



# A Generalized Estimating Equations Approach to Model Heterogeneity and Time Dependence in Capture-Recapture Studies

Md. Abdus Salam Akanda<sup>1</sup>, Russell Alpizar-Jara<sup>2</sup>

<sup>1</sup>Department of Statistics, University of Dhaka, Dhaka 1000, Bangladesh  
Corresponding author.  
E-mail: [akanda@du.ac.bd](mailto:akanda@du.ac.bd)


<sup>2</sup>Research Center in Mathematics and Applications, CIM-UE., Department of Mathematics, University of Évora, 7000-671 Évora, Portugal

## ABSTRACT

Individual heterogeneity in capture probabilities and time dependence are fundamentally important for estimating the closed animal population parameters in capture-recapture studies. A generalized estimating equations (GEE) approach accounts for linear correlation among capture-recapture occasions, and individual heterogeneity in capture probabilities in a closed population capture-recapture individual heterogeneity and time variation model. The estimated capture probabilities are used to estimate animal population parameters. Two real data sets are used for illustrative purposes. A simulation study is carried out to assess the performance of the GEE estimator. A Quasi-Likelihood Information Criterion (QIC) is applied for the selection of the best fitting model. This approach performs well when the estimated population parameters depend on the individual heterogeneity and the nature of linear correlation among capture-recapture occasions.

## KEYWORDS

capture-recapture, heterogeneity, generalized estimating equations, quasi-likelihood information criterion, population parameters

 © 2017 Md. Abdus Salam Akanda, Russell Alpizar-Jara

This is an open access article distributed under the Creative Commons Attribution-NonCommercial-NoDerivs license

## INTRODUCTION

Capture-recapture methods are extensively used in animal population estimation and in other fields such as quality control and epidemiology (Briand et al. 1997; Chao et al. 2001). A set of models and their inference procedures have been proposed in the capture-recapture literature for the estimation of animal population size from capture-recapture data (Seber 2002; Williams et al. 2002; Chao & Huggins 2005). These models have been mainly classified into two groups: closed and open populations, but this work focuses on closed population models. The closed population models arise when the population is assumed to be constant during the period of study, such that immigration, emigration, births, and deaths remain fixed. The general closed population model in capture-recapture studies (Otis et al. 1978) is denoted by  $M_{tth}$ , where ( $t$ ) is used to represent time effect, ( $b$ ) behavioural response, and ( $h$ ) individual inherent heterogeneity to capture. This paper deals with estimating the closed population size using a sub-model

of the type  $M_{th}$ , where individual heterogeneity and time effect are considered; there is no behavioral response to the capture and the capture probabilities depend on covariates. There are various advantages of models incorporating covariates such as: (i) the models provide a clear explanation of the sources of heterogeneity, and each covariate effect can be assessed; and (ii) if all relevant covariates are included, then these models generally yield better estimators with respect to bias and precision (Chao & Huggins 2005).

A broad variety of approaches have been considered when fitting capture-recapture closed population model  $M_{th}$ , including the sample coverage models (Chao et al. 1992), martingale methods (Lloyd & Yip 1991), latent class and log-linear models (Agresti 1994), the use of individual covariates in generalized linear models (GLM) (Huggins 1989), finite mixture models (Pledger 2000), and robust P-spline approach (Stoklosa & Huggins 2012). Pollock et al. (1984) proposed an estimation procedure to cope with individual heterogeneity,

modelling capture probabilities based on individual categorical covariates, such as age group and sex. Huggins (1989, 1991) extended the case to continuous covariates, developing a conditional likelihood model in terms of observable characteristics of the capture individuals and assuming independence among the capture occasions. Moreover, the time effects modelled as a function or as factors of covariates, such as the recorded air temperature on the capture-recapture occasion, can also be measured. For example, Stoklosa & Huggins (2012) and King & Brooks (2008) have addressed time dependence in capture probabilities using P-splines and Bayesian inference, respectively, through environmental (time-dependent) covariates only. The use of estimating equations to model individual heterogeneity has also been discussed recently. For example, Zhang (2012) and Hwang & Huggins (2005) examined the individual heterogeneity effects on the animal population estimation using capture-recapture closed population model  $M_h$  by not only solving the estimating equations, but by also assuming independence of the capture-recapture occasions. Akanda & Alpizar-Jara (2014a) proposed a generalized estimating equations (GEE) approach which accounts for individual heterogeneity and dependency among capture occasions, but their approach mainly focused on the  $M_{ph}$  model. They also showed that the performance of the GEE approach is better than the mixed effects approach considering the closed population capture-recapture model,  $M_h$  (Akanda & Alpizar-Jara 2014b).

Capture-recapture data may be correlated over capture time. The estimators in capture-recapture studies may be biased, failing to account for this correlation. Some sort of dependencies among capture-recapture occasions can be considered in the capture-recapture literature through the behavioural effect's model, such as trap happiness and trap shyness (Yang & Chao 2005; Pradel & Sanz-Aguilar 2012). Here a closed population capture-recapture model of the type  $M_{ih}$  is built that allows the modelling of individual and environmental characteristics. The performance of the GEE approach (Liang & Zeger 1986) in the capture-recapture closed population model,  $M_{ih}$ , is assessed through a simulation study, and the estimated population parameters for two different real data sets. A logit-link function is assumed to model capture probabilities, considering the log of odds (the ratio of the probability of an event capturing to the probability of not capturing) as a linear function of the explanatory covariates, as in Huggins (1989). The GEE approach uses individual observed characteristics to model individual heterogeneity, environmental characteristics to model time variation, and also accounts for linear correlation among capture-recapture occasions. A quasi-likelihood procedure is used to estimate the regression parameters related with the animal population size and capture probabilities.

The models with notations considered in this work are presented in the next section. Section 3 illustrates the methodology with two real data sets. The model section procedures are described in Section 4. A simulation study in Section 5 is presented to assess the performance of the GEE approach

in the closed population  $M_{ih}$  model. Finally, some concluding remarks are given in Section 6.

## 1. NOTATIONS AND MODELS

Suppose that the total number of individuals in a capture-recapture experiment is  $N$  over the capture-recapture occasions  $j = 1, 2, \dots, m$ . Let  $Y_{ij}$  be the indicator variable, considering 1 if the  $i$ th individual is trapped on the  $j$ th capture occasion and 0 otherwise. Let  $\tau_i = \sum_{j=1}^m Y_{ij}$  be the number of times where the  $i$ th individual has been trapped in the capture-recapture closed population study. Individual observable covariate  $x_i$  for the  $i$ th individual (for example, sex, age, weight, body length, etc.), and observable environmental covariate  $z_j$ , that only depends on the  $j$ th capture occasion (such as air temperature, humidity, rainfall, etc.) are considered. Suppose the probability ( $P_{ij}$ ) of the  $i$ th individual is trapped on the  $j$ th capture occasion is,

$$P_{ij} = \mu_{ij} = \Pr(Y_{ij} = 1 | x_i, z_j) = h(\beta_0 + \beta_1 x_i + \beta_2 z_j) \quad (1)$$

for  $i = 1, 2, \dots, N; j = 1, 2, \dots, m$  where  $h(u) = \{1 + \exp(-u)\}^{-1}$  is the logistic function and

$$X_i = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_i & x_i & \dots & x_i \\ z_1 & z_2 & \dots & z_m \end{bmatrix}'$$

is the design matrix. The design matrix  $X_i$  can be generalized to construct various closed population models in the capture-recapture studies (Akanda & Alpizar-Jara 2014a). The model (1) is a restricted model  $M_{ih}$  of Huggins (1991) but not equivalent to any models of Otis et al. (1978). In  $M_{ih}$  type of model, individual and time variation is explained by an individual covariate  $x_i$  and an environmental covariate  $z_j$  respectively. The probability that the  $i$ th individual does not capture on the  $j$ th capture occasion is  $1 - P_{ij} = (1 - \mu_{ij})$ , and  $P_{ij}(1 - P_{ij}) = \mu_{ij}(1 - \mu_{ij})$  is the variance of  $Y_{ij}$ . Let  $V_i = A_i^{1/2} R_i(\alpha) A_i^{1/2}$  where  $V_i$  is a variance-covariance matrix and  $A_i$  is a diagonal matrix of  $\text{Var}(Y_{i1}), \text{Var}(Y_{i2}), \dots, \text{Var}(Y_{im})$  of order  $m \times m$ .  $R_i(\alpha)$  is a correlation structure among  $Y_{i1}, Y_{i2}, \dots, Y_{im}$  to explain the average dependence of individuals being captured from one occasion to another occasion, where  $\alpha$  is the intraclass correlation coefficient,  $|\alpha| < 1$ . In the case of a capture-recapture experiment,  $\alpha$  is the correlation between two consecutive capture occasions. The identity matrix in the variance function of a generalized linear models (GLM) assumes independence and this is substituted in a GEE with a correlation matrix,  $R_i(\alpha)$  (Hardin & Hilbe 2013). Therefore, the GEE approach takes into account the dependence among the observations by specifying the correlation structure. This structure is used to estimate the covariance matrix (Zeger & Liang 1986). A GEE approach allows various types of correlation structures  $R_i(\alpha)$  and as a property, this approach provides unbiased estimates in analysing the correlated binary data (Diggle et al. 2013). Some common specifications for  $\text{corr}(Y_j)$  are as follows:

- Independence correlation structure: This correlation structure assumes that all pairwise correlation coefficients are

zero, that is,  $corr(Y_{ij}, Y_{ik}) = 0; j \neq k$ ; thus,  $R_i(\alpha) = I$ , where  $I$  is an identity matrix of order  $m \times m$ . Correlation coefficient is assumed to be zero, hence no estimate of  $\alpha$  is obtained.

- Exchangeable correlation structure: This structure assumes that all pairwise coefficients of correlation are equal, that is,  $corr(Y_{ij}, Y_{ik}) = \alpha; j \neq k$ .
- Autoregressive correlation structure: This structure assumes that the coefficients of correlation decay exponentially over capture time, that is,  $corr(Y_{ij}, Y_{ik}) = \alpha^{|j-k|}; j \neq k$ .
- Unstructured or pairwise correlation structure: This correlation structure assumes that all pairwise coefficients of correlation are not same, that is,  $corr(Y_{ij}, Y_{ik}) = \alpha_{jk}; j \neq k$ .

Let  $D_i$  be the matrix of derivatives  $\partial \mu_i / \partial \beta'$ , where,  $\mu_i = (\mu_{i1}, \mu_{i2}, \dots, \mu_{im})$ ; hence,  $D_i = A_i X_i$ . If  $x_i$  and  $z_j$  were observable for each individual in the population, the vector of parameters  $\beta = (\beta_0, \beta_1, \beta_2)'$  for the model (1) can be obtained by solving the generalized estimating equations as follows:

$$U(\beta) = \sum_{i=1}^N D_i' V_i^{-1} (Y_i - \mu_i) = 0 \quad (2)$$

However, the ultimate purpose is to estimate  $N$ , the total number of unknown individuals in the population. Also, the covariates  $x_i$  and  $z_j$  are unknown for the individuals that are not captured in any capture occasions. Let  $i = 1, 2, \dots, n$ , be a set of captured separate individuals at least once and  $i = n+1, \dots, N$  be a set of uncaptured individuals in the capture-recapture occasions. Thus,  $Y_{ij}$  is considered under the captured individuals ( $n$ ) (i.e.,  $T_{ij} \geq 1$ ) with their corresponding individual covariates as detailed by Zhang (2012) and Huggins (1989). Therefore, the vector of parameters  $\beta$  of the model (1) can be estimated by solving the estimating equations, which is known as generalized estimating equations (GEE) in capture-recapture studies (Akanda & Alpizar-Jara 2014a):

$$U(\beta) = \sum_{i=1}^n D_i' V_i^{-1} (Y_i - \mu_i) = 0 \quad (3)$$

The generalized estimating equations (3) is fitted by computing an initial estimate of the covariance matrix ( $V_i$ ) first and the vector of regression coefficients ( $\beta$ ) applying an ordinary generalized linear model. The working correlation matrix is then updated based on these regression parameters and the variance-covariance matrix is recalculated. The vector of estimated coefficients is updated, and until convergence, these steps are repeated (Zeger & Liang 1986). At the convergence process, the coefficients of regression are consistent and offer valid standard errors even though there is misspecification of the correlation structure (Zeger & Liang 1986). Let  $\hat{\beta}$  be the resulting estimator of  $\beta$  and let  $\hat{P}_{ij} = h(X_i, \hat{\beta})$ . Then, the population size can be estimated following the method of Huggins (1989) that is based on the Horvitz and Thompson estimator (1952). The population size  $N$  is estimated by  $\hat{N}_{th} = \sum_{i=1}^n \hat{\pi}_i(\beta)^{-1}$ , where,  $\hat{\pi}_i(\beta) = 1 - \prod_{j=1}^m (1 - \hat{p}_{ij})$  is the probability of being trapped at least once given the individual covariates. The variance of  $\hat{N}_{th}$  can be estimated by  $\widehat{var}(\hat{N}_{th}) = \sum_{i=1}^n \pi_i(\beta)^{-2} (1 - \pi_i(\beta)) + \Delta(\beta)' \Gamma(\beta) \Delta(\beta)$ , where  $\Gamma(\beta)$  is a conditional information matrix and  $\Delta(\beta) = \sum_{i=1}^n \pi_i(\beta)^{-2} \partial \pi_i(\beta) / \partial \beta$  is a vector

with all quantities evaluated at  $\hat{\beta}$ .

## 2. ILLUSTRATIVE EXAMPLES

### 2.1. Example: Deer mice data

The first example concerns the captures of deer mice (*Peromyscus maniculatus*). V. Reid collected the data set at East Stuart Gulch Colorado associated with covariates age, sex, and weight (in grams). A rectangular grid of 9x11 traps was used, with 50-foot (15.2-m) trap spacing. The data are well known and have been analyzed in numerous capture-recapture literature (Otis et al. 1978; Huggins 1991; Huggins & Yip 1997; Stanley & Richards 2005). The data set consists of  $n = 38$  distinct deer mice. There are 17 female and 21 male deer mice, of which there are 11 adults, 3 semi-adults and 24 young ones. The semi-adults are recorded as adults in this analysis. The numbers of deer mice caught for  $m = 6$  occasions ( $n_1$  to  $n_6$ ) are 15, 20, 16, 19, 25, 25 and  $\sum_{n_k} = 120$ . The recorded capture frequencies ( $f_j$  to  $f_6$ ) are 9, 6, 7, 6, 6 and 4. The average capture frequencies for females and males are 3.41 and 2.76, respectively, and for the young ones and adults are 3.54 and 2.50, respectively. The average weight is 14.53 grams and the sample standard deviation is 4.84. This data is used to apply the GEE approach in the capture-recapture model, using covariates. The following equation is applicable to the GEE approach for this data,

$$\ln \left( \frac{P_{ij}}{1 - P_{ij}} \right) = \beta_0 + \beta_{age} \times age_{(i)} + \beta_{sex} \times sex_{(i)} + \beta_{wt} \times weight_{(i)} + \beta_t \times time_{(j)} \quad (4)$$

$i = 1, 2, \dots, n; j = 1, 2, \dots, m$ , where,  $\beta_{age}$ ,  $\beta_{sex}$ ,  $\beta_{wt}$  and  $\beta_t$  denote the age, sex, weight and time effect respectively. Model parameters are estimated assuming various correlation structures among capture-recapture occasions like independence, autoregressive, exchangeable, and pairwise correlation structures. The parameter estimation results are reported in Table 1 applying the GEE approach in the capture-recapture closed population model  $M_{th}$ . The parameter estimation is carried out using the R package (R Development Core Team 2016). Odds ratio (OR) describes the strength of dependence or association between categorical variables. The odds ratio for continuous explanatory covariates indicates the effect of changes of one-unit in the explanatory covariate.

The odds ratios, in these results, indicate that an individual in the young age group is more likely to be trapped than the individual in the adult age group for any given working correlation structure, keeping all other covariates constant. The odds of trapping the individual in the young age group are about (1/0.171) 5.85 times higher than the individual in the adult age group, but (1/0.146) 6.85 times higher for pairwise correlation structure model. The capture probability of males is significantly higher than the capture probability of females. For example, the capture probability of males is about double the capture probability of females. The odds of trapping are increased by 15% for one gram increase of weight, but 18% for pairwise correlation structure. The probability of trapping

Table 1. GEE estimates for the deer mice capture-recapture data under various working correlation structures

Cov.	Independent correlation				Exchangeable correlation			
	Coff.	Std. Err.	P-value	O.R.	Coff.	Std. Err.	P-value	O.R.
age	-1.767	0.531	0.001	0.171	-1.775	0.654	0.007	0.170
sex	0.753	0.291	0.010	2.124	0.759	0.358	0.034	2.136
weight	0.143	0.053	0.007	1.154	0.144	0.065	0.028	1.154
time	0.233	0.085	0.006	1.262	0.233	0.080	0.004	1.262
cons.	-2.532	0.722	0.000	...	-2.539	0.853	0.003	...
	Autoregressive correlation				Pairwise correlation			
	Coff.	Std. Err.	P-value	O.R.	Coff.	Std. Err.	P-value	O.R.
age	-1.760	0.551	0.001	0.172	-1.926	0.628	0.002	0.146
sex	0.761	0.302	0.012	2.140	0.632	0.341	0.064	1.881
weight	0.142	0.055	0.010	1.153	0.164	0.062	0.009	1.178
time	0.233	0.087	0.007	1.262	0.220	0.070	0.002	1.246
cons.	-2.530	0.748	0.001	....	-2.625	0.812	0.001	...

increases when changing from one occasion to another occasion. There is a 26% increase in the risk of trapping for changing from one occasion to another occasion, but slightly lower for pairwise correlation structure model. This finding may suggest that deer mice are trap happy, as the probability of trapping increases from one occasion to another occasion.

**2.2. Example: House mice data**

The second example concerns the captures of house mice (*Mus musculus*). The house mice data are originally collected by Cou-lombe and are described and analysed in Otis et al. (1978). The two covariates sex (female or male) and age (juvenile, semi-adult or adult) are related to this data. Complete capture information is given in program CAPTURE (Rexstad & Burnham 1991). For this data set, a total of 173 individuals are captured. Two records in the analysis are excluded because the covariates for the two mice are missing. Therefore, this analysis consists of  $n = 171$  distinct house mice that are captured at least once. Juveniles and semi-adults are grouped together into a ‘young’ class because there are only 8 juveniles. There are 77 non-adults (45 males, 32 females) and 94 adults (41 males, and 53 females). The numbers of house mice catch for  $m = 10$  occasions ( $n_1$  to  $n_{10}$ ) were 68, 60, 62, 52, 73, 41, 76, 35, 76 and 38, and  $\sum n_k = 581$ . The recorded capture frequencies ( $f_1$  to  $f_{10}$ ) are 2, 62, 40, 31, 16, 13, 5, 1, 0 and 1. On average, the capture frequencies for females and males are 3.72 and 3.08, respectively, and for adults and non-adults are 3.81 and 2.90, respectively. The methods of Otis et al. (1978) show the way to select the model, where the capture probabilities are not homogeneous and depend on capture time. Therefore, the  $M_{th}$  model is selected for this data to apply the GEE approach in capture-recapture studies. According to the model formulation of  $M_{th}$ , the following equation is useful for the available data,

$$\ln\left(\frac{P_{ij}}{1-P_{ij}}\right) = \beta_0 + \beta_{age} \times age_{(i)} + \beta_{sex} \times sex_{(i)} + \beta_t \times time_{(j)} \quad (5)$$

$i = 1, 2, \dots, n; j = 1, 2, \dots, m$ , where,  $\beta_{age}$ ,  $\beta_{sex}$  and  $\beta_t$  denote the age, sex, and time effect respectively. The estimation results for this data using the GEE approach are summarized in Table 2. Note that the age, sex and time are all statistically significant covariates at 5% level of significance (P-values are mentioned in Table 2).

The odds ratios indicate that the adult age group individuals are more likely to be trapped than the young age group individuals for any given correlation structure, keeping all other covariates constant. The odds of trapping increase 46% for the individual in adult age group than the individual in young age group, but 45% increase for autoregressive and pairwise correlation structures. The capture probability of males is significantly lower than for females. According to the odds ratios, the odds of capturing females are about (1/.793) 1.26 times higher than those of males. The risk of trapping decreases for changing one occasion on another occasion that is the probabilities of capture depend on time which supports the findings of Otis et al. (1978). This finding may also suggest that house mice are trap shy.

**3. MODEL SELECTION**

Time correlation plays an important role when analysing data sets. Hence, a correlation structure that builds up the most parsimonious model in GEE analysis needs to be considered. As a property, the GEE approach provides unbiased estimates. Hence, one may select the best fitting model by observing the relative efficiency of the estimated model coefficients. Quasi-likelihood information criterion (QIC) is also applicable for the selection of best fitting model when GEE approach is used in

capture-recapture studies (Akanda 2014a). The QIC is a modified version of the usual Akaike's information criteria (AIC), which allows comparisons of GEE models and selection of a working correlation structure (Pan 2001). The most parsimonious model and best correlation structure are selected based on the smallest value of QIC. The estimation results show that the standard errors of various estimated parameters are dissimilar. Table 3 compares the relative efficiencies of the estimated parameters for several models with respect to the parameters estimated by using independence working correlation structure model and QIC for several models for the applied data sets.

For the deer mice data, it is revealed that most of the estimates obtained under the independence correlation structure model are more efficient as compared to the other estimates except for capture time. The relative efficiency of capture time (under independent correlation structure) is 1.063 in exchangeable correlation structure model and 1.214 in pair-

wise correlation structure model. The QIC suggests that pairwise correlation structure model has the lowest QIC (297.81), and thus, is chosen as the best fitting model for this data set. Under the pairwise correlation structure, the estimated population is 39.17 with standard error 1.13. For the house mice data, all the covariates are more efficient under the autoregressive correlation structure as per the relative efficiencies. The model selection criterion QIC also suggests that autoregressive correlation structure model may be the best choice for this data set. The estimated population size is 175.08 with standard error 2.07 considering the autoregressive correlation structure. Huggins (1989) examined this data set and modelled the individual heterogeneity as a function of the age category and sex of the individuals. He estimated the population size 176.9 with standard error 2.01. The continuous-time sample coverage method for model  $M_{th}$  (Chao & Lee 1993) yields a population size estimate of 172 with an estimated standard error 3.3. All these

Table 2. GEE estimates for the house mice capture-recapture data under various working correlation structures

Cov.	Independent correlation				Exchangeable correlation			
	Coff.	Std. Err.	P-value	O.R.	Coff.	Std. Err.	P-value	O.R.
age	0.379	0.105	0.000	1.461	0.379	0.102	0.000	1.461
sex	-0.232	0.104	0.026	0.793	-0.232	0.100	0.021	0.793
time	-0.041	0.018	0.024	0.960	-0.041	0.018	0.024	0.960
cons.	-0.545	0.138	0.000	...	-0.544	0.136	0.000	...
Cov.	Autoregressive correlation				Pairwise correlation			
	Coff.	Std. Err.	P-value	O.R.	Coff.	Std. Err.	P-value	O.R.
age	0.379	0.099	0.000	1.454	0.375	0.101	0.000	1.454
sex	-0.230	0.098	0.019	0.794	-0.223	0.099	0.025	0.801
time	-0.039	0.017	0.024	0.962	-0.044	0.018	0.016	0.957
cons.	-0.551	0.131	0.000	....	-0.529	0.139	0.000	...

Table 3. Quasi-likelihood information criterion (QIC) under various correlation structure models and relative efficiencies of the estimated coefficients under independence correlation structure model

Cov.	Independence	Exchangeable	Autoregressive	Pairwise
Deer mice data				
QIC	298.75	298.73	298.85	297.81
age	1.000	0.812	0.964	0.846
sex	1.000	0.813	0.964	0.853
weight	1.000	0.815	0.964	0.855
time	1.000	1.063	0.977	1.124
House mice data				
QIC	2173.59	2173.59	2172.61	2173.49
age	1.000	1.029	1.061	1.040
sex	1.000	1.040	1.061	1.051
time	1.000	1.000	1.059	1.000

estimation procedures imply that almost none or only a few individuals were missed in the capture-recapture experiment. Moreover, the GEE estimation results of the two examples agree with the results of Otis et al. (1978), but the proposed approach accounts for time dependence in addition to the heterogeneous capture probabilities.

#### 4. SIMULATION STUDY

A Monte Carlo (MC) simulation study is conducted in capture-recapture closed population  $M_{th}$  model to assess the performance of the GEE approach. The simulation program is written in R program (R Development Core Team, 2016) and the analyses are run on Intel(R) Core(TM) i5-3320M CPU computer. The number of Monte Carlo replicates ( $B = 1000$ ), population size ( $N = 100, 200, \text{ and } 500$ ), mean capture probability ( $p = 0.3 \text{ and } 0.5$ ), number of capture occasions ( $m = 6 \text{ and } 10$ ), and coefficient of correlation ( $\alpha = -0.5, -0.3, -0.1, 0, 0.1, 0.3, 0.5$ ) are used as input factors for the autoregressive correlation structure in the simulation study. A correlated capture history  $Y_{ij}$  is generated applying the proposed method of Qaqish (2003) and considering an autoregressive correlation structure. The individual's captured probabilities depended on the sex and weight, and also allowed for an environmental covariate for each occasion. The simulated individuals are assigned their sex with probability 0.5 and the weights are normally distributed with mean 15 and variance 4. The normal environmental covariates with mean 2 and variance 1 are used. Estimator performance is assessed based on the root mean square error ( $RMSE = \sqrt{\text{var}(\hat{N}) + \text{bias}^2}$ ). The main simulation results are presented in Table 4 and Table 5.

The performance of the GEE estimator for estimating population size ( $\hat{N}$ ) is good in capture-recapture studies, when there is no linear correlation ( $\alpha = 0$ ) among capture-recapture occasions and for the high average capture probability 0.5 ( $p = 0.5$ ). In such cases, this estimator produces low standard error, the absolute value of PRB, the coefficient of variation and RMSE. The performance of GEE estimator is poor and it is difficult to obtain reliable estimates when the average capture probability is low ( $p = 0.3$ ). The estimated size of population with its standard error depend on the number of capture-recapture occasions ( $m$ ), average capture probability ( $p$ ) and linear correlation ( $\alpha$ ) among capture-recapture occasions. For a fixed capture occasion and average capture probability, this estimator estimate higher population size with a lower standard error for negative linear correlation in comparison for an equal strength of positive linear correlation. The simulation results also show that the GEE approach underestimate the population size for positive linear correlation, and overestimate for negative linear correlation among capture-recapture occasions at the higher average capture probability ( $p = 0.5$ ). For a fixed capture-recapture occasion and average capture probability, the estimated population size and linear correlation among capture-recapture occasions are inversely related. The performance of estimators for  $m = 10$  capture-recapture occasions is superior to the per-

formance for capture-recapture occasions producing lower CV, RMSE and the absolute value of PRB. In general, the simulation results evidently show that the performance of the GEE estimator to estimate the size of population with its standard error varies on the number of capture-recapture occasions, average capture probability, and linear correlation among capture-recapture occasions.

#### 5. CONCLUSION

Individual heterogeneity in capture probabilities and time dependence are fundamentally important for estimating the population parameters (such as capture probability, population size, etc.) in capture-recapture studies. In studies of this type, the strength of linear correlation among responses is unknown, and many statistical techniques are used in the capture-recapture studies assuming independence among responses, while ignoring time correlation. The GEE approach plays an important role in analysing correlated capture-recapture data to get unbiased estimates. Within-cluster correlations are also used in this approach to increase estimator's efficiency allowing the repeated measures information. In this article, the GEE approach has been evaluated for adjusting the capture probabilities of a heterogeneous population and accounting for correlation structures among capture occasions. The estimation results and model selection criteria (QIC) show that correlation structure depends on data, and hence, the best model should be selected on the basis of considering various possible correlation structures. The GEE estimator performs well for the high capture probabilities, but the estimates seem to be unreliable for low capture probabilities. Simulation study also shows that the estimated population parameters vary on the number of capture-recapture occasions, average capture probability and the nature of linear correlation among capture-recapture occasions. The suggestion of Hwang and Huggins (2005), that estimators dealing with heterogeneity in capture probabilities should be considered, is in agreement with these analyses. Moreover, the correlation structure among the capture-recapture occasions should also be considered. Therefore, the results presented in these analyses underpin the importance of considering the correlation among capture occasions and heterogeneity in capture probabilities in capture-recapture studies. This approach may be extended to open population models to estimate the sampling animal population parameters in capture-recapture methodology.

**Acknowledgement:** The authors are greatly indebted to the Erasmus Mundus Mobility with Asia (EMMA), EU and FCT, Portugal for funding this research. The authors would like to thank XXI Congresso Anual da Sociedade Portuguesa de Estatística (SPE). The authors are very grateful to the anonymous referees for their careful reading of the manuscript and several suggestions that considerably improved the presentation. The authors have also declared no conflict of interest.

Table 4. Simulation results (1000 repetitions) of  $m = 6$  capture occasions considering autoregressive correlation structure for the GEE approach. Averages of the numbers of captured individuals, ( $\bar{n}$ ); the estimated population size, ( $AVE(\hat{N})$ ); standard errors of the estimated population size, ( $SE(\hat{N})$ ); nominal-based 95% confidence intervals (CI), i.e.,  $\hat{E}(\hat{N}) \pm 1.96 \times SE(\hat{N})$ ; percentage relative bias, ( $PRB = 100 \times \frac{E(\hat{N}) - N}{N}$ ), where  $E(\hat{N})$  is estimated by  $AVE(\hat{N})$ ; percentage coefficient of variation ( $CV = 100 \times \frac{SE(\hat{N})}{E(\hat{N})}$ ), and root mean square error, ( $RMSE = \sqrt{Var(\hat{N}) + Bias^2}$ ).

$\bar{p}$	$\alpha$	N	$\bar{n}$	$AVE(\hat{N})$	$SE(\hat{N})$	95%CI	PRB	CV	RMSE
0.30	-0.3	100	94.1	104.92	3.07	98.91-110.92	4.92	3.31	5.79
0.30	-0.1	100	90.7	99.88	3.39	93.23-106.53	-0.12	3.18	3.40
0.30	0.0	100	88.6	96.92	3.73	89.61-104.23	-3.08	3.12	4.84
0.30	0.1	100	85.9	93.30	4.18	85.10-101.49	-6.71	3.03	7.90
0.30	0.3	100	79.2	84.24	4.44	75.54-92.94	-15.76	2.74	16.37
0.30	0.5	100	70.6	73.37	4.84	63.88-82.87	-26.63	2.29	27.06
0.50	-0.5	100	99.9	101.62	0.34	100.96-102.28	1.62	1.28	1.66
0.50	-0.3	100	99.8	101.37	0.59	100.22-102.53	1.37	1.26	1.49
0.50	-0.1	100	99.1	100.66	1.03	98.64-102.67	0.66	1.24	1.22
0.50	0.0	100	98.6	100.11	1.29	97.58-102.64	0.11	1.24	1.30
0.50	0.1	100	97.6	99.06	1.60	95.91-102.20	-0.94	1.20	1.86
0.50	0.3	100	94.5	95.63	2.34	91.04-100.21	-4.37	1.10	4.96
0.50	0.5	100	89.3	90.03	3.00	84.15-95.91	-9.97	0.93	10.41
0.30	-0.3	200	188.2	209.51	4.26	201.15-217.87	4.75	2.32	10.42
0.30	-0.1	200	181.6	199.90	4.85	190.40-209.40	-0.05	2.24	4.85
0.30	0.0	200	176.8	193.27	5.27	182.94-203.60	-3.36	2.19	8.55
0.30	0.1	200	171.9	186.19	5.69	175.05-197.33	-6.91	2.11	14.94
0.30	0.3	200	158.8	168.51	6.11	156.54-180.48	-15.75	1.91	32.08
0.30	0.5	200	141.2	146.41	6.72	133.24-159.58	-26.80	1.58	54.01
0.50	-0.5	200	199.9	203.19	0.46	202.28-204.09	1.59	0.90	3.22
0.50	-0.3	200	199.5	202.69	0.87	200.99-204.39	1.35	0.89	2.83
0.50	-0.1	200	198.3	201.38	1.39	198.66-204.10	0.69	0.87	1.96
0.50	0.0	200	197.1	200.06	1.81	196.51-203.61	0.03	0.86	1.82
0.50	0.1	200	195.3	198.05	2.26	193.63-202.47	-0.98	0.84	2.98
0.50	0.3	200	189.4	191.55	3.24	185.20-197.90	-4.23	0.76	9.05
0.50	0.5	200	178.6	179.88	4.44	171.18-188.58	-10.06	0.63	20.60
0.30	-0.3	500	470.5	523.52	6.70	510.39-536.65	4.70	1.47	24.45
0.30	-0.1	500	453.4	498.60	7.74	483.43-513.77	-0.28	1.42	7.87
0.30	0.0	500	442.1	482.75	8.42	466.25-499.25	-3.45	1.38	19.20
0.30	0.1	500	429.8	464.98	8.85	447.63-482.33	-7.00	1.33	36.12
0.30	0.3	500	397.4	421.29	9.90	401.89-440.70	-15.74	1.20	79.33
0.30	0.5	500	352.5	365.01	10.93	343.59-386.43	-27.00	0.99	135.43
0.50	-0.5	500	499.8	507.89	0.78	506.35-509.42	1.58	0.57	7.92
0.50	-0.3	500	498.7	506.66	1.34	504.04-509.27	1.33	0.56	6.79
0.50	-0.1	500	495.6	503.14	2.29	498.65-507.62	0.63	0.55	3.88
0.50	0.0	500	492.9	500.08	2.83	494.54-505.63	0.02	0.54	2.83
0.50	0.1	500	488.3	494.98	3.59	487.95-502.02	-1.00	0.53	6.17
0.50	0.3	500	473.4	478.55	5.24	468.29-488.81	-4.29	0.48	22.08
0.50	0.5	500	446.1	449.17	7.06	435.34-463.00	-10.17	0.39	51.32

Table 5. Simulation results (1000 repetitions) of  $m = 10$  capture occasions considering autoregressive correlation structure for the GEE approach. Averages of the numbers of captured individuals, ( $\bar{n}$ ); the estimated population size, ( $AVE(\bar{N})$ ); standard errors of the estimated population size, ( $SE(\bar{N})$ ); nominal-based 95% confidence intervals (CI), that is,  $\hat{E}(\bar{N}) \pm 1.96 \times SE(\bar{N})$ ; percentage relative bias, ( $PRB = 100 \times \frac{E(\bar{N}) - N}{N}$ ), where  $E(\bar{N})$  is estimated by  $AVE(\bar{N})$ ; percentage coefficient of variation ( $CV = 100 \times \frac{SE(\bar{N})}{E(\bar{N})}$ ), and root mean square error, ( $RMSE = \sqrt{Var(\bar{N}) + Bias^2}$ ).

$\bar{p}$	$\alpha$	N	$\bar{n}$	$AVE(\bar{N})$	$SE(\bar{N})$	95%CI	PRB	CV	RMSE
0.30	-0.3	100	97.2	102.18	1.05	100.12-104.23	2.18	1.71	2.42
0.30	-0.1	100	96.2	100.98	1.48	98.08-103.88	0.98	1.68	1.77
0.30	0.0	100	95.4	100.06	1.70	96.72-103.40	0.06	1.65	1.71
0.30	0.1	100	94.1	98.58	2.06	94.56-102.61	-1.42	1.60	2.50
0.30	0.3	100	90.7	94.60	2.67	89.36-99.83	-5.40	1.47	6.03
0.30	0.5	100	83.9	87.10	3.61	80.02-94.18	-12.90	1.23	13.39
0.50	-0.5	100	98.0	100.11	0.02	100.07-100.15	0.11	0.33	0.11
0.50	-0.3	100	98.0	100.11	0.05	100.00-100.21	0.11	0.33	0.12
0.50	-0.1	100	98.0	100.08	0.18	99.72-100.43	0.08	0.33	0.20
0.50	0.0	100	97.9	100.03	0.28	99.49-100.58	0.03	0.32	0.28
0.50	0.1	100	97.8	99.93	0.43	99.09-100.76	-0.07	0.32	0.43
0.50	0.3	100	97.2	99.29	0.92	97.49-101.09	-0.71	0.31	1.16
0.50	0.5	100	95.0	97.07	1.64	93.86-100.28	-2.93	0.27	3.36
0.30	-0.3	200	196.3	204.24	1.50	201.30-207.17	2.12	1.21	4.49
0.30	-0.1	200	194.3	201.73	2.10	197.62-205.83	0.86	1.17	2.71
0.30	0.0	200	192.7	199.80	2.34	195.22-204.39	-0.10	1.15	2.35
0.30	0.1	200	190.5	197.20	2.90	191.52-202.88	-1.40	1.12	4.03
0.30	0.3	200	183.1	188.79	3.91	181.14-196.45	-5.61	1.02	11.87
0.30	0.5	200	170.1	174.35	5.03	164.49-184.21	-12.83	0.85	26.14
0.50	-0.5	200	198.0	200.22	0.03	200.15-200.28	0.11	0.23	0.22
0.50	-0.3	200	198.0	200.21	0.07	200.07-200.35	0.10	0.23	0.22
0.50	-0.1	200	197.9	200.15	0.24	199.69-200.61	0.08	0.23	0.28
0.50	0.0	200	197.8	200.04	0.40	199.26-200.82	0.02	0.22	0.40
0.50	0.1	200	197.6	199.82	0.63	198.58-201.06	-0.09	0.22	0.66
0.50	0.3	200	196.3	198.52	1.34	195.90-201.14	-0.74	0.21	2.00
0.50	0.5	200	192.0	194.12	2.36	189.49-198.75	-2.94	0.18	6.34
0.30	-0.3	500	493.9	510.55	2.25	506.14-514.97	2.11	0.76	10.79
0.30	-0.1	500	489.0	504.44	3.29	497.99-510.89	0.89	0.74	5.53
0.30	0.0	500	484.7	499.33	3.90	491.68-506.98	-0.13	0.72	3.96
0.30	0.1	500	479.0	492.58	4.70	483.37-501.79	-1.49	0.70	8.78
0.30	0.3	500	460.5	471.31	6.30	458.96-483.66	-5.74	0.63	29.38
0.30	0.5	500	428.1	435.24	7.72	420.11-450.37	-12.95	0.52	65.22
0.50	-0.5	500	498.0	500.53	0.05	500.43-500.63	0.11	0.15	0.53
0.50	-0.3	500	498.0	500.50	0.14	500.23-500.77	0.10	0.14	0.52
0.50	-0.1	500	497.9	500.35	0.40	499.56-501.14	0.07	0.14	0.54
0.50	0.0	500	497.6	500.10	0.59	498.94-501.25	0.02	0.14	0.60
0.50	0.1	500	497.1	499.54	0.96	497.65-501.42	-0.09	0.14	1.07
0.50	0.3	500	493.8	496.24	2.09	492.15-500.32	-0.75	0.13	4.30
0.50	0.5	500	483.2	485.47	3.87	477.89-493.04	-2.91	0.11	15.0



## References

- Agresti, A. (1994) Simple capture-recapture models permitting unequal catchability and variable sampling effort. *Biometrics*, 50(2), 494–500.
- Akanda, M.A.S. & Alpizar-Jara, R. (2014a) A generalized estimating equations approach for capture-recapture closed population models. *Environmental and Ecological Statistics*, 21(4), 667–688.
- Akanda, M.A.S. & Alpizar-Jara, R. (2014b) Estimation of capture probabilities using generalized estimating equations and mixed effects approaches. *Ecology and Evolution*, 4(7), 1158–1165.
- Briand, L.C., Emam, K.E., Freimut, B. & Oliver (1997) Quantitative evaluation of capture-recapture models to control software inspections, *Proceedings of the Eighth International Conference on Software Reliability Engineering*, Albuquerque, NM, 234–244.
- Chao, A., Lee, S.M. & Jeng, S.L. (1992) Estimating population size for capture-recapture data when capture probabilities vary by time and individual animal. *Biometrics*, 48(1), 201–216.
- Chao, A. & Lee, S.M. (1993) Estimating population size for continuous time capture-recapture models via sample coverage. *Biometrical Journal*, 35(1), 29–45.
- Chao, A., Tsay, P.K., Lin, S.H., Shau, W.Y. & Chao, D.Y. (2001) The applications of capture-recapture models to epidemiological data. *Stat. Med.*, 20(20), 3123–57.
- Chao, A. & Huggins, R.M. (2005) Modern closed-population capture-recapture models. In: C. Amstrup, T.L. McDonald and B.F.J. Manly (eds), *Handbook of capture-recapture analysis*. Princeton University Press, Princeton, NJ, 58–87.
- Diggle, P., Heagerty, P., Liang, K.Y. & Zeger, S. (2013) *Analysis of longitudinal data*. 2nd Edition, Oxford University Press, New York.
- Hardin, J. W. & Hilbe, J.M. (2013) *Generalized estimating equations*, 2nd Edition. Chapman and Hall/CRC Press, Boca Ratan, FL.
- Horvitz, D.G. & Thompson, D.J. (1952) A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.*, 47(260), 663–685.
- Huggins, R.M. (1989) On the statistical analysis of capture experiments. *Biometrika*, 76(1), 133–140.
- Huggins, R.M. (1991) Some practical aspects of a conditional likelihood approach to capture experiments. *Biometrics*, 47(2), 725–732.
- Huggins, R.M. & Yip, P.S.F. (1997) Statistical analysis of removal experiments with the use of auxiliary variables. *Stat. Sin.*, 7(3), 705–712.
- Hwang, W.H. & Huggins, R.M. (2005) An examination of the effect of heterogeneity on the estimation of population size using capture-recapture data. *Biometrika*, 92(1), 229–233.
- King, R. & Brooks, S.P. (2008) On the Bayesian estimation of a closed population size in the presence of heterogeneity and model uncertainty. *Biometrics*, 64(3), 816–824.
- Liang, K.Y. & Zeger, S.L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1), 13–22.
- Lloyd, C. & Yip, P. (1991) A unification of inference from capture-recapture studies through martingale estimating functions. In: Godambe, V.P. (eds) *Estimating functions*. Oxford: Clarendon Press, 65–88.
- Otis, D.L., Burnham, K.P., White, G.C. & Anderson, D.R. (1978) Statistical inference from capture data on closed animal populations. *Wildlife Monographs*, 62, 1–135.
- Pan, W. (2001) Akaike's information criterion in generalized estimating equations. *Biometrics*, 57(1), 120–125.
- Pledger, S. (2000) Unified maximum likelihood estimates for closed capture-recapture models using mixtures. *Biometrics*, 56(2), 434–442.
- Pollock, K., Hines, J. & Nichols, J. (1984) The use of auxiliary variables in capture-recapture and removal experiments. *Biometrics*, 40(2), 329–340.
- Pradel, R. & Sanz-Aguilar, A. (2012) Modeling trap-awareness and related phenomena in capture-recapture studies. *PLoS ONE*, 7(3), e32666.
- Qaqish, B.F. (2003) A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations. *Biometrika*, 90(2), 455–463.
- R Development Core Team (2016) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. Available from: <http://www.r-project.org/>.
- Rexstad, E. & Burnham, K. (1991) *User's guide for interactive program CAPTURE*. Colorado Cooperative Fish and Wildlife Research Unit, FortCollins.
- Seber, G.A.F. (2002) *The estimation of animal abundance and related parameters*. 2nd Edition, The Blackburn Press, London, Edward Arnold.
- Stanley, T.R. & Richards, J.D. (2005) Software review: a program for testing capture-recapture data for closure. *Wildlife Society Bulletin*, 33(2), 782–785.
- Stoklosa, J. & Huggins, R.M. (2012) A robust p-spline approach to closed population capture-recapture models with time dependence and heterogeneity. *Comput. Stat. Data Anal.*, 56(2), 408–417.
- Williams, B.K., Nichols, J.D. & Conroy, M.J. (2002) *Analysis and management of animal populations*. Academic Press, San Diego, California.
- Yang, H.C. & Chao, A. (2005) Modeling animals behavioral response by Markov chain models for capture-recapture experiments. *Biometrics*, 61(4), 1010–1017.
- Zeger, S.L. & Liang, K.Y. (1986) Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 42(1), 121–130.
- Zhang, S. (2012) A GEE approach for estimating size of hard-to-reach population by using capture-recapture data. *Statistics*, 46(2), 175–183.