International Journal of Development Mathematics Vol 1 Issue 1 Page No. 256 - 268

INTERNATIONAL JOURNAL OF DEVELOPMENT MATHEMATICS

ISSN: 3026-8656 (Print) | 3026-8699 (Online)

journal homepage: https://ijdm.org.ng/index.php/Journals

Population Size Estimation for Open Population Capture-Recapture Using Generalized Estimating Equations

Pascalis K. Matthew [a*], Danjuma Jibasen[b] and Ikeme J. Dike[b].

[a]*Department of Mathematical Sciences, Taraba State University, Jalingo, Nigeria.*

[b]*Department of Statistics, Modibbo Adama University, Yola, Nigeria.*

ARTICLE INFO

Article history:
Received 10 January 2024
Received in revised form 24 Fecruary 2024
Accepted 12 March 2024

Keywords:
Capture-recapture, Open oopulation, Population-size, GEE, QIC, RMSE

MSC 2020 Subject classification:
62J12, 62J99, 62P10, 62P12.

ABSTRACT

This study applied a Generalized Estimating Equation (GEE) approach accounting for heterogeneity, behavioural and environmental effects to estimate the population-size for open population capture-recapture experiment. A non-spatial open population capture-recapture data was simulated using three different trapping occasions (m =4, 5, 6) with population size (N=1000). The GEE estimates for behavioral/heterogeneity covariates and heterogeneity/behavioral/environmental covariates models under different working correlation structures were also presented. The simulation results, show that, for the heterogeneity/behavioral model, the root mean square error (RMSE) for (m = 4, 5, 6) recapture occasions when N=1000 are (199.2171, 199.2573, 1193.2222) and for the heterogeneity/behavioral/environmental covariates model, the RMSE for (m= 4, 5, 6) recapture occasions when N=1000 are (199.2071, 199.2252, 1193.2052). The comparison results between behavioral/heterogeneity covariates model and heterogeneity/behavioral/environmental covariates model show that, the heterogeneity/behavioral/environmental covariates model performs better in estimating the population-size in terms of accuracy as indicated by the smallest values of RMSE. The study concludes that; environmental factors play an important role when estimating the population-size for open population capture-recapture experiments.

1. Introduction

Capture-recapture is a study where repeated observations are collected on the same individuals over time. The method is used for estimating population size and other parameters that are based on ratios of marked to unmarked individuals (*Wesson et al*., 2023). When the study population consists of *N* elements, *N* being unknown and finite, capture-recapture methods can be applied to estimate the size of *N* (*Koivuniemi et al*., 2019). The population under study is sampled two or more times, so that, at each time, the unmarked animals caught are marked uniquely (usually with a numbered leg band in bird studies); animals previously marked have their capture recorded and then all of the animals are released back into the population (*Wesson et al*., 2017). The experimenter has the complete capture history of each animal handled at the end of the study. Capture-recapture methods are extensively used in animal population estimation and in other fields such as quality control and epidemiology (Akanda and Alpizar-Jara, 2017). A set of models and their inference procedures have been proposed in the capture-recapture literatures for the estimation of animal population size from capture-recapture data (*Chowdhury et al*., 2019). There are two types of capture-recapture models;(a) models that assume the population remains constant throughout the period of study (closed population models), and (b) models where individuals can be added due to birth, immigration, or reintroduction, and can be removed from the population through death, emigration, or other reasons (open population models). Open population models can also be categorized according to whether the recaptures are observations on live animals or dead animals. The capture–recapture data collected on the same individuals across successive capture occasions, are binary longitudinal or repeated measurements data (Akanda and Alpizar-Jara, 2014). These repeated observations are often correlated over time.

The dependency or correlation structure may be induced by incorporating individual heterogeneity. Failure to account for this dependency may provide biased estimates (*Chao et al*., 2021). The Generalized Estimating Equations (GEE) approach is a powerful statistical tool that extends the application of capture-recapture models to open populations, allowing for more accurate population estimates. The GEE approach addresses the limitations of closed population models by providing a framework to estimate population size while considering these

* Corresponding author. Tel.: +2347039321996
E-mail address: pkadaro@yahoo.com (P. K. Matthew)
https://doi.org/10.62054/ijdm/0101.19

demographic changes. *Worthington et al*. (2018) developed a multi-state model to estimate the size of a closed population from capture–recapture studies. The simulation results show that the estimate of population size can be biased when movement between states is not accounted for. *Mamadou et al*. (2019) proposed a new algorithm for estimating the parameters of a robust design with a fairly large number of capture occasions and a simple parametric bootstrap variance estimator. *Kevin et al*. (2017) used capture-recapture techniques to estimate the number of men who have sex with men and among female sex workers in 11 selected towns in Uganda. The study utilized conventional 2-source capture-recapture (CRC) to estimate the population of female sex workers and men who have sex with men. *You et al*. (2021) proposed a modern method to estimate population size based on capture-recapture designs of k-samples. The study covers models assuming a single constraint (identification assumption) on the k-dimensional distribution such that the target quantity is identified and the statistical model is unrestricted.

The study presents solutions for linear and non-linear constraints commonly assumed to identify capture-recapture models, including no k-way interaction in linear and log-linear models, independence or conditional independence. *Kreshpaj et al*. (2021) used capture-recapture methods to estimate the magnitude of under-reporting of occupational injuries (OIs) among precarious and non-precarious workers in Sweden in 2013. The study demonstrated empirically in Sweden that under-reporting of OIs is 50% higher among precariously employed workers. Capture-recapture study was used to estimate the population of female sex workers in Ghana (*Guure et al*.,2021). It was also used in Bhutan (*Khandu et al*., 2021), in Juba (*Okiria et al*., 2019), in Hai Phong province (*Nguyen et al*.,2021). *Wesson et al.* (2023) used capture-recapture methods to estimate the size of hidden populations by leveraging the proportion of over-lapping of the population on independent lists. The study used simulations to model when capture-recapture methods deliver biased or unbiased estimates and compare model selection criteria. The study concluded that, conventional capture-recapture model selection fails with sparse cell counts. This naïve approach to model selection should be replaced with the implementation of multiple different models in order triangulate the truth in real-world applications. Huang and Pan (2021) proposed a novel method to model the mean and within-subject correlation coefficients for longitudinal binary data, simultaneously, by taking into account the constraints of the upper bounds. A joint GEE method was developed for the purpose and the resulting correlation coefficients were shown to satisfy the constraints. Asymptotic normality of the parameter estimators was derived and simulation studies were made under various scenarios, showing that the proposed joint GEE method works very well even if the working co-variance structures are mis-specified. *Zhuoran et al*. (2023) used GEE to investigate the association between exposures and hearing loss, because it is robust to misspecification of the correlation matrix.

The study proposed to model the correlation coefficients and use second-order GEE to estimate the correlation parameters. In simulation studies, the proposed method achieves an efficiency gain which is larger for the coefficients of the covariates corresponding to the within-cluster variation than the coefficients of cluster-level covariates. *Asep et al*. (2020) used GEE as one of the longitudinal data regression methods for composite index data on children's welfare (CWCI) in West Java Province. The study concluded that the research is important as an alternative approach of longitudinal data regression analysis when a correlation problem is found in the condition of observed data. *Nasrin et al*. (2021) proposed a method for the analysis of spatially clustered real data based on GEE which were originally developed for analyzing longitudinal data. A simulation study was conducted, and the proposed GEE approach yields better results than the popular conditional auto-regressive model from both perspectives of parameter estimation and spatial process capturing. *Goldenberg et al*. (2023) used bivariate and multivariable logistic regression with GEE to model the association of unstable housing exposure and evictions with intimate partner violence (IPV) and workplace violence among a community-based longitudinal cohort of cisgender and transgender Female sex workers in Vancouver, Canada, from 2010 through 2019. The study concludes that, Female sex workers face a high burden of unstable housing and evictions, which are linked to increased odds of intimate partner and workplace violence. Increased access to safe, women-centered, and nondiscriminatory housing is urgently needed. The applications of GEE approach by (Zhang, 2012; Akanda and Alpizar-Jara, 2014a; 2017; *Chao et al*., 2021) were all focused on closed population capture-recapture models. The aim of this paper is to apply GEE approach to estimate the population size of an open population capture-recapture data, taking into account individual heterogeneity, behavioural effects and environmental covariates.

2. Methods

2.1. Linking GEE to capture-recapture

Let N be the total number of individuals in the population of study, and let m be the possible number of capture occasions, $(m \geq 2)$.Let Y_{ij} the indicator of the i^{th} individual being caught in the j^{th} occasion, that is,

$Y_{ij} = 1$ if i^{th} individual is captured in the j^{th} occasion, and $Y_{ij} = 0$, otherwise. Let T_i be the number of samples that the i^{th} individual belongs to, then let $T_i = \sum_{j=1}^{m} Y_{ij}$ .Let $Y_{ij} = (y_{i1},...,y_{im})^T$ be the $m \times 1$ random vector to record the capture history of the i^{th} individual for the *m* occasions. Covariates z_i and x_{ij} are associated with the i^{th} individual and the j^{th} occasion, for $i = 1,...,N$, $j = 1,...,m$. Let z_i be an individual measurable covariate, such as age, sex or weight, etc. Let x_{ij} be a measurable occasion related covariate for the i^{th} individual, and let

$$X_i = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ z_i & z_i & \cdots & z_i \\ x_{i1} & 0 & \cdots & 0 \\ 0 & x_{i2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & x_{im} \end{bmatrix}^T$$

be the covariance matrix.

2.1.1. Zhang (2012) models

Zhang (2012) proposed the probability P_{ij} that the i^{th} individual is captured on j^{th} occasion is given by

$$P_{ij} = \Pr(Y_{ij} = 1 | z_i, x_j) = h(\beta_0 + \beta_1 z_i + \beta_{j+1} x_{ij}) \tag{1}$$

for $i = 1,...,N$, $j = 1,...,m$, where $h(u) = \frac{e^u}{1+e^u} = (1+(e^{-u}))^{-1}$ is the logistic function. If $x_{i_1 j} = x_{i_2 j} = x_j$ for $i_1 \neq i_2$, model (2) may be used for simplicity

$$P_{ij} = \Pr(Y_{ij} = 1 | z_i, x_j) = h(\beta_0 + \beta_1 z_i + \beta_{j+1} x_j) \tag{2}$$

for $i = 1,...,N$, $j = 1,...,m$. Then the GEE model by Zhang (2012) is given by

$$h(\beta) = \sum_{i=1}^{n}\sum_{j=1}^{m}(Y_{ij} - \mu_{ij}) + \sum_{i=1}^{n}\sum_{j=1}^{m}(Y_{ij} - \mu_{ij}) z_i \tag{3}$$

2.1.2. Akanda and Alpizar-Jara (2014) models

Akanda and Alpizar-Jara (2014) introduced v_{ij} into the model (2) which is an indicator that, the i^{th} individual is captured prior to j^{th} occasion, and which depends on occasions as well as the individual i.e

$$v_{ij} = \begin{cases} 1, & caught \\ 0, & otherwise \end{cases} \tag{4}$$

The probability P_{ij} that the i^{th} individual is captured on j^{th} occasion is given by

$$P_{ij} = \Pr(Y_{ij} = 1 | z_i, x_j, v_{ij}) = h(\beta_0 + \beta_1 z_i + \beta_{j+1} x_j + \beta_{bh} v_{ij}) \tag{5}$$

for $i = 1,...,N$, $j = 1,...,m$. Then the GEE model by Akanda and Alpizar-Jara (2014) is given by

$$h(\beta) = \sum_{i=1}^{n}\sum_{j=1}^{m}(Y_{ij} - \mu_{ij}) + \sum_{i=1}^{n}\sum_{j=1}^{m}(Y_{ij} - \mu_{ij}) z_i + \sum_{i=1}^{n}\sum_{j=1}^{m}(Y_{ij} - \mu_{ij}) v_{ij} \tag{6}$$

2.1.3. Matthew *et al*. (2023) model

Matthew *et al*. (2023) introduced f_{ij} into the (4) which denote the environmental covariates such as mean temperature, humidity or rainfall recorded for i^{th} individual at j^{th} occasions.

Therefore, the probability P_{ij} that the i^{th} individual is captured on j^{th} occasion is given by

$$P_{ij} = \Pr(Y_{ij} = 1 | z_i, x_j, v_{ij}, f_{ij}) = h(\beta_0 + \beta_1 z_i + \beta_{j+1} x_j + \beta_{bh} v_{ij} + \beta_{ev} f_{ij}) \tag{7}$$

for $i = 1,..., N$, $j = 1,..., m$. Then the GEE model by *Matthew et al*. (2023) is given by

$$h(\beta) = \sum_{i=1}^{n}\sum_{j=1}^{m}\left(Y_{ij} - \mu_{ij}\right) + \sum_{i=1}^{n}\sum_{j=1}^{m}\left(Y_{ij} - \mu_{ij}\right)z_i + \sum_{i=1}^{n}\sum_{j=1}^{m}\left(Y_{ij} - \mu_{ij}\right)v_{ij} + \sum_{i=1}^{n}\sum_{j=1}^{m}\left(Y_{ij} - \mu_{ij}\right)f_{ij} = 0 \tag{8}$$

2.4. Fitting Generalized Estimating Equation Models

The following steps were used in fitting the proposed GEE models as suggested (Wentao,2019):

1. **Linear predictor and best link function:** A link transformation function is specified that allows the response variable to be expressed as a function of parameter estimate β of an additive model, which model the expected value of the marginal response for the population $\mu_i = \mathrm{E}(Y_i)$ to a linear combination of the covariates. The choice of link function depends primarily on the distribution specified.
2. **Distribution of the response variable:** The GEE allows the specifications of the distribution from the exponential family of distributions, including Normal, inverse Normal, Poisson, Binomial, Gamma distribution and negative Binomial. Prior knowledge of the distribution of the response variable is important.
3. **Structure of the correlation within the response variable:** Although GEE is robust when it comes to misspecification of the correlation structure, it is important to take precaution when it comes to specifying the structure. The working correlation employed are as follows: (a) Independent Structure(b)Exchangeable Structure (c) Autoregressive Structure (d) Unstructured Structure

2.5. Model Selection of Generalized Estimating Equations

A modification of AIC known as the Quasi-Likelihood Information Criterion (QIC) was introduced by Pan (2001). The QIC is constructed on the basis of independence of the working correlation assumption. The QIC can be used to select the appropriate working correlation structure.
Then the simplified version of QIC given by Pan (2001) as cited by Wang et.al (2015) is denoted by

$$QIC = -2Q(\hat{\mu}; I) + 2p \tag{9}$$

Therefore, QIC is used to select an optimal correlation structure. The model with the smallest QIC will be chosen as the best parsimonious model with the best correlation structure (Pan, 2001; *Wang et al.,* 2015).

2.6. Population Size Estimator and the Variance

Let p_{ij} $(i = 1,..., n;\ j = 1,..., m)$ represent the probability of capture for the i^{th} individual during the j^{th} capture occasion. The open-population capture–recapture models employed here relate a differentiable function of the p_{ij} to a linear function of external covariates (*Matthew et al*., 2023). These covariates include sex, weight, behaviour effect and mean temperature. Individual capture probabilities were modelled with a logistic regression, so that,

$$P_{ij} = \frac{e^{\beta_0 + \beta_{sex} sex_i + \beta_{wt} wt_i + \beta_{bh} v_{ij}}}{1 + e^{\beta_0 + \beta_{sex} sex_i + \beta_{wt} wt_i + \beta_{bh} v_{ij}}} \tag{10}$$

for behavioural/heterogeneity covariates and

$$P_{ij} = \frac{e^{\beta_0 + \beta_{sex} sex_i + \beta_{wt} wt_i + \beta_{bh} v_{ij} + \beta_{bh} f_{ij}}}{1 + e^{\beta_0 + \beta_{sex} sex_i + \beta_{wt} wt_i + \beta_{bh} v_{ij} + \beta_{bh} f_{ij}}} \tag{11}$$

for behavioural/heterogeneity/environmental covariates.

The estimates for the parameters β_i are obtained by the method of maximum likelihood. Given a set of observed capture histories, I_{ij} , where $I_{ij}=1$ if the i^{th} individual was seen at the j^{th} occasion and $I_{ij}=0$ otherwise.

The estimated capture probability and standard error for the i^{th} individual at j^{th} occasion is labelled as $\hat{p}_{ij}$ and $\hat{\pi}_{\hat{p}_{ij}}$, respectively. So that, $E\left[\hat{p}_{ij}\right]=p_{ij}$ and $E\left[\hat{\pi}_{\hat{p}_{ij}}\right]=\pi_{\hat{p}_{ij}}$. The true number of individuals alive at trap j^{th} occasion is N_j . Assuming capture is statistically independent across individuals (i.e., across subscript i), $E\left[I_{ij}\right]=p_{ij}$ and $\operatorname{var}(I_{ij})=p_{ij}(1-p_{ij})$. (McDonald and Amstrup, 2001) give the estimate of population size at the j^{th} capture occasion, denoted $\hat{N}_j$ $(j=2,\ldots,m)$, as

$$\hat{N}_j=\sum_{i=1}^{n}\frac{I_{ij}}{\hat{p}_{ij}} \tag{12}$$

$$E\left[\hat{N}_j\right]\approx N_j+\sum_{i=1}^{N_j}\frac{1}{p_{ij}^2}\left(\pi_{\hat{p}_{ij}}^2-\operatorname{cov}(I_{ij},\hat{p}_{ij})\right) \tag{13}$$

So that $E\left[\hat{N}_j\right]\approx N_j$ provided that $\operatorname{cov}(I_{ij},\hat{p}_{ij})\approx\pi_{\hat{p}_{ij}}^2$ for all i .The approximate variance of the estimator $\hat{N}_j$ is given by

$$Var\left(\hat{N}_j\right)=\sum_{i=1}^{n}\left(\frac{I_{ij}(1-\hat{p}_{ij})}{\hat{p}_{ij}^2}+\frac{I_{ij}\pi_{\hat{p}_{ij}}^2}{\hat{p}_{ij}^3}+\frac{I_{ij}(1-\hat{p}_{ij})\pi_{\hat{p}_{ij}}^2}{\hat{p}_{ij}^4}\right) \tag{14}$$

The performances of the estimators are based on the comparisons of root mean square error (RMSE). RMSE is given by

$$RMSE=\sqrt{Var(\hat{N})+bias^2} \tag{15}$$

3. Results and Discussion

3.1. Simulation 1

In order to illustrate the GEE approach on open population capture-recapture data, a non-spatial open population data was simulated using R with the following parameters; $n=1000$, $\phi=0.8$, $\lambda=1.2$, $p=0.4$, and $m=4$ trapping occasions was used to simulate the recapture history. The individual discrete covariate, sex, was simulated based on the Binomial distribution *BIN(1000, 0.5)*. The individual continuous covariate, weight, and the environmental covariate, temperature, were simulated based on the Normal distribution *N(15, 5)* and *N(30, 10)* respectively. The indicator variable, behavioural effect was generated from the capture history. The simulation results for the different covariate structures are presented in Tables 1 and 2 respectively.

Table 1 GEE estimates for $p_{ij} = h(\beta_0 + \beta_{sex} sex_i + \beta_{wt} wt_i + \beta_{bh} v_{ij})$ model m=4, N=1000

Structure	Exchangeable		Independence	
Coefficients	Estimate	Std Err	Estimate	Std Err
β_0	0.339550	0.07326	0.326927	0.08639
β_{sex}	-0.01646	0.03692	-0.009717	0.04142
β_{wt}	0.000230	0.00615	0.000763	0.00672
β_{bh}	-0.02239	0.03556	-0.006817	0.04066
QIC	**265.5100**		266.7600	
Structure	Unstructured		AR-1	
Coefficients	Estimate	Std Err	Estimate	Std Err
β_0	0.38657	0.06616	0.342293	0.07839
β_{sex}	-0.03098	0.03383	-0.01662	0.03894
β_{wt}	-0.00152	0.00573	0.000318	0.00636
β_{bh}	-0.01865	0.03242	-0.02236	0.03774
QIC	265.550		266.8600	

The information provided in Table 1 shows the GEE estimates for individual heterogeneity and behavioural covariates model under different working correlation structure.

Table 2 GEE estimates for $p_{ij} = h(\beta_0 + \beta_{sex} sex_i + \beta_{wt} wt_i + \beta_{bh} v_{ij} + \beta_{tp} f_{ij})$ model m=4, N=1000

Structure	Exchangeable		Independence	
Coefficients	Estimate	Std Err	Estimate	Std Err
β_0	0.269506	0.10867	0.253823	0.12663
β_{wt}	0.000301	0.00618	0.000859	0.00676
β_{bh}	-0.02161	0.03580	-0.02181	0.04114
β_{sex}	-0.01738	0.03668	-0.01060	0.04112
β_{tp}	0.004731	0.00436	0.004905	0.00433
QIC	**267.3100**		268.5500	
Structure	Unstructured		AR-1	
Coefficients	Estimate	Std Err	Estimate	Std Err
β_0	0.329570	0.09693	0.273890	0.11651
β_{wt}	-0.00153	0.00575	0.000432	0.00641
β_{bh}	-0.00173	0.03287	-0.02143	0.03812
β_{sex}	-0.03204	0.03359	-0.01748	0.03866
β_{tp}	0.00387	0.00462	0.004589	0.00436
QIC	267.6500		267.6100	

The information provided in Table 2 shows the GEE estimates for individual heterogeneity, behavioural and environmental covariates model under different working correlation structure.

Table 3 Parameter estimates for open population Capture-Recapture models m=4, N=1000

Model	Struc	$\hat{N}$	$S.E$	RMSE
$p_{ij} = h(\beta_0 + \beta_{sex} sex_i + \beta_{wt} wt_i + \beta_{bh} v_{ij})$	Excha	253	17.893	199.2171
$p_{ij} = h(\beta_0 + \beta_{sex} sex_i + \beta_{wt} wt_i + \beta_{bh} v_{ij} + \beta_{tp} f_{ij})$	Unstr	260	17.432	199.2071

3.2. Discussion on Simulation 1

From the information provided in Table 1, which displays the GEE estimates for the individual heterogeneity and behavioural covariates model under different working correlation structures, it can be observed that, the QIC values for exchangeable, independence, unstructured and AR-1 are 265.51, 266.76, 265.55 and 266.86 respectively. The model under exchangeable working correlation structure was selected as the best model, because it has the smallest value of QIC, and was used to calculate the parameter estimates for open population capture-recapture data given in Table 3. From the information provided in Table 2, which displays the GEE estimates for the individual heterogeneity, behavioural and environmental covariates model under different working correlation structures, it can be observed that, the QIC values for exchangeable, independence, unstructured and AR-1 are 267.31, 268.55, 267.65 and 267.61 respectively. The model under exchangeable working correlation structure was selected as the best model, because it has the smallest value of QIC, and was used to calculate the parameter estimates for open population capture-recapture data given in Table 3. From Tables 1 and 2, the positive values of the estimates for the covariates ($\beta_{wt}, \beta_{sex}, \beta_{bh}, \beta_{tp}$) influences the dependent variable p_{ij} positively i.e as the values of the independent variables ($\beta_{wt}, \beta_{sex}, \beta_{bh}, \beta_{tp}$) increases, the dependent variable p_{ij} also increase. While, the negative values of the estimates for the covariates (β_{wt}, $\beta_{sex}, \beta_{bh}, \beta_{tp}$) influences the dependent variable p_{ij} negatively i.e as the values of the independent variables (β_{wt}, $\beta_{sex}, \beta_{bh}, \beta_{tp}$) decreases, the dependent variable p_{ij} also decrease. From the information provided in Table 3, the estimated population size $\hat{N}$ =253 with a *SE*($\hat{N}$)=17.893 and *RMSE*=199.2171 when individual heterogeneity/behavioural covariates model is considered under exchangeable working correlation structure. The estimated population size $\hat{N}$ =260 with a *SE*($\hat{N}$)= 17.432 with *RMSE*=199.2071 when behavioural/individual heterogeneity/environmental covariates model is considered under unstructured working correlation structure. Table 3 also displayed, the comparison between $p_{ij} = h(\beta_0 + \beta_{sex} sex_i + \beta_{wt} wt_i + \beta_{bh} v_{ij})$ by Akanda and Alpizar-Jara (2014) and $p_{ij} = h(\beta_0 + \beta_{sex} sex_i + \beta_{wt} wt_i + \beta_{bh} v_{ij} + \beta_{tp} f_{ij})$ by Matthew *et al*. (2023). The model proposed by (Matthew *et al*., 2023) performs better in estimating the population size in terms of accuracy as indicated by the smallest value of *RMSE*=199.2071.

3.3. Simulation 2

In order to illustrate the GEE approach on open population capture-recapture data, a non-spatial open population data was simulated using R with the following parameters; $n = 1000$, $\phi =$ 0.8, $\lambda =$ 1.2, $p =$ 0.4, and $m = 5$ trapping occasions was used to simulate the recapture history. The individual discrete covariate, sex, was simulated based on the Binomial distribution *BIN(1000, 0.5)*. The individual continuous covariate, weight, and the environmental covariate, temperature, were simulated based on the Normal distribution *N(15, 5)* and *N(30, 10)* respectively. The indicator variable, behavioural effect was generated from the capture history. The simulation results for the different covariate structures are presented in Tables 4 and 5 respectively.

Table 4 GEE estimates for $p_{ij} = h(\beta_0 + \beta_{sex} sex_i + \beta_{wt} wt_i + \beta_{bh} v_{ij})$ model *m*=5, *N*=1000

Structure	Exchangeable	Independence

Coefficients	Estimate	Std Err	Estimate	Std Err
β_0	0.20471	0.08144	0.45219	0.06937
β_{sex}	0.07210	0.06178	-0.08882	0.03475
β_{wt}	-0.01270	0.00872	-0.00256	0.00559
β_{bh}	-0.08663	0.05161	-0.00364	0.03350
QIC	**293.634**		293.930	
Structure	Unstructured		AR-1	
Coefficients	Estimate	Std Err	Estimate	Std Err
β_0	0.47546	0.08830	0.45424	0.07094
β_{sex}	-0.08396	0.03860	-0.08818	0.03495
β_{wt}	-0.00343	0.00480	-0.04665	0.00552
β_{bh}	-0.00398	0.00450	-0.00397	0.03352
QIC	296.600		293.940	

The information provided in Table 4 shows the GEE estimates for individual heterogeneity and behavioural covariates model under different working correlation structure.

Table 5 GEE estimates for $p_{ij} = h(\beta_0 + \beta_{sex} sex_i + \beta_{wt} wt_i + \beta_{bh} v_{ij} + \beta_{tp} f_{ij})$ model m=5, N=1000

Structure	Exchangeable		Independence	
Coefficients	Estimate	Std Err	Estimate	Std Err
β_0	0.42960	0.08297	0.41018	0.07872
β_{wt}	-0.0052	0.00510	-0.00380	0.00560
β_{bh}	-0.0593	0.03531	-0.04660	0.03347
β_{sex}	-0.0878	0.03459	-0.09010	0.03463
β_{tp}	0.00242	0.00246	0.003020	0.00266
QIC	**294.750**		294.900	
Structure	Unstructured		AR-1	
Coefficients	Estimate	Std Err	Estimate	Std Err
β_0	0.45511	0.10176	0.41399	0.08119
β_{wt}	-0.00885	0.00484	-0.00420	0.00553
β_{bh}	-0.06156	0.04514	-0.04660	0.03354
β_{sex}	-0.08472	0.03839	-0.08930	0.03480
β_{tp}	0.00150	0.00246	0.00289	0.00263
QIC	297.470		294.850	

The information provided in Table 5 shows the GEE estimates for individual heterogeneity/behavioural/environmental covariates model under different working correlation structure.

Table 6 Parameter estimates for open population Capture-Recapture models m=5, N=1000

Model	Struc	$\hat{N}$	$S.E$	RMSE
$p_{ij} = h(\beta_0 + \beta_{sex} sex_i + \beta_{wt} wt_i + \beta_{bh} v_{ij})$	Excha	249	17.625	199.2573
$p_{ij} = h(\beta_0 + \beta_{sex} sex_i + \beta_{wt} wt_i + \beta_{bh} v_{ij} + \beta_{tp} f_{ij})$	Excha	250	17.608	199.2252

3.4. Discussion on Simulation 2

From the information provided in Table 4, which displays the GEE estimates for the individual heterogeneity and behavioural covariates model under different working correlation structures, it can be observed that, the QIC values for exchangeable, independence, unstructured and AR-1 are 293.634, 293.93, 296.60 and 293.94 respectively. The model under exchangeable working correlation structure was selected as the best model, because it has the smallest value of QIC, and was used to calculate the parameter estimates for open population capture-recapture data given in Table 6. From the information provided in Table 5, which displays the GEE estimates for the individual heterogeneity, behavioural and environmental covariates model under different working correlation structures, it can be observed that, the QIC values for exchangeable, independence, unstructured and AR-1 are 294.75, 294.90, 297.47 and 294.85 respectively. The model under exchangeable working correlation structure was selected as the best model, because it has the smallest value of QIC, and was used to calculate the parameter estimates for open population capture-recapture data given in Table 6. From Tables 4 and 5, the positive values of the estimates for the covariates ($\beta_{wt}, \beta_{sex}, \beta_{bh}, \beta_{tp}$) influences the dependent variable p_{ij} positively i.e as the values of the independent variables ($\beta_{wt}, \beta_{sex}, \beta_{bh}$, β_{tp}) increases, the dependent variable p_{ij} also increase. While, the negative values of the estimates for the covariates ($\beta_{wt}, \beta_{sex}, \beta_{bh}, \beta_{tp}$) influences the dependent variable p_{ij} negatively i.e as the values of the independent variables ($\beta_{wt}, \beta_{sex}, \beta_{bh}, \beta_{tp}$) decreases, the dependent variable p_{ij} also decrease. From the information provided in Table 6, the estimated population size $\hat{N}$ =249 with a *SE*($\hat{N}$)= 17.625 and *RMSE*=199.2573 when individual heterogeneity/behavioural covariates model is considered under exchangeable working correlation structure. The estimated population size $\hat{N}$ =250 with a *S E*($\hat{N}$)= 17.608 with *RMSE*=199.2252 when behavioural/individual heterogeneity/environmental covariates m odel is considered under unstructured working correlation structure. Table 6 also presented, the comparison bet ween $p_{ij} = h(\beta_0 + \beta_{sex} sex_i + \beta_{wt} wt_i + \beta_{bh} v_{ij})$ by Akanda and Alpizar- Jara, (2014) and $p_{ij} = h(\beta_0 + \beta_{sex} sex_i + \beta_{wt} wt_i + \beta_{bh} v_{ij} + \beta_{tp} f_{ij})$ by Matthew *et al*. (2023). The model proposed by (Matthew *et al*., 2023) performs better in estimating the population size in terms of accuracy as indicated by the smallest value of *RMSE*=199.2252.

3.5. Simulation 3

In order to illustrate the GEE approach on open population capture-recapture data, a non-spatial open population data was simulated using R with the following parameters; $n = 1000$, $\phi =$ 0.8, $\lambda =$ 1.2, $p =$ 0.4, and $m = 6$ trapping occasions was used to simulate the recapture history. The individual discrete covariate, sex, was simulated based on the Binomial distribution *BIN(1000, 0.5)*. The individual continuous covariate, weight, and the environmental covariate, temperature, were simulated based on the Normal distribution *N(15, 5)* and *N(30, 10)* respectively. The indicator variable, behavioural effect was generated from the capture history. The simulation results for the different covariate structures are presented in Tables 7 and 8 respectively.

Table 7 GEE estimates for $p_{ij} = h(\beta_0 + \beta_{sex} sex_i + \beta_{wt} wt_i + \beta_{bh} v_{ij})$ model *m*=6, *N*=1000

Structure	Exchangeable		Independence	
Coefficients	Estimate	Std Err	Estimate	Std Err
β_0	0.24842	0.05385	0.24811	0.05484
β_{sex}	0.01835	0.03068	0.01670	0.03162
β_{wt}	-0.00487	0.00592	-0.0484	0.00589
β_{bh}	-0.00449	0.02774	0.00124	0.02945
QIC	316.260		**316.130**	
Structure	Unstructured		AR-1	
Coefficients	Estimate	Std Err	Estimate	Std Err
β_0	0.27007	0.05295	0.25001	0.05470
β_{sex}	0.02753	0.03186	0.01646	0.03157
β_{wt}	-0.00806	0.00527	-0.00490	0.00589
β_{bh}	-0.01177	0.0565	0.00034	0.02936
QIC	317.000		316.430	

The information provided Table 7 shows the GEE estimates for individual heterogeneity and behavioural covariates model under different working correlation structure.

Table 8 GEE estimates for $p_{ij} = h(\beta_0 + \beta_{sex} sex_i + \beta_{wt} wt_i + \beta_{bh} v_{ij} + \beta_{tp} f_{ij})$ model m=6, N=1000

Structure	Exchangeable		Independence	
Coefficients	Estimate	Std Err	Estimate	Std Err
β_0	0.28840	0.06677	0.27960	0.06930
β_{wt}	0.00443	0.00525	0.00560	0.00550
β_{bh}	0.03332	0.02512	0.03750	0.02960
β_{sex}	0.03015	0.03284	0.03170	0.03071
β_{tp}	0.00248	0.00270	0.00230	0.00268
QIC	320.339		**319.850**	
Structure	Unstructured		AR-1	
Coefficients	Estimate	Std Err	Estimate	Std Err
β_0	0.33759	0.06480	0.28626	0.07288
β_{wt}	0.00862	0.00540	-0.00517	0.00586
β_{bh}	0.03764	0.02470	-0.00011	0.07288
β_{sex}	0.03274	0.03030	0.016151	0.03164
β_{tp}	0.00183	0.00260	-0.00229	0.00271
QIC	320.647		319.990	

The information provided in Table 8 shows the GEE estimates for individual heterogeneity/behavioural/environmental covariates model under different working correlation structure.

Table 9 Parameter estimates for open population Capture-Recapture models m=6, N=1000

Model	Struc	$\hat{N}$	$S.E$	RMSE
$p_{ij} = h(\beta_0 + \beta_{sex} sex_i + \beta_{wt} wt_i + \beta_{bh} v_{ij})$	Indep	243	17.246	1193.2222
$p_{ij} = h(\beta_0 + \beta_{sex} sex_i + \beta_{wt} wt_i + \beta_{bh} v_{ij} + \beta_{tp} f_{ij})$	Indep	243	17.206	1193.2053

3.6. Discussion on Simulation 3

From the information provided in Table 7, which displays the GEE estimates for the individual heterogeneity and behavioural covariates model under different working correlation structures, it can be observed that, the QIC values for exchangeable, independence, unstructured and AR-1 are 316.26, 316.13, 317.00 and 316.43 respectively. The model under independence working correlation structure was selected as the best model, because it has the smallest value of QIC, and was used to calculate the parameter estimates for open population capture-recapture data given in Table 9. From the information provided in Table 8, which displays the GEE estimates for the individual heterogeneity, behavioural and environmental covariates model under different working correlation structures, it can be observed that, the QIC values for exchangeable, independence, unstructured and AR-1 are 320.339, 319.85, 320.647 and 319.99 respectively. The model under independence working correlation structure was selected as the best model, because it has the smallest value of QIC, and was used to calculate the parameter estimates for open population capture-recapture data given in Table 9. From Tables 7 and 8, the positive values of the estimates for the covariates ($\beta_{wt}, \beta_{sex}, \beta_{bh}, \beta_{tp}$) influences the dependent variable p_{ij} positively i.e as the values of the independent variables ($\beta_{wt}, \beta_{sex}, \beta_{bh}$, β_{tp}) increases, the dependent variable p_{ij} also increase. While, the negative values of the estimates for the covariates ($\beta_{wt}, \beta_{sex}, \beta_{bh}, \beta_{tp}$) influences the dependent variable p_{ij} negatively i.e as the values of the independent variables ($\beta_{wt}, \beta_{sex}, \beta_{bh}, \beta_{tp}$) decreases, the dependent variable p_{ij} also decrease. From Table 9, the estimated population size $\hat{N}$ =243 with a $SE(\hat{N})$=17.246 and $RMSE$=1193.2222 when individual heterogeneity/behavioural covariates model is considered under independence working correlation structure. The estimated population size $\hat{N}$ =243 with a $SE(\hat{N})$=17.2067 and $RMSE$=1193.2053 when behavioural/individual heterogeneity/environmental covariates model is considered under independence working correlation structure. From Table 9, the comparison between $p_{ij} = h(\beta_0 + \beta_{sex} sex_i + \beta_{wt} wt_i + \beta_{bh} v_{ij})$ by Akanda and Alpizar-Jara (2014) and $p_{ij} = h(\beta_0 + \beta_{sex} sex_i + \beta_{wt} wt_i + \beta_{bh} v_{ij} + \beta_{tp} f_{ij})$, by *Matthew et al*.(2023), shows that, the model proposed by (*Matthew et al*., 2023) performs better in estimating the population size in terms of accuracy as indicated by the smallest value of *RMSE=1193*.2053.

4. Conclusion

This study, applied a Generalized Estimating Equation (GEE) approach accounting for heterogeneity, behavioural effects and environmental effects to estimate the population size for open population capture-recapture experiment is proposed. A non-spatial open population capture-recapture data was simulated using three different trapping occasions *(m = 4, 5, 6)* with population size *(N =1000)*. The GEE estimates for behavioural/individual heterogeneity covariates and behavioural/heterogeneity/environmental covariates model under different working correlation was also presented. The study results from the simulations, show that, the comparison between $p_{ij} = h(\beta_0 + \beta_{sex} sex_i + \beta_{wt} wt_i + \beta_{bh} v_{ij})$ by Akanda and Alpizar- Jara (2014) and $p_{ij} = h(\beta_0 + \beta_{sex} sex_i + \beta_{wt} wt_i + \beta_{bh} v_{ij} + \beta_{tp} f_{ij})$ by Matthew *et al*.(2023), shows that, the model proposed by (Matthew *et al*., 2023) performs better in estimating the population size in terms of accuracy as indicated by the smallest value of *RMSE*. The study concludes that; environment plays important role when estimating the population size of open population capture-recapture experiments.

References

Akanda, M.A.S., and Alpizar-Jara, R. (2017). A generalized Estimating Equations Approach to model heterogeneity and Time dependence in Capture Recapture Studies. *European Journal of Ecology, 3*(1), 9–17.

Koivuniemi, M., Kurkilahti, M., Niemi, M., Auttila, M., and Kunnasranta, M. (2019). A mark–recapture approach for estimating population size of the endangered ringed seal (*Phoca hispida saimensis*). *PLoS ONE, 14*(3), 64-82.

Chowdhury, R.I., and Islam, M.A. (2017). Generalized Estimating Equation. In: Analysis of Repeated Measures Data. Springer, Singapore.

Akanda, M.A.S., and 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., and Alpizar-Jara, R. (2017). A generalized Estimating Equations Approach to model heterogeneity and Time dependence in Capture Recapture Studies. *European Journal of Ecology*, *3*(1), 9–17.

Chao, L., Ye, S., Qian, X., Stephen, L.R., Hui, H., and Yongtao, G. (2022). Mean corrected generalized estimating equations for longitudinal binary outcomes with report bias *Statistical Methods in Medical Research*. *31*(2) 315–333.

Wesson, P., Reingold, A,, and Mc Farland, W. (2017). Theoretical and empirical comparisons of methods to estimate the size of hard-to-reach populations: a systematic review. *AIDS Behavoiur 45*(3),415-440.

Worthington, H., Rachel S., Mc Crea,, Ruth, K., and Griffiths, R.A. (2018).Estimation of Population Size When Capture Probability Depends on Individual States. *Journal of Agricultural, Biological, and Environmental Statistics*, *24*(1), 154–172.

Mamadou, Y.O.I., Louis., P.R., and Greg, R. (2019). Capture-Recapture Methods for Data on the Activation of Applications on Mobile Phones. *Journal of the American Statistical Association*. *114*(525), 105-114.

Kevin, A., Reena, H. D., and Hladik, W. (2017). Capture-Recapture Among Men Who Have Sex with Men and Among Female Sex Workers in 11 Towns in Uganda. *Creative Commons Open Access Journal*.

Kreshpaj, B., Bodin, T., Wegman, D.H., Matilla, S.N., Burstrom, B., Kjellberg, K., *et al.* (2021). Under-reporting of non-fatal occupational injuries among precarious and non-precarious workers in Sweden. *Occup Environ Med*,*79*(1),3-9.

You, Y., Laan M Van Der, Collender, P., Cheng, Q., Hubbard, A., Jewell, N.P., Hu, Z.T., Mejia, R., and Remais, J.(2021). Estimation of population size based on capture-recapture designs and evaluation of the estimation reliability. ArXiv.2105.05373v1[maths.ST]

Nguyen, L.T., Patel, S., Nguyen, N.T., Gia, H.H., Raymond, H.F., Hoang, V.T.H., *et al*. (2021). Population size estimation of female sex workers in Hai Phong, Vietnam:use of three source capture–recapture method. *J Epidemiol Glob Health*,*11*(2),194–199.

Okiria, A.G., Bolo, A., Achut, V., Arkangelo, G.C., Michael, A.T.I., Katoro, J.S., et.al. (2019). Novel approaches for estimating female sex worker population size in conflict affected South Sudan. JMIR *Public Heal Surveill*,5(1),1–8.

Khandu, L., Tobgay, T., Kinley, K., Choida, N., Tashi, T., Wangdi, T., *et al*. (2021). Characteristics and population size estimation of female sex workers in Bhutan. *Sex Transm Dis*,*48*(10),754–60.

Guure, C., Dery, S., Afagbedzi, S., Tun, W., Weir, S.S., Quaye, S., *et al*. (2021). National and subnational size estimation of female sex workers in Ghana 2020: comparing 3-source capture-recapture with other approaches. *PLoS ONE*, 16(9),1–16.

Huang, Y., and Pan, J. (2021). Joint generalized estimating equations for longitudinal binary data. *Computational Statistics and Data Analysis*, ***155***, 107110. https://doi.org/10.1016/j.csda.2020.107110

Wesson, P., Jewell, N.P., Willi McFarland., M.D., and Glymour., M.M. (2023). Evaluating tools for capture-recapture model selection to estimate the size of hidden populations: It works in practice, but does it work in theory? *Annals of Epidemiology*, *77*(4), 24–30.

Zhuoran, W.E., Hanbing, Z., Sharon, C., Gary, C., and Molin, W. (2023). Generalized Estimating Equations for Hearing Loss Data with Specified Correlation Structures *Stat.ME*

Nasrin, L., Mohammad, S.A., and Syed, S. H. (2021). A Generalized Estimating Equations Approach for Modeling Spatially Clustered Data. *Austrian Journal of Statistics*: *50*(36),36-52.

Goldenberg, S.M., Buglioni, N., Krusi, A., Frost, E., Moreheart, S., Braschel, M., and Shannon, K. (2023). Housing Instability and Evictions Linked to Elevated Intimate Partner and Workplace Violence Among Women Sex Workers in Vancouver, Canada: Findings of a Prospective, Community-Based Cohort, 2010–2019. *American ournal of public health*. *113*(4), 442-452.

Asep, S. A., Inge, W., and Hilda, N. (2020). Analysis of Longitudinal Regression Model Using the Generalized Estimating Equation (GEE) for the Child Welfare Composite Index (CWCI) in West Java. *Advances in Social Science, Education and Humanities Research, volume 550.* Proceedings of the 1st International Conference on Mathematics and Mathematics Education (ICMMED), 1-8.

Chao, L., Ye, S., Qian, X., Stephen, L.R., Hui, H., and Yongtao, G. (2021). Mean corrected generalized estimating equations for longitudinal binary outcomes with report bias *Statistical Methods in Medical Research*. *31*(2) 315–333.

Zhang, S. (2012). A GEE approach for estimating size of hard-to-reach population by using capture-recapture data. *Journal of Theoretical and Applied Statistics*. *46*(2), 175–183.

Pan, W. (2001). Akaike's Information Criterion in Generalized Estimating Equations. *Biometrics, 57*(1), 120-125.

Wentao, G.(2019). Bootstrap-Adjusted Quasi-Likelihood Information Criteria for Mixed Model Selection. PHD Dissertation, Bowling Green State University.

Wang, Y., Orla, M., Wang, Z., Bhatnagar, S.R., Schultz, J., and Moodie, E., M. (2015). The Perils of Quasi-Likelihood Information Criteria, *Stat Journal*, *4*(2), 246-254.

McDonald, T. L., and Amstrup, S. C. (2001). Estimation of population size using open capture-recapture models. *Journal of Agricultural, Biological, and Environmental Statistics*, *6*(2), 206-220.

Matthew, P.K., Jibasen, D., and Dike, I.J. (2023). Estimating the Population Parameters for Open Population Capture-Recapture Data Using Generalized Estimating Equations Approach. Proceedings of the American University of Nigeria 1st International Conference. *1*(1);498-516.