_m + \sum_{i=1}^{n_{1m}+n_{2m}} \frac{1}{\sum_j p_{ij}^{(g)} + 0.5(\sum_j p_{ij}^{(g)})^2}, n_f + \sum_{i=1}^{n_{1f}+n_{2f}} \frac{1}{\sum_j p_{ij}^{(g)} + 0.5(\sum_j p_{ij}^{(g)})^2}\right) = c, \tag{4}$$

where n_m and n_f denote the number of uniquely identified subjects for the first population (m) and for the second (f), respectively. Call $\beta_g(\beta)$ the solution to (4).

With this set up, the constrained MLE can be found by maximizing the likelihood

$$L(\beta) = \prod_{i=1}^{n_1+n_2} \frac{\prod_{j=1}^{S} p_{ij}^{Y_{ij}}(1-p_{ij})^{1-Y_{ij}}}{e^{-\sum_{j=1}^{S} p_{ij}}\left(\sum_{j=1}^{S} p_{ij} + 0.5(\sum_{j=1}^{S} p_{ij})^2\right)},$$

with

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = X_{ij}'\beta + \beta_g(\beta)X_{ig},$$

for $\beta \in \mathcal{R}^p$.

In order to compute standard errors we have to take into account the fact that the sample is biased and some subjects are never observed, as suggested initially by Böhning (2008) and extended to the GC scenario by Farcomeni (2017). The estimator's variance is the sum of two terms:

$$\mathrm{Var}(\hat{N}) = \mathrm{Var}(E[\hat{N}|\Delta_i]) + E[\mathrm{Var}(\hat{N}|\Delta_i)], \tag{5}$$

where Δ_i is an indicator of having used the i-th subject for estimation. 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 $\hat{p}_i = \exp(-\sum_{j=1}^{S}\hat{p}_{ij})(1+\sum_{j=1}^{S}\hat{p}_{ij}/2)\sum_{j=1}^{S}\hat{p}_{ij}$. Detailed derivation steps are given in Farcomeni (2017). The second addend shall be obtained through resampling of the observed data with replacement. For each bootstrap sample one can record the final population size estimate, and an unbiased estimate of the second term of (5) is given by the variance of the resampled estimates.

3.2 Dealing with measurement error

The strategy outlined in the previous section gives us a generalized Chao estimator with external information, but ignores the possibility of measurement error. To do so, we first briefly recall SIMEX procedure for homoscedastic measurement error. First, SIMulation is performed by fixing $0 = \lambda_1 < \lambda_2 < \cdots \lambda_k$ and generating *pseudo errors* from a zero-centered Gaussian random variable with variance $\lambda_j \sigma^2$, where σ^2 is the estimated or known measurement error variance, adding them to the covariate measured with error, and finally estimating the population size (e.g., as in the previous section). This gives perturbed population size estimates, say $\hat{N}_{jt}$. The operation is repeated T times to reduce the effects of random sampling, and consequently an averaged population size estimate $\hat{N}_j = 1/T \sum_t N_{jt}$ is associated with each λ_j. An appropriate model $\hat{N}_j = f(\lambda_j)$ is then fit, e.g., a quadratic model of the kind $E[\hat{N}_j] = \alpha + \beta_1 \lambda_j + \beta_2 \lambda_j^2$. Finally, it is straighforward to check that the denoised population size estimate corresponds to the case $\lambda_j = -1$, therefore measurement error correction is obtained by extrapolating the linear model to estimate $\hat{N} = \hat{\alpha} - \hat{\beta}_1 + \hat{\beta}_2$.

In our case we have more than one measurement method, and proceed to generalize SIMEX to this scenario. Let σ_j^2 denote the measurement error variance for the j-th measurement method. Suppose there are r measurement methods (in our case, $r = 3$). These variances are easily estimated from the replicated measurements. To implement our SIMEX method we fix $0 = \lambda_{1j} < \lambda_{2j} < \cdots < \lambda_{kj}$. Of course for simplicity we might use the same values for each measurement method. We then obtain the cartesian product of these values, in order to consider all possible r^k combinations. Our heteroscedastic SIMEX approach proceeds as follows:

- For each combination of values $(\lambda_{u_1 1}, \lambda_{u_2 2}, \ldots, \lambda_{u_r r})$ we generate pseudo errors $U_{itj} \sim N(0, \lambda_{u_j j} \sigma_j^2)$ and add these to the contaminated measurements W_{itj}. We thus obtain a simulated data set.
- For each simulated data set we estimate the population size, and repeat this operation B times for each combination $(\lambda_{u_1 1}, \lambda_{u_2 2}, \ldots, \lambda_{u_r r})$.
- At the end of the procedure, the mean value for the population size for each scenario is recorded, call it $\hat{N}_{u_1, u_2, \ldots, u_r}$. It is straightforward to see that the error in variables variance is $(1 + \lambda_{uj})\sigma_j^2$ for $j = 1, \ldots, r$. Hence if $\lambda_{uj} = -1$ we can *extrapolate* the estimate that would have been obtained without measurement error.
- To do so, we first estimate a polynomial regression model with interactions as

$$E[\hat{N}_{u_1, u_2, \ldots, u_r}] = \alpha + \sum_{j=1}^{r} \beta_j \lambda_{u_j j} + \sum_{j=1}^{r} \beta_{2j} \lambda_{u_j j}^2 + \sum_{j<h} \beta_{jh} \lambda_{u_j j} \lambda_{u_h h};$$

  and then extrapolate to the case $\lambda = (-1, -1, \cdots, -1)$.

For estimation of standard errors we use a modified SIMEX information criterion as follows. First, for each $(\lambda_{u_1 1}, \lambda_{u_2 2}, \ldots, \lambda_{u_r r})$, including obviously

the $(0, 0, 0, \ldots, 0)$ case, we estimate the estimator's variance as the average (over the B replicates) of each estimator's variance. This is done as described in the previous section. Let $s_{u_1 u_2 \ldots u_r}$ denote the standard error for a fixed λ combination, after averaging over the T replicates. Call $\bar{s}$ the average of the values obtained above. As suggested by Carrol et al (2006), the standard error of the SIMEX estimator can be found by extrapolation of the difference $s_{u_1 u_2 \ldots u_r} - \bar{s}$ to $\lambda = (-1, -1, \cdots, -1)$; where we use the same polynomial regression model with two-way interactions as above.

4 Simulations

In this section we illustrate the methodology proposed through a simulation study. In Subsection 4.1 we focus on the performance of the generalized Chao Estimator as external information, either precise or misspecified, is taken into account. In Subsection 4.2 we focus on further use of our proposed measurement error correction tool.

4.1 Using external information

We focus on use of the gender ratio constraint within the generalized Chao estimator frameowork. Our aim is explicitly that of showing the advantage in taking into account external information, even when this might not be exactly correct.

We obtain capture probabilities for each observation through the following logit parameterization:

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = X_{i0} + \alpha + X_{i1} + 2X_{i2} + 0.5X_{i3} \text{ for } j = 1, \ldots, S \qquad (6)$$

where, in equation (6), X_{i2} is a dichotomous variable representing the gender of the $i-th$ subject, X_{i1}, X_{i3}, X_{i0} are zero centered Gaussian random variables. It must be pointed out that X_{i0} is only aimed at generating unobserved heterogeneity, and thus it is ignored in any analysis (mimicking the situation of the presence of subject-specific unmeasured covariates). Finally, a fixed gender ratio c is obtained at population level by setting $\Pr(X_{i2} = 1) = c/(1 + c)$.

We obtain different simulation set ups by varying parameters as follows:

- Population size: $N = 1000, 2000$.
- Capture occasions: $S = 10, 20, 50$.
- Gender ratio: $c = 1.02,\ 1.2,\ 1.5$.
- Intercept of model (6): $\alpha = -2, -4$.

By combining all the possible parameters we obtaine 36 different scenarios and, for each scenario, we generated the data $B = 500$ times. Generated sampling fractions n/N oscillate between an *average* minimum of 0.4 (for $S = 10$ and $\alpha = -4$) to an *average* maximum of 0.8 (for the case $S = 50$ and

$\alpha = -2$). Obviously, the iteration-specific sampling fraction is random, with some variability, with a minimum very close to zero and a maximum above 95% in our iterations.

For each iteration within each scenario we generate data, record subjects observed at least once, and apply four times our constrained Chao estimator, the first time assuming a correct gender ratio and three times with a shifted (hence, wrong) gender ratio. We compare with the generalized unconstrained Chao estimator and the basic Chao estimator (which therefore ignores covariates). Tables 1, 2, 3 and 4 report the median and the MAD of the values of the different generalizations of Chao estimators within each scenario.

Table 1 Simulation study: median and MAD (in brackets) of the population size estimates when $N = 1000$, $\alpha = -2$. $\hat{N}_c$ denotes our constrained GC estimator with external information, $\hat{N}_{c+const}$ the same assuming an incorrect gender ratio $c + const$. $\hat{N}_{cov}$ is the unconstrained GC estimator and $\hat{N}_{basic}$ the classical Chao estimator. Results are based on 500 iterations.

N	S	True c	α	$\hat{N}_c$	$\hat{N}_{c+0.05}$	$\hat{N}_{c+0.1}$	$\hat{N}_{c+0.3}$	$\hat{N}_{cov}$	$\hat{N}_{basic}$
1000	10	1.02	-2	1009.8	985.5	962.9	890.2	754.5	701.8
				(40.2)	(38.5)	(37.5)	(32.8)	(36.3)	(22.3)
1000	20	1.02	-2	1007.5	982.7	960.5	888.8	832.5	788.9
				(36.2)	(35)	(33.5)	(31)	(40.7)	(28)
1000	50	1.02	-2	1003.7	979.9	958.4	892.3	905.9	881.1
				(30.8)	(29.3)	(28.4)	(27.3)	(31.2)	(22.7)
1000	10	1.2	-2	1005.6	986.6	969	911.6	773.4	726.6
				(35.2)	(34)	(33)	(30)	(36.5)	(22.1)
1000	20	1.2	-2	1007.2	987.8	969.8	910.7	846.5	805.4
				(35.2)	(33.9)	(33.2)	(30.8)	(37.4)	(24.2)
1000	50	1.2	-2	1001.6	983	965.8	910.6	914.1	892.8
				(30.2)	(29.5)	(29.1)	(26.4)	(29.8)	(22)
1000	10	1.5	-2	1004	990.5	978.1	935.3	798.6	753.7
				(35)	(34)	(33.5)	(31.4)	(34.6)	(19.6)
1000	20	1.5	-2	1008.9	995.4	982.4	938.1	861.6	826.9
				(31.4)	(31.1)	(30.6)	(29.4)	(36.4)	(23.6)
1000	50	1.5	-2	1003.3	989.9	977.4	934.9	924.7	903.5
				(29.6)	(29)	(28.6)	(26.7)	(27)	(21.2)

Results are clearly that, in most of the cases, use of external information can lead to a much lower bias, without increasing the standard error. Consequently, the constrained MSE can be substantially smaller than the unconstrained MSE. Not surprisingly, this holds true even if the external information is not precise, to a certain extent. This is particularly apparent for smaller sampling fractions (e.g., for the cases in which $S = 10$ in our tables), where rather precise estimates are obtained while unconstrained methods seem to underestimate by a large extent the true population size. This is particularly important as, unlike our motivating application, relatively small sampling fractions in practice are the rule rather than the exception. Finally, as noted also elsewhere, use of covariates might be beneficial also when not taking into account external information, as the unconstrained GC estimator $\hat{N}_{cov}$ seems to constantly dominate $\hat{N}_{basic}$ in terms of MSE.

Table 2 Simulation study: median and MAD (in brackets) of the population size estimates when $N = 1000$, $\alpha = -4$. $\hat{N}_c$ denotes our constrained GC estimator with external information, $\hat{N}_{c+const}$ the same assuming an incorrect gender ratio $c + const$. $\hat{N}_{cov}$ is the unconstrained GC estimator and $\hat{N}_{basic}$ the classical Chao estimator. Results are based on 500 iterations.

N	S	True c	α	$\hat{N}_c$	$\hat{N}_{c+0.0.5}$	$\hat{N}_{c+0.1}$	$\hat{N}_{c+0.3}$	$\hat{N}_{cov}$	$\hat{N}_{basic}$
1000	10	1.02	-4	849.3	829.9	812	753.9	558.2	473.8
				(57.2)	(55.9)	(54.8)	(50.4)	(69.9)	(27.2)
1000	20	1.02	-4	931.2	909.3	889.7	825.3	644.4	558.8
				(49.1)	(47.8)	(46.8)	(43.7)	(69.7)	(28.4)
1000	50	1.02	-4	986	962.2	940.5	870.6	748	670.9
				(49.7)	(47.9)	(46.4)	(41.2)	(57.9)	(29.1)
1000	10	1.2	-4	849.6	834.3	819.6	772.2	590.7	501.1
				(49.6)	(48.8)	(47.7)	(44.8)	(83.4)	(27.8)
1000	20	1.2	-4	930.4	913.5	897.8	845	667	586.3
				(43.7)	(42.8)	(41.8)	(39.6)	(62.3)	(26.7)
1000	50	1.2	-4	984	965.2	947.7	890.5	760.3	690.5
				(40.5)	(40)	(38.8)	(35.4)	(57.2)	(29)
1000	10	1.5	-4	844.9	833.9	823.8	788.6	628.7	538.7
				(45)	(44.4)	(43.8)	(41.7)	(89.5)	(27.2)
1000	20	1.5	-4	927.4	915.3	904.3	865.2	699	623.1
				(42.1)	(41.8)	(41.1)	(41.1)	(60)	(27.2)
1000	50	1.5	-4	981.7	968.5	956.3	914.7	788.8	722.2
				(36.2)	(35.9)	(35.5)	(34.4)	(53.4)	(27.2)

Table 3 Simulation study: median and MAD (in brackets) of the population size estimates when $N = 2000$, $\alpha = -2$. $\hat{N}_c$ denotes our constrained GC estimator with external information, $\hat{N}_{c+const}$ the same assuming an incorrect gender ratio $c + const$. $\hat{N}_{cov}$ is the unconstrained GC estimator and $\hat{N}_{basic}$ the classical Chao estimator. Results are based on 500 iterations.

N	S	True c	α	$\hat{N}_c$	$\hat{N}_{c+0.0.5}$	$\hat{N}_{c+0.1}$	$\hat{N}_{c+0.3}$	$\hat{N}_{cov}$	$\hat{N}_{basic}$
2000	10	1.02	-2	2014.7	1965.6	1922.3	1778.5	1502.6	1407.2
				(40.2)	(38.5)	(37.5)	(32.8)	(36.3)	(22.3)
2000	20	1.02	-2	2027	1976.7	1931.2	1785.7	1657	1582
				(36.2)	(35)	(33.5)	(31)	(40.7)	(28)
2000	50	1.02	-2	2011.9	1964.5	1921.1	1783	1806.3	1764.9
				(30.8)	(29.3)	(28.4)	(27.3)	(31.2)	(22.7)
2000	10	1.2	-2	2005.9	1968.1	1932.9	1816.3	1541.2	1449
				(35.2)	(34)	(33)	(30)	(36.5)	(22.1)
2000	20	1.2	-2	2018.5	1979.4	1943.4	1824.8	1677.6	1608.2
				(35.2)	(33.9)	(33.2)	(30.8)	(37.4)	(24.2)
2000	50	1.2	-2	2006.1	1968.3	1933.8	1820	1821.2	1781.2
				(30.2)	(29.5)	(29.1)	(26.4)	(29.8)	(22)
2000	10	1.5	-2	2000.8	1974.2	1949.2	1862.9	1586.2	1506.7
				(35)	(34)	(33.5)	(31.4)	(34.6)	(19.6)
2000	20	1.5	-2	2011	1983.2	1957.4	1870.2	1717.6	1650.2
				(31.4)	(31.1)	(30.6)	(29.4)	(36.4)	(23.6)
2000	50	1.5	-2	2003.8	1977.3	1952.5	1867.7	1844.7	1809.8
				(29.6)	(29)	(28.6)	(26.7)	(27)	(21.2)

4.2 Using external information further correcting for measurement error

We now study the performance of our SIMEX-corrected constrained estimator when the available covariates are subject to measurement errors. To do so,

Table 4 Simulation study: median and MAD (in brackets) of the population size estimates when $N = 2000$, $\alpha = -4$. $\hat{N}_c$ denotes our constrained GC estimator with external information, $\hat{N}_{c+const}$ the same assuming an incorrect gender ratio $c + const$. $\hat{N}_{cov}$ is the unconstrained GC estimator and $\hat{N}_{basic}$ the classical Chao estimator. Results are based on 500 iterations.

N	S	True c	α	$\hat{N}_c$	$\hat{N}_{c+0.0.5}$	$\hat{N}_{c+0.1}$	$\hat{N}_{c+0.3}$	$\hat{N}_{cov}$	$\hat{N}_{basic}$
2000	10	1.02	-4	1688.2	1648.7	1613.5	1497.5	1116	945.1
				(73.9)	(72.1)	(71)	(64.9)	(99.6)	(35.8)
2000	20	1.02	-4	1850.5	1807.5	1767.3	1640.2	1286.5	1119
				(72.5)	(70.3)	(69.1)	(63.8)	(96.3)	(43.1)
2000	50	1.02	-4	1969.2	1920.9	1878.4	1738.8	1483.9	1341.2
				(65.6)	(63)	(59.9)	(53.6)	(73.8)	(42.2)
2000	10	1.2	-4	1689.8	1659.4	1631	1536.4	1170.6	1004.6
				(69.1)	(68)	(66.8)	(62.6)	(114.2)	(35.2)
2000	20	1.2	-4	1846.3	1812.8	1781.5	1677.7	1327.1	1169.6
				(67.8)	(66.1)	(64.6)	(60.4)	(89.8)	(39.7)
2000	50	1.2	-4	1968.9	1931.2	1896.7	1781.5	1515.2	1382.5
				(53.8)	(53.4)	(52.4)	(47.9)	(72.5)	(35.2)
2000	10	1.5	-4	1687.8	1665.8	1645.2	1575.3	1225.5	1078.4
				(67.2)	(67.2)	(66.2)	(63.1)	(99.3)	(37.4)
2000	20	1.5	-4	1844.9	1821.1	1798.3	1721.4	1383	1244.4
				(56.1)	(55.4)	(54.5)	(52.3)	(86.6)	(37.7)
2000	50	1.5	-4	1962.2	1935.2	1910.1	1824.9	1563.2	1446.4
				(54.3)	(53.7)	(53.3)	(49.5)	(71.8)	(39.4)

we considered four different scenarios and run a brief simulation study. For each scenario we generate the data $B = 500$ times and compare the SIMEX constrained estimator (labeled as N_c^{SIMEX}) with the uncorrected constrained estimator (labeled as N_c as before). This will give a direct appraisal of the consequences of correcting for measurement error after constrained estimation.

To apply the SIMEX approach, we fixed a grid of values of $\lambda = 0.1, 0.5, 0.7, 1, 2$. For each combination of *pseudo errors* the procedure is repeated $T = 10$ times. We obtain capture probabilities for each observation through a logit parameterization given by

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = X_{i0} + \alpha + X_{i1} + 2X_{i2} + X_{i3} \text{ for } j = 1, \ldots, S. \qquad (7)$$

In equation (7), X_{i1} is a dichotomous variable representing the gender of the $i-th$ subject, while X_{i0}, X_{i2}, X_{i3} are zero centered Gaussian random variables which are subject to measurement error. Operatively speaking these have been contaminated by adding white noise to each observation. As in the previous case, X_{i0} is only aimed at generating unobserved heterogeneity and, as a consequence, is ignored in each analysis.

Table 5 clearly shows how correcting for measurement error further reduces bias when estimating the true population size. The advantage is clear even if we are partially taking into account unspecified sources of individual heterogeneity and the external information. This could be expected as measurement error, to the very list, is a source of unobserved heterogeneity. The advantages of the SIMEX correction are more and more evidence as either the capture

Table 5 Simulation study: median and MAD (in brackets) of the population size estimates when $N = 1000$, $\alpha = -2, -4$ and $S = 10, 20$. $\hat{N}_c^{SIMEX}$ denotes our constrained GC estimator after a correction for measurement error, $\hat{N}_c$ the constrained estimator. Results are based on 500 iterations.

N	S	True c	α	$\hat{N}_c^{SIMEX}$	$\hat{N}_c$
1000	10	1.02	-2	876.7	865.4
				(54.4)	(46.0)
1000	20	1.02	-2	921.6	912.3
				(46.6)	(41.3)
1000	10	1.02	-4	673.8	652.5
				(69.7)	61.0
1000	20	1.02	-4	753.0	738.0
				(65.1)	(55.8)

probability (represented by the parameter α in equation (7)) or the capture occasions S decrease.

5 Data Analysis

For comparison and assessment of the importance of using external information and correcting for measurement error, we work with the same data introduced in Farcomeni (2017), and we focus on similar estimators. The Maldives Whale Shark Research Programme (`http://mwsrp.org`) collected data for a period of six months in the area surrounding South Ari atoll in Maldives. Sightings of whale sharks, together with several additional information, were collected every day. A more detailed description of the motivation and rationale for the study (and for assuming that the population is closed within the observation period) is given in Farcomeni (2017). We note here that as often happens with studies on whale sharks at their aggregation points, surfacing sharks are mostly jouvenile males. Sharks were identified by matching the unique spots patterns on their skin, using photo images from possibly different perspectives. Skin spots are unique to each individual and they have been argued to not change during the lifetime of the animal. Length was assessed at each encounter with one or more of three methods. In fact at first length was simply eyeballed by an expert at sighting, and the guess was recorded. If possible, two researchers dove to swim with the whale shark and measure it with a tape. Both measurement methods are clearly only approximate. When available, a laser photogrammetry device was used by a researcher. Laser photogammetry involves using a camera with lasers at both sides, at a distance of 50 centimeters apart. The lasers project two green dots on to the flank of the shark, which are then captured in a photograph. The photograph is then processed in photoshop to establish the distance between the green dots. This measurement is then used to establish the distance between the 5th (most posterior) gill slit and the anterior edge of the first dorsal fin. A formula is finally used to determine the full length of the shark based on the gill to dorsal measurement. Full details can be found in Rohner et al (2011, 2015). Laser

Table 6 Population size estimate $\hat{N}$, standard errors for $\hat{N}$, assumed gender ratio (where NA means: No Assumption and corresponds to the estimate in Farcomeni (2017)) and RMSEA for the Whale Shark data set

$\hat{N}$	S.E.	Gender Ratio	RMSEA
222	61.84	NA	0.073
355	102.29	1.02:1	0.073
351	108.05	1.05:1	0.073
357	110.82	1.10:1	0.074
335	103.87	0.98:1	0.073

photogrammetry is definitely more precise than subjective judgement and use of tapes while swimming with the animal, but yet also prone to measurement error. Additionally, laser photogrammetry measures are available only for a subset of sightings. Other covariates used in Farcomeni (2017) were gender (with a gender ratio of approximately 10:1), cloud cover (complete/no), and a classical behavioural effect. A classical behavioural effect is an homogeneous modifier of the sighting probability after the first identification, and it can be included in the logistic parameterization simply via a dummy variable

$$X_{ij} = I\left(\sum_{t=1}^{j-1} Y_{it} > 0\right),$$

where $I(\cdot)$ denotes the indicator function. A total of $n = 112$ sharks was identified at least once during $S = 181$ sampling occasions. In Farcomeni (2017) a final estimate of $\hat{N} = 222$ with standard error 61.84 is reported.

We now use GC estimators based on the same covariates as Farcomeni (2017) and the same data set, but take also into account measurement error through our multivariate SIMEX procedure, and include external information on the gender ratio (which is reported as 1.02:1 in Chang et al (1997)).

First we ignore measurement error and report on use of external information. In Table 6 we report population size estimates and their standard error for a gender ratio of 1.02:1, and also perform a sensitivity analysis based on other values for the constraint. Note that whale sharks are not gender changing hence in accordance with Fisher's principle the gender ratio can not be too far from 1:1.

From Table 6 we can see that there is a huge effect of external information on the population size estimate. The very small number of females catched did not lead Farcomeni (2017) to discriminate between differential surfacing habits (and hence, catchability) of females and males with respect to true population imbalance. When information on gender ratio is included, the final population estimates are approximately 60% larger than the one reported in Farcomeni (2017). Also the standard errors are slightly larger, which could be expected in front of larger size estimates. In the table we also report goodness of fit as obtained through the RMSEA statistic. Fit is acceptable in all cases. Notably, same fit is obtained with and without constraint. This is not suprising as the constraint is acting basically only on the predictions for the missing cells.

The estimates in Table 6 are still prone to bias due to measurement error. We fix a gender ratio of 1.02:1 and use our heteroscedastic SIMEX approach to additionally correct for measurement error. The final estimate is $\hat{N} = 344$ (standard error: 123.91), slightly lower than the one reported in Table 6 in correspondence of the gender ratio 1.02:1. The biasing effect of measurement error is definitely present but not substantial. We finally note that in all cases very large standard errors are obtained, indicating that even if there were several sampling occasions a strong uncertainty is associated with the estimated size of this elusive population.

6 Conclusions

When one group has a strongly differential catchability with respect to another one, usually a logistic parameterization is sufficient in detecting this difference and providing unbiased population size estimators. We have argued in this paper that when a subgroup is very hard to detect, there might not be enough individuals for the model to be able to discriminate between unequal catchability and group unbalance at population level. External information, at least in terms of reasonable ranges for group size ratios, might be available and could be very useful. We have seen that a substantial decrease in MSE can be achieved with external information. The final population size estimates can be substantially different. In our original data application we have seen that including information about a gender ratio produces an increase of about 60% in $\hat{N}$, assuming ratios between 0.98 and 1.10. Other choices for gender ratios might contradict Fisher's principle, and hence not be plausible. The external information on gender ratio was indeed used to reduce bias due to the very scarce number of females observed during the survey, that is, to force the estimates take into account the fact that females are not observed because they do not surface, and not because there are only few in the population.

Another important issue is that of measurement error. In capture-recapture experiments animals might only be seen for a short time frame, or only photos of them might be available for covariate measurements. A consequence is that in several capture-recapture experiments measurement error might be present, and could bias the final estimates in unpredictable directions. The novelty of our approach is that it provides measurement error correction by using all available information even when different measurement methods have been used throughout the experiment. The SIMEX framework is particularly advantageous for use in conjunction with Chao and GC estimators since it allows us to use the entire data set for measurement error correction, before restricting to singletons and doubletons for population size estimation. As argued in Farcomeni (2017), the GC estimator guarantees a lower bound estimate *and* robustness againts unobserved heterogeneity.

We note that we could have followed other routes to tackle the two open problems discussed in this paper. For use of external information we could have exploited penalized likelihoods. This would have been computationally

more cumbersome, but with more slack for the user for a balance between adherence to the constraint and likelihood maximization. Other methods for measurement error correction are also available, including for instance regression calibration (RC). An advantage of RC is that simple expressions are available for correction and the final standard errors, therefore being less computationally intense than SIMEX. One of the limitations in our framework is that closed form expressions might not be available anymore.

Finally, one approach which can very naturally take into account measurement error and external information is the Bayesian framework. In this work we have not considered the Bayesian framework as we found very difficult to guarantee the same properties of Chao and GC estimators, that is, that a certain posterior summary could be used as a lower bound estimator for the population size. Extension of Chao and GC estimators to the Bayesian framework is definitely grounds for further work.

Acknowledgements The authors are 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 referees for kind comments and suggestions.

References

Bartolucci F, Forcina A (2006) A class of latent marginal models for capture–recapture data with continuous covariates. Journal of the American Statistical Association 101:786–794

Bartolucci F, Lupparelli M (2008) Focused information criterion for capture-recapture models for closed populations. Scandinavian Journal of Statistics 35:629–649

Böhning D (2008) A simple variance formula for population size estimators by conditioning. Statistical Methodology 5:410–423

Böhning D, Vidal-Diez A, Lerdsuwansri R, Viwatwongkasem C, Arnold M (2013) A generalization of Chao's estimator for covariate information. Biometrics 69:1033–1042

Carrol RJ, Ruppert D, Stefanski L (2006) Measurement error in nonlinear models. Chapman and Hall, London

Chang W, Leu M, Fang L (1997) Embryos of the whale shark, Rhincodon typus: Early growth and size distribution. Copeia 1997:444–446

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, Colwell RK (2017) Thirty years of progeny from Chao's inequality: estimating and comparing richness with incidence data and incomplete sampling. Statistics and Operations Research Transactions 41:3–54

Farcomeni A (2011) Recapture models under equality constraints for the conditional capture probabilities. Biometrika 98:237–242

Farcomeni A (2017) Fully general Chao and Zelterman estimators with application to a whale shark population. Journal of the Royal Statistical Society (Series C) p in press

Farcomeni A, Tardella L (2012) Identifiability and inferential issues in capture-recapture experiments with heterogeneous detection probabilities. Electronic Journal of Statistics 6:2602–2626

Gould W, Stefanski L, Pollock K (1999) Use of simulationextrapolation estimation in catcheffort analyses. Canadian Journal of Fisheries and Aquatic Sciences 56:1234–1240

Gustafson P (2003) Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments. Chapman and Hall/CRC Press

Huggins R, Hwang W (2010) A measurement error model for heterogeneous capture probabilities in markrecapture experiments: An estimating equation approach. Journal of Agricultural, Biological and Environmental Statistics 15:198–208

Hwang W, Huang S (2003) Estimation in capture-recapture models when covariates are subject to measurement errors. Biometrics 59:1113–1122

Keogh RH, White IR (2014) A toolkit for measurement error correction, with a focus on nutritional epidemiology. Statistics in medicine 33

King R, Bird SM, Brooks SP, Hutchinson SJ, Hay G (2005) Prior information in behavioral capture-recapture methods: demographic influences on drug injectors' propensity to be listed in data sources and their drug-related mortality. American Journal of epidemiology 162:694–703

King R, Bird SM, Overstall AM, Hay G, Hutchinson SJ (2014) Estimating prevalence of injecting drug users and associated heroin-related death rates in england by using regional data and incorporating prior information. Journal of the Royal Statistical Society: Series A (Statistics in Society) 177:209–236

McCrea RS, Morgan BJT (2014) Analysis of capture-recapture data. CRC Press

Meekan M, Bradshaw C, Press M, McLean C, Richards A, Quasnichka S, Taylor J (2006) Population size and structure of whale sharks (rhincodon typus) at ningaloo reef western australia. Marine Ecology Progress Series

Rohner CA, Richardson AJ, Marshall AD, Weeks SJ, Pierce SJ (2011) How large is the world's largest fish? Measuring whale sharks *rhincodon typus* with laser photogrammetry. Journal of Fish Biology 78:378–385

Rohner CA, Richardson AJ, Prebble CE, Marshall AD, Bennett MB, Weeks SJ, Cliff G, Wintner SP, Pierce SJ (2015) Laser photogrammetry improves size and demographic estimates for whale sharks. PeerJ 3

Stoklosa J, Dann P, Huggins RM, Hwang WH (2016) Estimation of survival and capture probabilities in open population capture–recapture models when covariates are subject to measurement error. Computational Statistics & Data Analysis

White I, Frost C, Tokunaga S (2001) Correcting for measurement error in binary and continuous variables using replicates. Statistics in Medicine

20:3441–3457

Francesco Dotto is Research Fellow at the Department of Statistical Sciences of Sapienza - University of Rome. He has recently worked on robust statistics and, more specifically, robust model based clustering methods. He has interests in modeling and statistical applications for ecology and socio-economic sciences.

Alessio Farcomeni is Associate Professor at the Department of Public Health and Infectious Diseases of Sapienza - University of Rome. His research interests and main contributions focus on population size estimation, longitudinal data analysis, multiple testing, and robust statistics; with applications to ecology, biomedicine, and socio-economic sciences.