

Population size estimation and heterogeneity in capture–recapture data: a linear regression estimator based on the Conway–Maxwell–Poisson distribution

Orasa Anan¹ · Dankmar Böhning¹ · Antonello Maruotti^{2,3}

Accepted: 16 March 2016 / Published online: 18 April 2016
© Springer-Verlag Berlin Heidelberg 2016

Abstract The purpose of the study is to estimate the population size under a truncated count model that accounts for heterogeneity. The proposed estimator is based on the Conway–Maxwell–Poisson distribution. The benefit of using the Conway–Maxwell–Poisson distribution is that it includes the Bernoulli, the Geometric and the Poisson distributions as special cases and, furthermore, allows for heterogeneity. Parameter estimates can be obtained by exploiting the ratios of successive frequency counts in a weighted linear regression framework. The results of the comparisons with Turing’s, the maximum likelihood Poisson, Zelterman’s and Chao’s estimators reveal that our proposal can be beneficially used. Furthermore, our proposal outperforms its competitors under all heterogeneous settings. The empirical examples consider the homogeneous case and several heterogeneous cases, each with its own features, and provide interesting insights on the behavior of the estimators.

Keywords Capture–recapture methods · Truncated-at-zero count data · Ratio plot · Heterogeneous populations

✉ Antonello Maruotti
a.maruotti@lumsa.it

Orasa Anan
oa2e12@soton.ac.uk

Dankmar Böhning
D.A.Bohning@soton.ac.uk

¹ School of Mathematics and Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, UK

² Dipartimento di Scienze Economiche, Politiche e delle Lingue Moderne, Libera Università Maria Ss. Assunta, Via Pompeo Magno 22, 001921 Roma, Italy

³ Centre for Innovation and Leadership in Health Sciences, University of Southampton, Southampton, UK

1 Introduction

Capture–recapture (CR) methods have been adopted in a wide range of applications, including ecology (Alunni-Fegatelli and Tardella 2013; Farcomeni 2011), epidemiology (Böhning et al. 2005), criminal activity (van Der Heijden et al. 2003; Farcomeni and Scacciatelli 2013), official statistics (Rocchetti et al. 2011; Gerritse et al. 2015) and, in general, in the estimation of the size of hidden populations. A recent review can be found in McCrea and Morgan (2014). CR analyses are based on the repeated sampling from a population and, consequently, on the use of recapture information to infer the number of uncaptured units. Throughout the paper, we consider the following CR setting. The target population is sampled over a certain number of capture occasions, and for each occasion, captured units are counted only once. Moreover, we consider a closed population, i.e. the unknown population size, say N , is assumed to be constant (with no births/deaths during sampling stages), missclassification is not allowed and all units act independently.

Formally, let X_i , $i = 1, \dots, N$ denote the number of times unit i is captured over the m sampling occasions, and let $p_x = \Pr(X_i = x)$. Also let f_x denote the frequency of units captured exactly x times, $x = 0, 1, \dots, m$. As $X_i = 0$ is not observed, the corresponding f_0 is unknown and might be replaced by its expected value Np_0 . Nevertheless, p_0 is usually unknown too and has to be estimated. As X_i takes only non-negative integer values, the Poisson model with parameter λ may represent a natural starting point. Clearly, this model is restrictive because it assumes a unit variance-to-mean ratio. Hence, even if the Poisson distribution can be recognized as an important tool to model count data, it may not be suitable for CR data, which are characterized by overdispersion/underdispersion, i.e. the variance is greater/lower than the corresponding sample mean, mainly due to unobserved heterogeneity (see e.g. Baksh et al. 2011). To account for heterogeneity in the estimation of the population size, the Poisson parameter is often considered as an unobserved random variable with a latent distribution $\lambda(t)$ (Chao 1987). Accordingly, the marginal distribution is provided as

$$p_x(\lambda) = \int_0^\infty \frac{\exp(-t)t^x}{x!} \lambda(t) dt \quad (1)$$

where the mixing distribution density $\lambda(t)$ is unknown. One way to model overdispersion is to consider the Gamma–Poisson mixture, where Poisson variables have means that follow a Gamma distribution. This yields a Negative Binomial marginal distribution. However, in the CR framework, the Negative Binomial distribution might not be appropriate as constraints on the dispersion parameter might lead to unrealistic estimates of f_0 and, moreover, it is limited to model over-dispersed data only, i.e. it is unable to fit under-dispersed data. Thus, to mitigate the potential bias in population size estimation due to heterogeneity, discrete (Pledger 2005; Bartolucci and Forcina 2006; Morgan and Ridout 2008) and continuous (Dorazio and Royle 2003; Niwitpong et al. 2013; Rocchetti et al. 2014) mixing distributions have been used.

We wish to contribute extending this branch of literature by proposing a more general count distribution that captures a wider range of dispersion settings than existing distributions. In detail, we look at a two-parameter generalized form of the Poisson

distribution, called the Conway–Maxwell–Poisson (CMP) distribution (Shmueli et al. 2005) to account for heterogeneity as it includes as special sub-models important distributions (i.e. the Poisson, the Bernoulli and geometric distributions) and generalizes the Poisson distribution allowing for overdispersion as well as underdispersion.

In the following, we will exploit heterogeneity (in the number of times a unit is captured) through a graphical device, namely the ratio-plot (Böhning et al. 2013). The ratio-plot is a graphical method for identifying the form of the heterogeneity distribution in CR data. In particular, it assesses if the homogeneous Poisson is appropriate or whether (or not) heterogeneity arises in the observed data. Furthermore, in this work we aim at extending the usefulness of the ratio-plot beyond its descriptive nature. Indeed, we will use the ratio-plot as a tool to obtain the estimate of p_0 (and, accordingly, of f_0 and N) in a heterogeneous population. We introduce a regression estimator based on the (log) ratio-plot which provides straightforward parameter estimates and derive its asymptotic variance. Formally, we use the relationship between the (log) ratios of successive capture probabilities to estimate model parameters through a (weighted) linear regression approach.

We illustrate the proposal by a large-scale simulation study in order to investigate the empirical behaviour of the proposed distribution with respect to several factors, such as the population size, the mixing distribution and the number of occasions (modelled by varying the mean of the count variable). To show the practical usefulness of the CMP regression, we compare its performance to a few alternative estimators, widely used in the CR framework. Finally, we test the proposal by analysing several real datasets. The outline of the paper is as follows. In Sect. 2, we specify the proposed model, along with the ratio-plot and the computational aspects of the adopted maximum likelihood regression-based algorithm. Properties of the proposed estimator are investigated in depth, and its asymptotic variance is derived. Furthermore, we summarize alternative population size estimators. In Sect. 3, we give a comparison of the performance of several model specifications under different data generation schemes by means of a simulation study. In Sect. 4, we present several real-data analyses. In Sect. 5, we point out some remarks, along with drawbacks that may arise by adopting the proposed methodology.

2 The Conway–Maxwell–Poisson distribution for capture–recapture data

2.1 The Conway–Maxwell–Poisson distribution

The CMP distribution, as an extension of the Poisson distribution, is a flexible model for analyzing count data, although it has been used less frequently as other generalizations. As discussed by Shmueli et al. (2005), the CMP distribution generalized the Poisson distribution allowing for under-dispersion as well as over-dispersion. Its probability mass function $\text{CMP}(\lambda, \nu)$ is given by

$$p_x = \frac{\lambda^x}{(x!)^\nu} \frac{1}{z(\lambda, \nu)}, \quad x = 0, 1, 2, \dots; \lambda > 0; \nu \geq 0$$

where the normalizing constant

$$z(\lambda, \nu) = \sum_{j=0}^{\infty} \frac{\lambda^j}{(j!)^\nu}$$

is a generalization of well-known infinite sums. The CMP distribution has been overlooked for long time due to the complexity in dealing with the infinite sum $z(\lambda, \nu)$, that is often approximated.

The case $\nu = 1$ corresponds to the Poisson distribution, as $z(\lambda, \nu) = e^\lambda$. For $\nu \rightarrow \infty$, the CMP distribution approaches the Bernoulli distribution with parameter $\lambda(1+\lambda)^{-1}$. We would like to point out that, with $\nu = 0$ and $0 < \lambda < 1$, $z(\lambda, \nu) = \frac{1}{1-\lambda}$ and, accordingly, the CMP distribution reduces to the geometric distribution with parameter $(1 - \lambda)$. At last, note that for $\nu = 0$ and $\lambda \geq 1$, $z(\lambda, \nu)$ does not converge, leading to an undefined distribution.

To complete the description on the CMP distribution, let us specify its moments by using an asymptotic approximation of $z(\lambda, \nu)$, as described in [Shmueli et al. \(2005\)](#),

$$E(X) \approx \lambda^{1/\nu} + \frac{1}{2\nu} - \frac{1}{2}$$

$$V(X) \approx \frac{1}{\nu} \lambda^{1/\nu}.$$

Using the [Guikema and Coffelt \(2008\)](#) specification, the dispersion can be written as

$$D(X) = \frac{V(X)}{E(X)} \approx \frac{\frac{\mu}{\nu}}{\mu + \frac{1}{2}\nu - \frac{1}{2}} \approx \frac{1}{\nu},$$

with $\mu = \lambda^{1/\nu}$. When $\nu < 1$, the variance can be shown to be greater than the mean and the dispersion > 1 . This is a result of overdispersed data. When $\nu = 1$, and the mean and variance are equal, the dispersion is equal to 1 (Poisson model). When $\nu > 1$, the variance is smaller than the mean and the dispersion is < 1 .

2.2 The ratio-plot

In this work, we avoid classical approaches to estimation of population size (see e.g. [Lindsay and Roeder 1987](#); [Böhning et al. 2005](#); [Bunge and Barger 2008](#)) and propose a method based on ratios of successive probability counts, namely,

$$r_x = (x + 1) \frac{p_{x+1}}{p_x}$$

which is a function of the observed count x .

In CR studies, the zero counts are truncated and, hence, the observed sample frequencies f_1, f_2, \dots arise from the zero-truncated distribution $\frac{p_x}{1-p_0}$. However, the ratio r_x for the truncated and the untruncated distribution is identical

$$r_x = (x + 1) \frac{p_{x+1}}{p_x} = (x + 1) \frac{p_{x+1}/(1 - p_0)}{p_x/(1 - p_0)}.$$

This is an important result as it makes the ratio applicable into a CR framework. The ratio for the CMP distribution has the following form

$$r_x = (x + 1) \frac{p_{x+1}}{p_x} = (x + 1) \frac{\frac{\lambda^{x+1}}{\{(x+1)!\}^v} \frac{1}{z(\lambda, \nu)}}{\frac{\lambda^x}{(x!)^v} \frac{1}{z(\lambda, \nu)}} = \lambda(x + 1)^{1-\nu} \tag{2}$$

and does not depend on the complex normalizing constant term $z(\lambda, \nu)$. Equation (2) suggests a non-linear relation between the ratio of successive probabilities and the count x . However, if we consider the ratio on the log-scale, we achieve a linear relationship. Accordingly,

$$\begin{aligned} \log(r_x) &= \log \left\{ (x + 1) \frac{p_{x+1}}{p_x} \right\} = \log\{\lambda(x + 1)^{1-\nu}\} \\ &= \log \lambda + (1 - \nu) \log(x + 1) = \beta_0 + \beta_1 \log(x + 1). \end{aligned} \tag{3}$$

From (3), we have that $\lambda = \exp(\beta_0)$ and $\nu = 1 - \beta_1$; however, due to $\nu \geq 0$ (or, equivalently, $1 - \nu \leq 1$), we must constrain $\beta_1 \leq 1$. There are no restrictions on β_0 , $\lambda > 0$ implies $\beta_0 \in (-\infty, +\infty)$.

In practice, we approximate capture probabilities by relative frequencies, therefore the ratio in (2) can be obtained by

$$r_x^* = (x + 1) \frac{\hat{p}_{x+1}}{\hat{p}_x} = (x + 1) \frac{f_{x+1}/N}{f_x/N} = (x + 1) \frac{f_{x+1}}{f_x},$$

as well as the ratio in (3) can be computed as

$$\log(r_x^*) = \log \left\{ (x + 1) \frac{f_{x+1}}{f_x} \right\},$$

where f_x is the frequency of count x and $N = \sum_{x=0}^m f_x$.

By plotting $\log(r_x^*)$ against $\log(x + 1)$, we derive a graphical diagnostic tool for detecting the validity of Conway–Maxwell–Poisson model. The resulting plot is known as the *log-ratio plot* (see Böhning et al. 2013 for further details). A log-ratio plot showing a positive slope indicates for the presence of overdispersion with respect to the Poisson distribution. On the other hand, in the case of underdispersion, the log-ratio plot displays a straight line with a negative slope. Finally, when the log-ratio plot displays a horizontal line, the equi-dispersion case is plausible, or, in other words, the Poisson distribution can be used to fit the data.

Other distributions as the Negative Binomial have been often considered to deal with heterogeneity. It has also a straight line behaviour when plotting ratios of successive capture probabilities against x , but fitting parameters has frequently boundary issues.

Taking $\log(r_x^*)$ helps but this destroys the linear characteristic with respect to x . Hence there are benefits of CMP distribution in comparison to the Negative Binomial as well.

2.3 Model inference

The use of the ratio in (3) goes beyond a simple graphical technique to check for under/over-dispersion in CR data. Indeed, it can be used as a tool for estimating model's parameters. Thus, let us consider our basic Eq. (3), we fit the following model

$$\log(r_x^*) = \underbrace{\beta_0 + \beta_1 \log(x + 1)}_{\text{Systematic}} + \underbrace{\epsilon_x}_{\text{Random}}, \quad (4)$$

where β_0 and β_1 are the intercept and the slope parameters respectively, and ϵ_x is the error term.

Commonly, a least square estimation (LS) method is used to provide estimates of β_0 and β_1 . However, model (4) does not satisfy the classical linear regression assumptions. In the first place, the response is discrete (although log-transformed), so we might consider a generalized linear model. However, this is inadvisable since an appropriate formulation as a generalized linear model leads to an autoregressive equation involving $\log f_x$ as an additional offset term in the linear predictor. These kinds of models experience difficulties in terms of the definition of the likelihood as well as in carrying out inference. Furthermore, CR frequencies often have $f_1 \gg f_2 > f_3 > \dots$, and, additionally, heteroskedasticity might occur in heterogeneous population due to e.g. unobserved information (see e.g. Rocchetti et al. 2014). All these issues are relevant and should be accounted for. Thus, we address them by using weighted least squares (WLS) techniques to estimate the regression parameters β_0 and β_1 , and accordingly λ and ν . These are obtained by minimising

$$\sum_{x=1}^{m-1} W_x [\log(r_x^*) - \beta_0 - \beta_1 \log(x + 1)]^2,$$

where W_x denotes the x -th element of an appropriate weight matrix. In other words, we take

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1} \mathbf{X}'\mathbf{W}\mathbf{Y}, \quad (5)$$

where

$$\mathbf{Y} = \begin{pmatrix} \log \frac{2f_2}{f_1} \\ \log \frac{3f_3}{f_2} \\ \vdots \\ \log \frac{mf_m}{f_{m-1}} \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & \log(2) \\ 1 & \log(3) \\ \vdots & \vdots \\ 1 & \log(m) \end{pmatrix}$$

and m is the maximum count used in the estimator.

The application of weighted least square requires the specification of $\mathbf{W} \approx cov(\mathbf{Y})^{-1}$ to reduce the mean square error. Following Meurant (1992) and Rocchetti et al. (2011), covariances between adjacent log-ratios do not play a large role in reducing mean square error, and thus we suggest to drop the off-diagonal terms in $cov(\mathbf{Y})$ in approximating \mathbf{W} , with little loss of efficiency. Accordingly

$$\mathbf{W} = \begin{bmatrix} \frac{1}{f_1} + \frac{1}{f_2} & 0 & \cdots & 0 \\ 0 & \frac{1}{f_2} + \frac{1}{f_3} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \frac{1}{f_{m-1}} + \frac{1}{f_m} \end{bmatrix}^{-1} \tag{6}$$

To see that (6) is the right choice, let $\mathbf{W}_x = [Var\{\log(r_x^*)\}]^{-1}$, we have

$$\begin{aligned} Var\{\log(r_x^*)\} &= Var\left[\log\left\{(x+1)\frac{\hat{p}_{x+1}}{\hat{p}_x}\right\}\right] \\ &= Var\{\log(x+1) + \log(\hat{p}_{x+1}) - \log(\hat{p}_x)\} \\ &= Var\{\log(\hat{p}_{x+1})\} + Var\{\log(\hat{p}_x)\} - 2Cov\{\log(\hat{p}_{x+1}), \log(\hat{p}_x)\}. \end{aligned}$$

Using the delta method

$$\begin{aligned} Var\{\log(r_x^*)\} &\approx \frac{1}{p_{x+1}^2} Var(\hat{p}_{x+1}) + \frac{1}{\hat{p}_x^2} Var(\hat{p}_x) - \frac{2Cov(\hat{p}_{x+1}, \hat{p}_x)}{\hat{p}_{x+1}\hat{p}_x} \\ &= \frac{1}{p_{x+1}^2} \left\{ \frac{\hat{p}_{x+1}(1 - \hat{p}_{x+1})}{n} \right\} + \frac{1}{\hat{p}_x^2} \left\{ \frac{\hat{p}_x(1 - \hat{p}_x)}{n} \right\} + \frac{2\hat{p}_{x+1}\hat{p}_x}{\hat{p}_{x+1}\hat{p}_x} \\ &= \frac{1 - \hat{p}_{x+1}}{n\hat{p}_{x+1}} + \frac{1 - \hat{p}_x}{n\hat{p}_x} + \frac{2}{n} \\ &= \frac{1}{n\hat{p}_{x+1}} - \frac{\hat{p}_{x+1}}{n\hat{p}_{x+1}} + \frac{1}{n\hat{p}_x} - \frac{\hat{p}_x}{n\hat{p}_x} + \frac{2}{n} \end{aligned}$$

where n is the number of observations from the target population.

Therefore, the variance of log-ratio is given by

$$Var\{\log(r_x^*)\} \approx \frac{1}{n\hat{p}_{x+1}} + \frac{1}{n\hat{p}_x}.$$

In practice, \hat{p}_{x+1} and \hat{p}_x can be estimated by relative observed frequency $\frac{f_{x+1}}{n}$ and $\frac{f_x}{n}$, respectively. Hence

$$\widehat{Var}\{\log(r_x^*)\} = \frac{1}{n\frac{f_x}{n}} + \frac{1}{n\frac{f_{x+1}}{n}} = \frac{1}{f_x} + \frac{1}{f_{x+1}}.$$

Thus, we get $\hat{\beta}_0$ and $\hat{\beta}_1$ from (5), in which \mathbf{W} is given by (6). Accordingly, the unknown f_0 can be then estimated by considering that

$$\begin{aligned}\log\left(\frac{f_1}{f_0}\right) &= \hat{\beta}_0 \\ \frac{f_1}{f_0} &= \exp(\hat{\beta}_0) \\ \hat{f}_0 &= f_1 \exp(-\hat{\beta}_0),\end{aligned}$$

where \hat{f}_0 is the unobserved frequency estimator. The linear regression estimator based on the Conway–Maxwell–Poisson distribution (LCMP) of target population size can be readily achieved as

$$\hat{N}_{LCMP} = n + \hat{f}_0 = n + f_1 \exp(-\hat{\beta}_0). \quad (7)$$

We also obtain an estimated probability of the count to be zero (unobserved) as

$$\hat{p}_0 = \hat{f}_0 / \hat{N}_{LCMP}.$$

We anticipate that \hat{N}_{LCMP} is asymptotically unbiased in the sense

$$\frac{E(\hat{N}_{LCMP})}{N} \rightarrow_{N \rightarrow \infty} 1,$$

if the sample arises from a Conway–Maxwell–Poisson distribution. The rationale for this is as follows. Suppose that β_0 would be known, then

$$\hat{N}_{LCMP} = n + f_0 = n + f_1 \exp(-\beta_0) \quad (8)$$

is unbiased as

$$\begin{aligned}E(\hat{N}_{LCMP}) &= N(1 - p_0) + p_1 N \exp(-\beta_0) = N[(1 - p_0) + p_1/\lambda] \\ &= N[(1 - p_0) + p_0] = N.\end{aligned}$$

For any matrix \mathbf{W} , the weighted least squares estimate in (5) is unbiased if \mathbf{W} is non-random, as

$$E\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1} \mathbf{X}'\mathbf{W}\mathbf{X} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}.$$

However, an efficient estimator is achieved only if $\mathbf{W} = \boldsymbol{\Sigma}^{-1}$, where $\boldsymbol{\Sigma}$ is the true variance-covariance matrix of \mathbf{Y} . If an estimator $\hat{\boldsymbol{\Sigma}}$ of $\boldsymbol{\Sigma}$ is used (as it is often the case in practice and also in our situation), efficiency is usually lost, but not asymptotic unbiasedness. For the latter, only a consistent estimate of $\boldsymbol{\Sigma}$ is needed. This is the case

for our situation. It is shown in Rocchetti et al. (2011) that using the weight-matrix in (6) leads to a gain in efficiency in comparison with the unweighted unbiased estimate

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y}.$$

Hence, we prefer to use (5) with weight matrix (6).

It is clear that some attention has to be paid to the fact that weights are estimated in reality and this is further addressed in the simulation study. We point out here that the Conway–Maxwell–Poisson distribution includes as a special case the geometric ($\nu = 0$) so that an associated weighted least-squares estimator is available for the geometric. It has the simple form $\widehat{\log \lambda} = \left(\sum_{x=1}^{m-1} W_x \log \frac{f_{x+1}}{f_x} \right) / \left(\sum_{x=1}^{m-1} W_x \right)$, where W_x is the x -th diagonal element of (6).

2.4 Variance estimation and confidence interval

Let \hat{N} be the population size estimator, according to Böhning (2008), the variance of $\hat{N}_{LCMP} = n + f_1 e^{-\hat{\beta}_0}$ arises from two sources; these are influenced by the random variable n and the estimator \hat{f}_0 . Using conditional moment techniques, a formula for the variance of population size estimator is given as:

$$Var(\hat{N}) = Var_n\{E(\hat{N}|n)\} + E_n\{Var(\hat{N}|n)\}, \tag{9}$$

where E_n and Var_n refer to the first and the second moment of the marginal distribution under observed data n . It is

$$E(\hat{N}|n) \approx n + \hat{f}_0,$$

with \hat{f}_0 non random, so that

$$Var_n\{E(\hat{N}|n)\} = \widehat{Var}_n\{n + \hat{f}_0\} = \widehat{Var}_n\{n\} = N(1 - p_0)p_0, \tag{10}$$

where the latter follows from the fact that $n \sim Binomial(N, 1 - p_0)$.

Since $E(n) = N(1 - p_0)$ and $p_0 = E(f_0/N)$, leading to $\hat{p}_0 = \frac{\hat{f}_0}{n + \hat{f}_0}$, we have that (10) can be estimated by

$$\widehat{Var}_n\{E(\hat{N}|n)\} = \frac{n \hat{f}_0}{n + \hat{f}_0} = \frac{n f_1 e^{-\hat{\beta}_0}}{n + f_1 e^{-\hat{\beta}_0}}. \tag{11}$$

Also, we assume as $E_n\{Var(\hat{N}|n)\}$ can be estimated by $Var(\hat{N}|n) = Var(\hat{f}_0) = Var\{f_1 e^{(-\hat{\beta}_0)}\}$, hence we have that

$$E_n\{Var(\hat{N}|n)\} = \widehat{Var}\{f_1 e^{(-\hat{\beta}_0)}\}. \tag{12}$$

By the conditional technique,

$$\text{Var}(f_1 e^{-\hat{\beta}_0}) = \text{Var}_{f_1}\{E(f_1 e^{-\hat{\beta}_0})|f_1\} + E_{f_1}\{\text{Var}(f_1 e^{-\hat{\beta}_0})|f_1\}, \quad (13)$$

thus

$$\begin{aligned} \text{Var}_{f_1}\{E(f_1 e^{-\hat{\beta}_0})|f_1\} &\approx \text{Var}(f_1 e^{-\hat{\beta}_0}) = (e^{-\hat{\beta}_0})^2 \text{Var}(f_1) \\ &= (e^{-\hat{\beta}_0})^2 N p_1 (1 - p_1) = (e^{-\hat{\beta}_0})^2 f_1 \left(1 - \frac{f_1}{N}\right), \end{aligned} \quad (14)$$

as well as, $E_{f_1}\{\text{Var}(f_1 e^{-\hat{\beta}_0})|f_1\}$ can be estimated by $\text{Var}\{(f_1 e^{-\hat{\beta}_0})|f_1\}$, so that

$$E_{f_1}\{\text{Var}(f_1 e^{-\hat{\beta}_0})|f_1\} \approx \text{Var}\{(f_1 e^{-\hat{\beta}_0})|f_1\} = f_1 \text{Var}(e^{-\hat{\beta}_0}). \quad (15)$$

Using the delta method, we achieved that $\text{Var}(e^{-\hat{\beta}_0}) = (e^{-\hat{\beta}_0})^2 \text{Var}(\hat{\beta}_0)$. Hence $E_{f_1}\{\text{Var}(f_1 e^{-\hat{\beta}_0})|f_1\} \approx f_1^2 (e^{-\hat{\beta}_0})^2 \text{Var}(\hat{\beta}_0)$, where $\text{Var}(\hat{\beta}_0)$ comes from the linear regression process. The approximated expression for the variance of the new estimator \hat{N}_{LCMP} is given as

$$\widehat{\text{Var}}(\hat{N}_{LCMP}) = \frac{n f_1 e^{-\hat{\beta}_0}}{n + f_1 e^{-\hat{\beta}_0}} + (e^{-\hat{\beta}_0})^2 f_1 \left(1 - \frac{f_1}{N}\right) + f_1^2 (e^{-\hat{\beta}_0})^2 \text{Var}(\hat{\beta}_0). \quad (16)$$

Finally, when N is large, the asymptotic variance estimator of \hat{N}_{LCMP} is

$$\widehat{\text{Var}}(\hat{N}_{LCMP}) = \frac{n f_1 e^{-\hat{\beta}_0}}{n + f_1 e^{-\hat{\beta}_0}} + (e^{-\hat{\beta}_0})^2 f_1 [1 + f_1 \text{Var}(\hat{\beta}_0)]. \quad (17)$$

A confidence level refers to the percentage of all possible samples that can be expected to include the true value of population size N . We used 95% confidence level to imply 95% of the confidence intervals, including the true population size estimator. It is simply to construct 95% confidence interval of N under the assumption that population distribution be approximately normal as:

$$\hat{N} \pm z_{0.975} SE(\hat{N}), \quad (18)$$

where $SE(\hat{N})$ denotes the standard error of \hat{N} , approximated by the asymptotic standard error; $\widehat{SE}(\hat{N}) = \sqrt{\widehat{\text{Var}}(\hat{N})}$, and $z_{0.975} = 1.96$.

2.5 Alternative estimators

Several estimators have been applied to estimate population size in CR data. This section focuses on well-known estimators based on homogeneous Poisson and heterogeneous models. Turing's estimator and the maximum likelihood estimator under

a Poisson model are considered as estimators in the homogeneous case. Estimators for heterogeneous populations as the Zelterman's estimator and the Chao's lower bound estimator are considered as well. In the simulation study and in the application section, estimator performances are compared with the LCMP estimator under several settings.

2.5.1 Turing's estimator

The application of Turing's estimator can be used under a homogeneous Poisson distribution. Then in terms of a homogeneous Poisson distribution with parameter λ we have

$$p_0 = e^{-\lambda} = \frac{\lambda e^{-\lambda}}{\lambda} = \frac{p_1}{E(X)} = \frac{E(f_1)/N}{E(S)/N} = \frac{E(f_1)}{E(S)}, \quad (19)$$

where $p_1 = \lambda e^{-\lambda}$ and $S = \sum_{x=1}^m f_x$. Replacing these expected values by their observed quantities we have

$$\hat{p}_0 = \frac{f_1}{S}. \quad (20)$$

We achieve Turing's estimator as

$$\hat{N}_{Turing} = \frac{n}{1 - f_1/S}. \quad (21)$$

The variance for Turing estimation (Lerdsuwansri 2012) is given by

$$\widehat{Var}(\hat{N}_{Turing}) = \frac{n \frac{f_1}{S}}{(1 + \frac{f_1}{S})^2} + \frac{n^2}{(1 + \frac{f_1}{S})^4} \left[\frac{f_1(1 - \frac{f_1}{S})}{S^2} + \frac{f_1^2}{S^3} \right]. \quad (22)$$

The benefits of Turing's estimator are that it is easy to calculate, its value can be obtained in a straightforward way, and there is no need for an iterative procedure. In addition, it uses all the information in the sample by means of S and f_1 , the latter being usually large.

2.5.2 Maximum likelihood estimator under the zero-truncated Poisson distribution

Let us assume that the capture–recapture count data X can be modelled as a zero-truncated Poisson distribution. Thus, population size can be estimated as

$$\hat{N}_{MLE} = \frac{n}{1 - \exp(-\hat{\lambda}_{MLE})}, \quad (23)$$

The maximum likelihood estimator $\hat{\lambda}_{MLE}$ can be obtained by using the EM-algorithm technique under the zero-truncated Poisson distribution (see Böhning et al. 2005). A simple variance estimate of (23) is given as

$$\widehat{\text{Var}}(\hat{N}_{MLE}) = \frac{\hat{N}_{MLE}}{\left\{ \exp\left(\frac{\sum_{x=1}^m x f_x}{\hat{N}_{MLE}}\right) - \frac{\sum_{x=1}^m x f_x}{\hat{N}_{MLE}} - 1 \right\}}. \quad (24)$$

2.5.3 Zelterman's estimator

Zelterman (1988) suggested an estimator under a truncated Poisson sampling estimator. This is a well-known robust estimator under potential unobserved heterogeneity. Zelterman suggests using $\hat{\lambda}_1 = \frac{2f_2}{f_1}$. As Kuhnert and Böhning (2009) pointed out, there are two reasons for choosing $\hat{\lambda}_1$ in Zelterman's parameter. Firstly, the majority of frequency count units are usually represented in terms of counts of once and twice (f_1 and f_2 are used), so that count data greater than two are likely to have little effect on this estimator. Secondly, $\hat{\lambda}_1$ is the closest neighbour of the target point of estimation f_0 . The Zelterman estimator of population size is ultimately provided as

$$\hat{N}_{Zel} = \frac{n}{1 - \exp\left(-\frac{2f_2}{f_1}\right)}. \quad (25)$$

Zelterman's estimator has been widely used since it is easy to understand, and it is a robust estimator because it uses only the first and second order of frequencies. However, it might be not a good estimator for long tail count data (Lanumteang 2011). Also, it can overestimate the population size under heterogeneity (Böhning and Schön 2005). The Variance of the Zelterman's Estimator is

$$\widehat{\text{Var}}(\hat{N}_{Zel}) = nG(\hat{\lambda}) \left[1 + nG(\hat{\lambda})\hat{\lambda}^2 \left(\frac{1}{f_1} + \frac{1}{f_2} \right) \right] \quad (26)$$

where $G(\hat{\lambda}) = \frac{\exp(-\hat{\lambda})}{\{1 - \exp(-\hat{\lambda})\}^2}$ and $\hat{\lambda} = \frac{2f_2}{f_1}$ (see Böhning 2008).

2.5.4 Chao's lower bound estimator

Chao (1987, 1989) introduced an alternative estimator of population size under unobserved heterogeneity of the Poisson parameter. Counts are assumed to be generated from a mixed Poisson model with arbitrary mixing density $g(\lambda)$; $p_x = \int_0^\infty \frac{e^{-\lambda} \lambda^x}{x!} g(\lambda) d\lambda$ where $x = 0, 1, 2, \dots$. Based on the Cauchy-Schwarz inequality, we achieve a lower bound for p_0 as: $\frac{p_1^2}{2p_2} \leq p_0$, multiplying those probabilities by N leads to $\frac{(Np_1)^2}{2(Np_2)} \leq Np_0$. Hence replacing Np_1 and Np_2 with the observed frequencies f_1 and f_2 leads to the lower bound estimator $\frac{f_1^2}{2f_2}$, that is

$$\hat{f}_0 = \frac{f_1^2}{2f_2} \quad (27)$$

and

$$\hat{N}_{Chao} = n + \frac{f_1^2}{2f_2}. \quad (28)$$

Note that only f_1 and f_2 are used in Chao's lower bound estimator. An important lower bound for estimating the population size is given.

An estimate of the variance of the Chao's estimator is given by

$$\widehat{Var}(\hat{N}_{Chao}) = \frac{1}{4} \frac{f_1^4}{f_2^3} + \frac{f_1^3}{f_2^2} + \frac{1}{2} \frac{f_1^2}{f_2} - \frac{1}{4} \frac{f_1^4}{(nf_2^2)} - \frac{1}{2} \frac{f_1^4}{f_2(2nf_2 + f_1^2)} \tag{29}$$

(see Böhning 2008). An extended version of the Chao's estimator has been recently proposed in Chiu et al. (2014). It also contains a variance estimate for Chao's estimator as well as a confidence interval construction based upon the log-normal distribution.

3 Simulation study

This section provides a comprehensive assessment of population size estimator performance. We compare the LCMP estimator proposed in this work with other well-established estimators highlighted in the previous section. We plan the simulation study to cover schemes with different underlying *null* models, with varying population size $N = 100; 1000; 10,000$ and levels of heterogeneity.

In detail, we consider the following data generation settings

- (i) The Poisson distribution: counts are generated from the Poisson distribution with parameters

$$\lambda \in \{0.5, 1.0, 1.5, 2.0, 2.5, 3.0\}$$

- (ii) The geometric distribution: counts are generated from the geometric distribution with parameters

$$\lambda \in \{0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$$

where $\lambda = 1 - p$, and p is a probability of success.

- (iii) The Conway–Maxwell–Poisson distribution: counts are generated from Conway–Maxwell–Poisson distribution with parameters

$$\lambda \in \{0.5, 1.0, 1.5, 2.0, 2.5, 3.0\}$$

$$\nu \in \{0.4, 0.6, 0.8\}$$

- (iv) The Negative Binomial distribution: counts are generated from a Negative Binomial distribution

$$p_x = \frac{\Gamma(x + k)}{\Gamma(x + 1)\Gamma(k)} (1 - \lambda)^k \lambda^x,$$

with parameters

$$\lambda \in \{0.2, 0.4, 0.6, 0.8\},$$

dispersion parameters

$$k \in \{2, 4, 6\},$$

expected value and variance given respectively by

$$E(X) = \frac{k\lambda}{1-\lambda} = \mu$$

and

$$\text{Var}(X) = \frac{k\lambda}{(1-\lambda)^2} = \mu + \frac{1}{k}\mu^2.$$

As the settings (i)–(ii)–(iii) covers situations where the data generation is from a special case of the CMP distribution, we include setting (iv) to investigate what happens if we leave the family, e.g. if we sample from a Negative Binomial distribution. We draw $B = 1000$ samples from each *null* model. Any occurrences of zero counts were truncated, and five estimators of population size were compared: the Turing's estimator (Turing), the maximum likelihood estimation under the zero-truncated Poisson model (MLEPoi), Chao's lower bound estimator (Chao), Zelterman's estimator (Zel) and weighted linear regression estimator under the zero-truncated Conway–Maxwell–Poisson model (LCMP).

Let $\hat{N}_{(b)}$ denotes the estimated value of the population size at replication b th where $b = 1, 2, 3, \dots, B$, we evaluate estimators performance in terms of relative bias

$$RBias(\hat{N}) = \frac{1}{N} [E(\hat{N}) - N] = \frac{1}{N} bias(\hat{N}), \quad (30)$$

relative variance

$$RVar(\hat{N}) = \frac{1}{N^2} \left\{ \frac{1}{B} \sum_{b=1}^B (\hat{N}_b - E[\hat{N}])^2 \right\} \quad (31)$$

and relative root mean square error

$$RRMSE\{\hat{N}\} = \frac{1}{N} \sqrt{\text{Var}\{\hat{N}\} + \{bias(\hat{N})\}^2}. \quad (32)$$

When the data generation process follows a Poisson distribution, all estimators are asymptotically unbiased with respect to the population size N (see Figs. 1, 2, 3). Sensible differences can be detected for *small* population sizes (e.g. $N = 100$). The estimators allowing for heterogeneity (i.e. Zel and Chao) are persistently biased and show the highest RRMSE values. On the other hand, the LCMP estimator performs in line with the homogeneous, Poisson-based estimators (i.e. Turing and MLEPoi). This is an expected result, as the Poisson distribution is a special case of the CMP one, which is however more general and may be suitable under different data generation settings, far from the Poisson case.

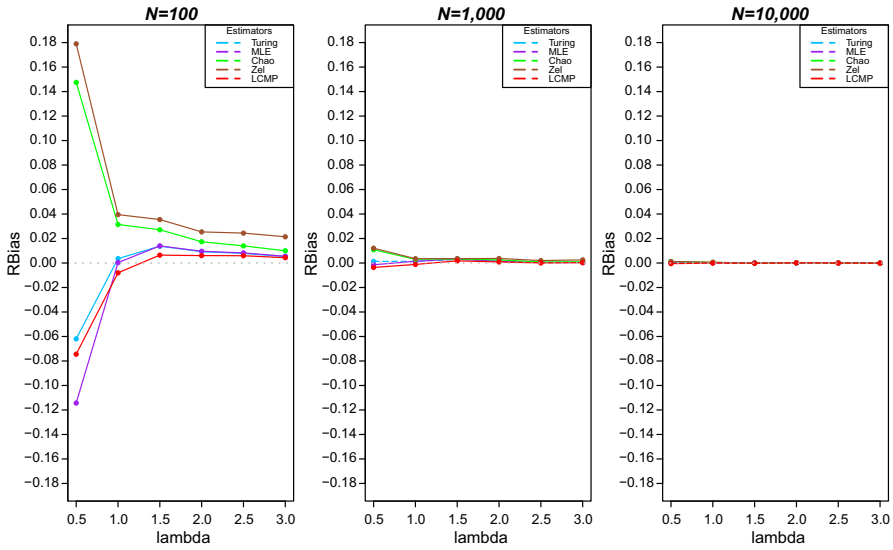


Fig. 1 Relative bias of five estimators for counts drawn from $Poi(\lambda)$

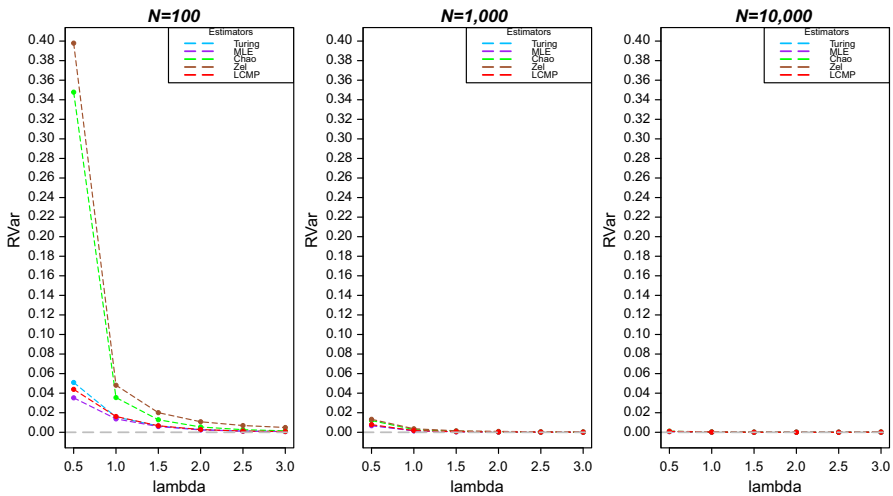


Fig. 2 Relative variance of five estimators of seven estimators for counts drawn from $Poi(\lambda)$

Indeed, by considering a geometric data generation process, the performance of all competing estimators is dramatically poor, no matter of the population size, with the exception of the LCMP estimator proposed in this work (see Figs. 4, 5, 6). The LCMP estimator provides unbiased estimates as the population size increases, at the price of a slightly higher variability, with respect to its competitors. Overall, the LCMP estimator clearly outperforms all the other estimators. Again, this is somehow expected, as the geometric distribution is a specific case nested in the CMP distribution. A comparable performance is reached by the Zelterman estimator for increasing values of λ .

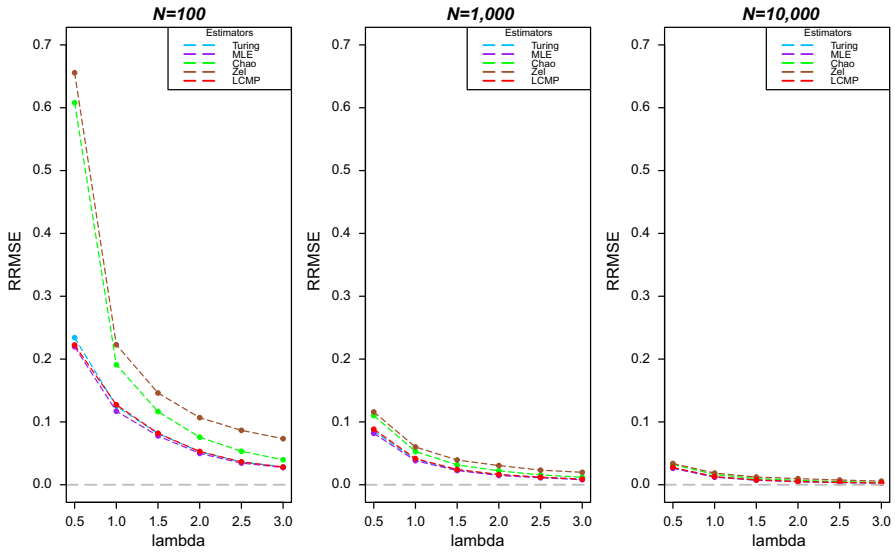


Fig. 3 Relative root mean square error of five estimators for counts drawn from $Poi(\lambda)$

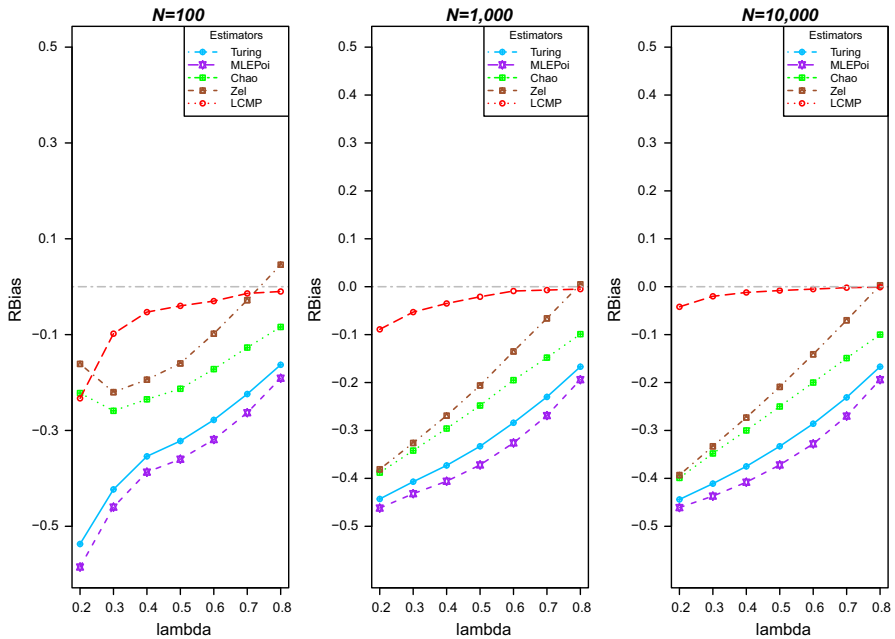


Fig. 4 Relative bias of five estimators for counts drawn from $Geo(\lambda)$; $\lambda = 1 - p$

We further test estimators performance under overdispersion and underdispersion, by generating data from a CMP distribution. We expect that the LCMP estimator, as well as Chao and Zelterman estimators, outperforms Turing and MLEPoi estimators under *heterogeneous* schemes. Results are displayed in Figs. 7, 8 and 9.

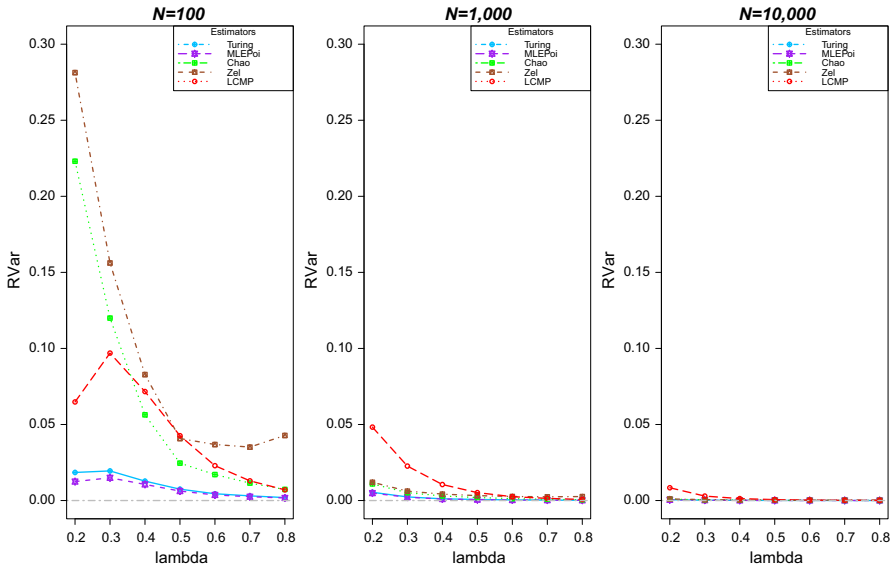


Fig. 5 Relative variance of five estimators for counts drawn from $Geo(\lambda)$; $\lambda = 1 - p$

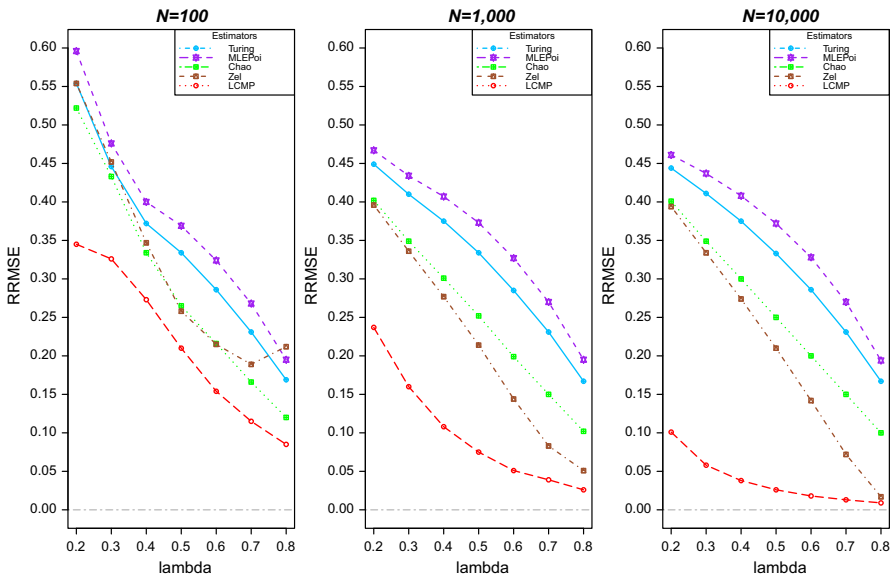


Fig. 6 Relative root mean square of five estimators for counts drawn from $Geo(\lambda)$; $\lambda = 1 - p$

Overall, it can be seen that the LCMP has the best performance when the population size is medium or large, whereas Turing and MLEPoi estimators underestimate the population size. Even the other heterogeneous estimators tend to underestimate the population size, providing reasonable results for $N = 10,000$ only. Indeed, the CMP distribution is a very general one and accounts for many (possible) data fea-

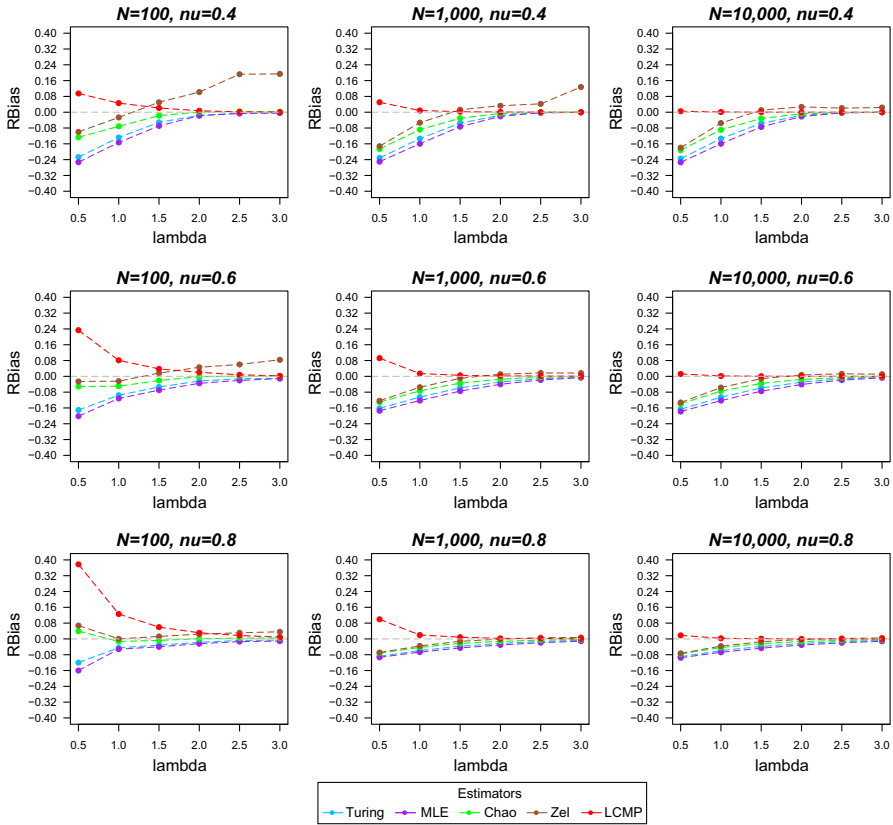


Fig. 7 Relative bias of five estimators for counts drawn from $CMP(\lambda, \nu)$

ture, that may not be captured by existing estimators. To corroborate the simulation results on the LCMP estimator, we provide plots to show convergence to normality under the CMP data generation process for $\lambda = 0.5; 1.0$ and $\nu = 0.4, ; 0.6; 0.8$ (see Fig. 10).

At last, it can be said that under a Negative-Binomial the MLEPoi and Turing’s estimators show a clear underestimation of population size, as well as Chao’s, whereas Zelterman’s estimator does not have a clear path. Similar results can be found in Lanunteang and Böhning (2011). The new estimator performs much better than its competitors, although it tends to overestimate for a small population size and low values of λ . Such an effect disappears as λ and/or k increase (see Figs. 11, 12, 13). To conclude the simulation study, the proposed LCMP estimator can be used even under a Negative Binomial data generation process.

4 Real data examples

In this section we apply different estimators to real data examples. We consider the following benchmark datasets: the cholera data (McKendrick 1926); the golf tees

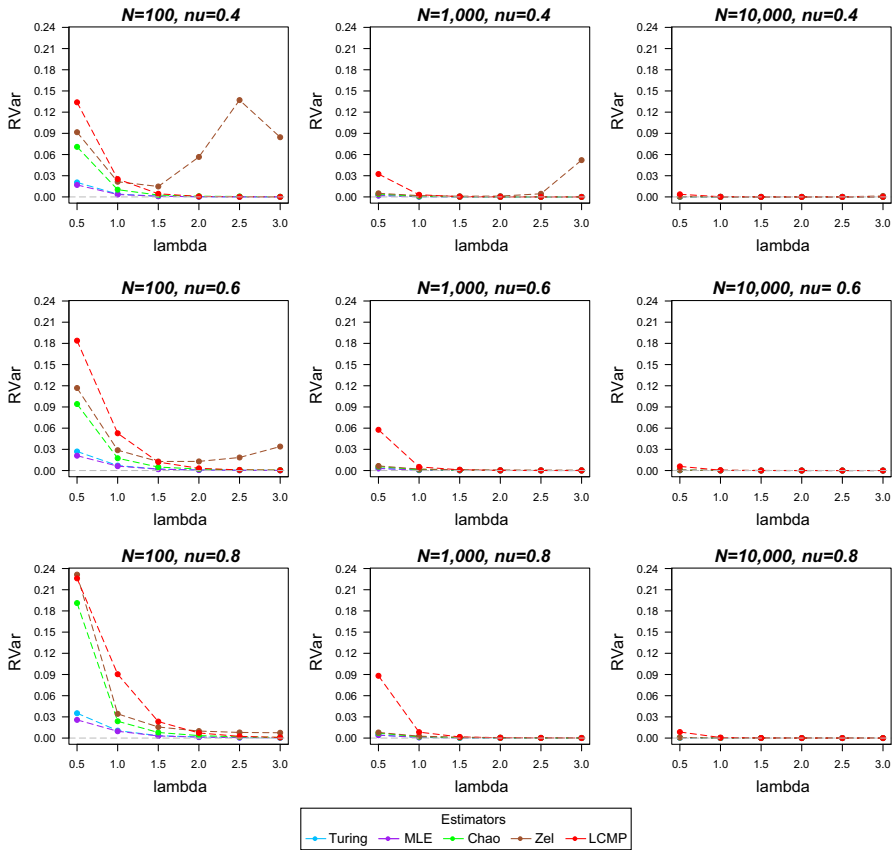


Fig. 8 Relative variance of five estimators for counts drawn from $CM P(\lambda, \nu)$

data (Borchers and Buckland 2002); artificial data used by Link (2003); drug users in Bangkok (Viwatwongkasem et al. 2008). Obtained results under different estimators are also compared to provide an overview of differences in estimating population size in CR data. Furthermore, we provide results for the standard error of the population size by using the asymptotic approximation computed in Sect. 2.4. Goodness-of-fit is investigated through plots and a chi-square goodness of fit test under the null hypothesis of a homogeneous zero-truncated Poisson data is computed. The maximum likelihood estimator under the geometric distribution (Niwitpong et al. 2013) is further considered to see difference with our proposals in terms of model fitting.

4.1 Cholera epidemic in India

The example stems from Mao and Lindsay (2003) and has been discussed previously in Blumenthal et al. (1978), Scollnik (1997), and others. A cholera epidemic affected a village with 223 households in India. Originally, the data were presented by McKendrick (1926) in his paper presentation to the Edinburgh Mathematical Society. Data are provided in Table 1.

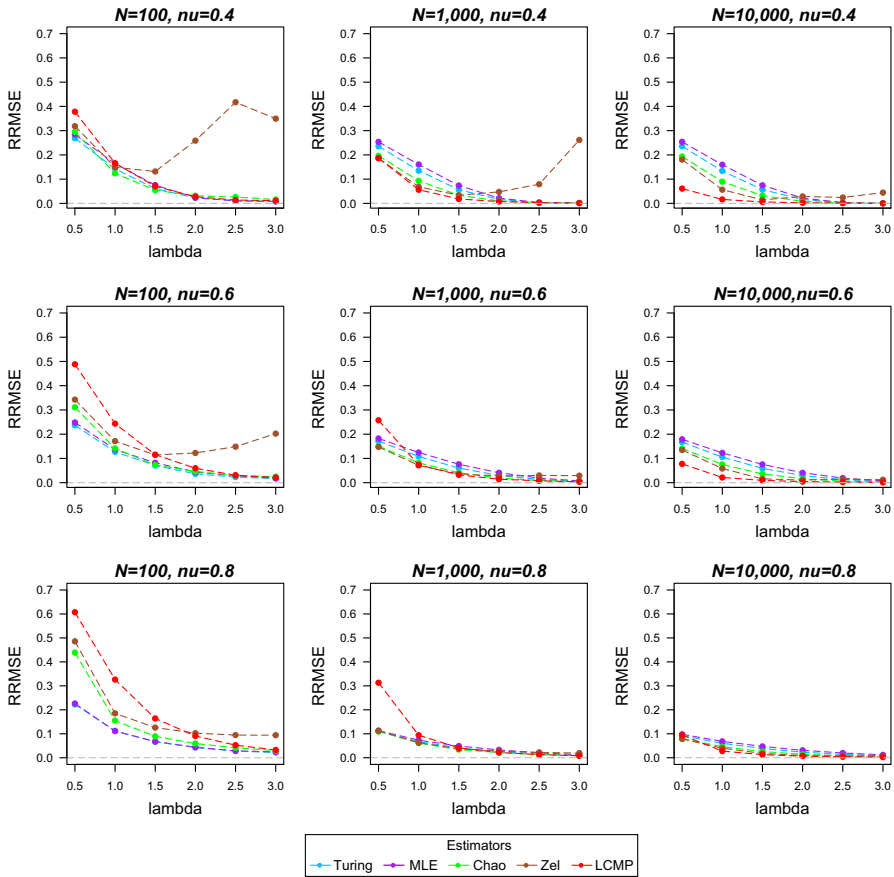


Fig. 9 Relative root mean square error of five estimators for counts drawn from $CMP(\lambda, \nu)$

Table 2 presents the various estimates of the total number and their associated 95% confidence intervals. For the cholera epidemic data evidence has been provided for homogeneity (see Fig. 14). Accordingly, estimates do not differ much; neither do their confidence intervals with the exception of Zelterman which has a large confidence interval. The LCMP approaches the Poisson distribution, as $\hat{\lambda} = 1.01$ and $\hat{\nu} = 1$, i.e. the proposed estimator can be used even if homogeneity is ensured and produces comparable results with homogeneous estimators (e.g. Turing) often used under the homogeneous population setting. This is also confirmed by the formal chi-squared test indicated that the cholera data follow homogeneity of a zero-truncated Poisson distribution with p value of 0.85. The graphical representation of estimated versus observed frequencies, provided in Fig. 18, supports this conclusion.

4.2 Golf-tees data

In a field experiment, $N = 250$ groups of golf tees were placed in a survey region, either exposed above the surrounding grass or hidden by it. They were surveyed by the

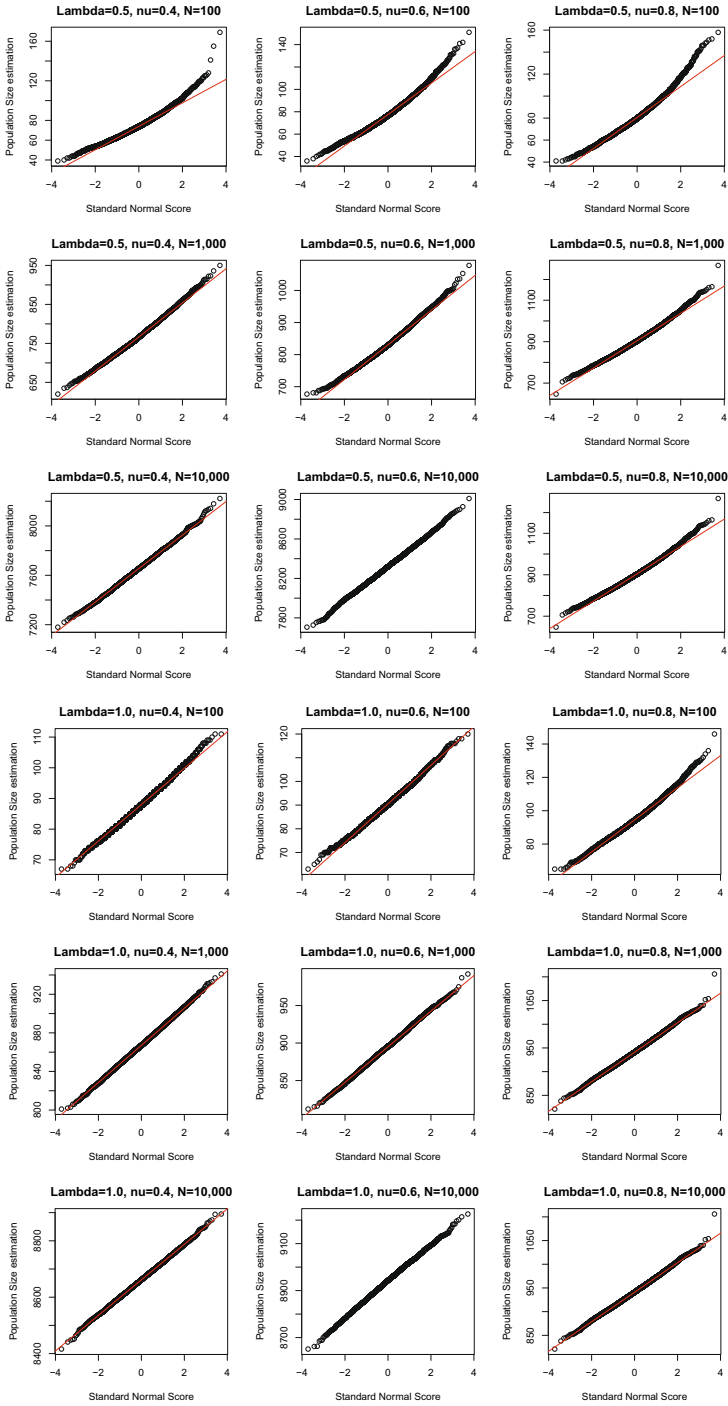


Fig. 10 Normality plot for LCMP estimator under the CMP distribution

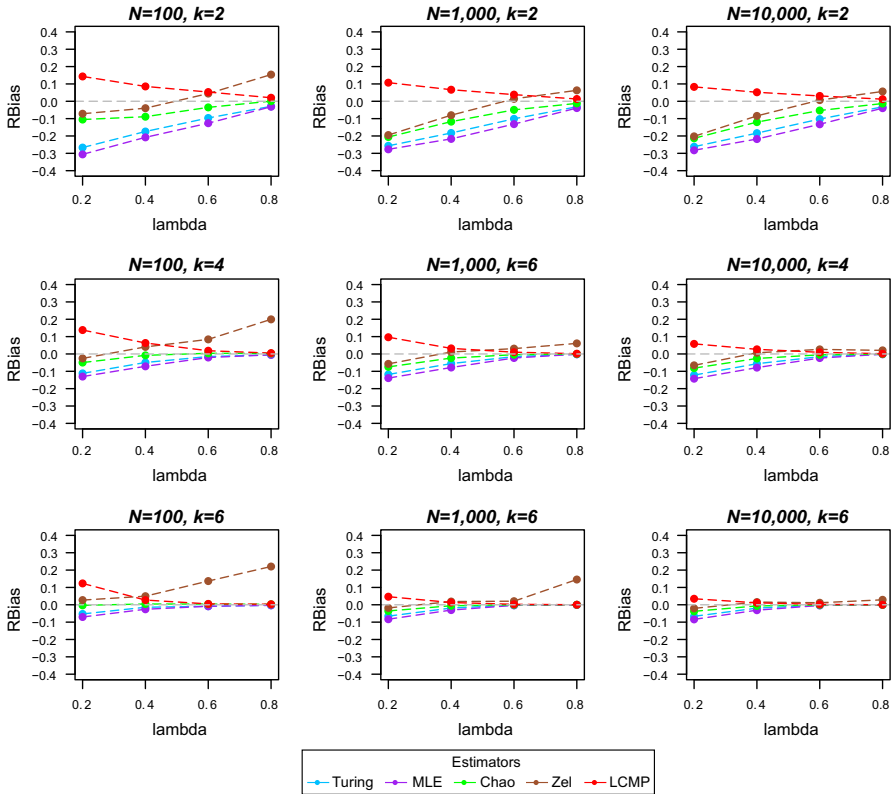


Fig. 11 Relative bias of five estimators for counts drawn from $NB(\lambda, k)$

1999 statistics honor class at the University of St Andrews (Scotland), see [Borchers and Buckland \(2002\)](#). A total of $n = 162$ groups of tees were observed, but a (potentially unknown) number is missed and needs to be estimated. Table 3 shows the corresponding frequency distribution. Figure 15 provides a plot of the log-ratios of successive frequencies and the count distribution.

It is clear that the log-ratio plot displays a linear relationship between log-ratios and log-counts, with a positive slope. It is reasonable to assume that a heterogeneous model would be suited to estimate population size, such as the LCMP estimator so far proposed. The chi-squared test reject the null hypothesis that the data follow a truncated Poisson with a p value < 0.001 . Moreover, f_1 is greater than f_2 and so on, leading to increased variance by increasing x values. Thus, the weighted least square model might be more suitable than the least square. The estimated regression parameter estimates are $\hat{\beta}_0 = -0.268$ and $\hat{\beta}_1 = 1$, i.e. $\hat{\beta}_1$ is on the boundary of the parameter space. Accordingly, the parameters of the zero-truncated Conway–Maxwell–Poisson model are $\hat{\lambda} = 0.765$ and $\hat{\nu} = 0$, i.e. the geometric distribution is obtained. Zelterman’s estimator shows the highest degree of accuracy in terms of having the smallest bias, followed by LCMP. Turing and MLEPoi provide the least accuracy since they show a very large bias, as expected as the log-ratio plot suggests

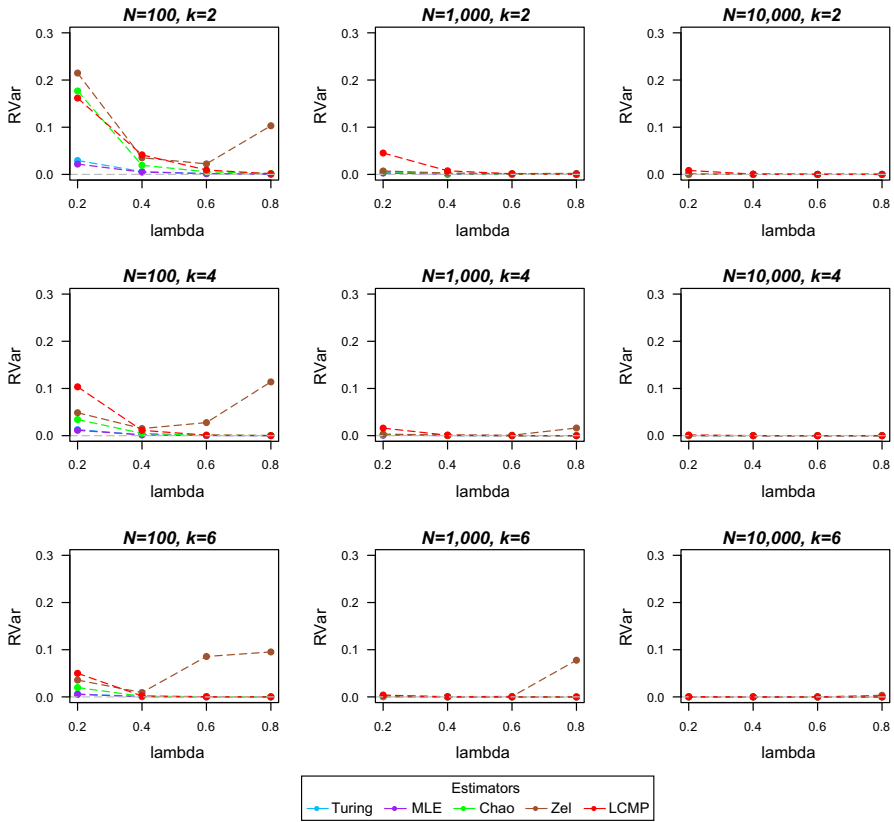


Fig. 12 Relative variance of five estimators for counts drawn from $NB(\lambda, k)$

to avoid any estimator based on a homogeneous model. We would remark that the imposed (and necessary) constraint on $\hat{\beta}_1$ may limit the capacity of the LCMP estimator to recover the *true* population size if the underlying count distribution is far from being geometrically-distributed. However, as the LCMP-based estimator allows for heterogeneity, it provides better estimates than homogeneous population-based estimators. In Fig. 18 we compare estimated frequencies under the LCMP estimator with the homogeneous MLEPoi one and, furthermore, we add estimated frequencies according the MLE under the Geometric distribution. It is even more clear from the graph that the truncated Poisson distribution is not suitable for these data (Table 4).

4.3 Link (2003) data

Her we refer to an artificial dataset considered in Link (2003), see Table 5. These data are of particular interest as they show substantial heterogeneity (see Fig. 16) with a *large* number of maximum recaptures. Thus, we expect that all the considered estim

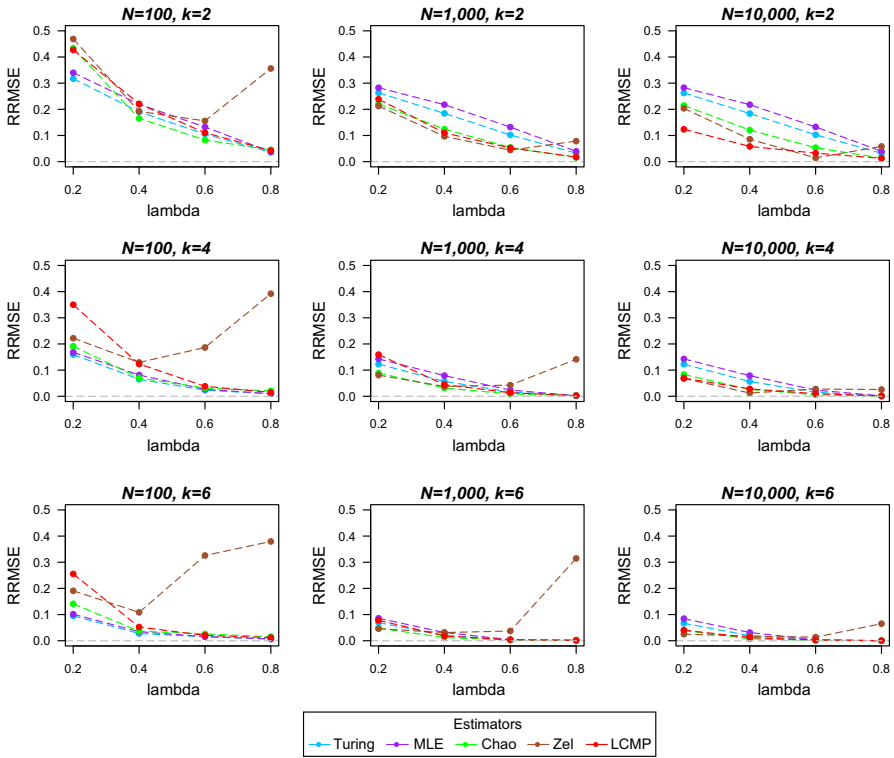


Fig. 13 Relative root mean square error of five estimators for counts drawn from $NB(\lambda, k)$

Table 1 Frequency distribution of the cholera epidemic data

x	1	2	3	4	n
f_x	32	16	6	1	55

Table 2 Cholera data: population size estimates

Model	\hat{N}	$SE(\hat{N})$	95% CI
Homogeneous Poisson			
Turing	88	11.69	66–111
MLEPoi	89	11.59	67–112
Heterogeneous structure			
Chao	87	13.28	61–114
Zelterman	88	17.05	55–122
LCMP	87	7.59	73–102
$(\hat{\lambda} = 1.01 \text{ and } \hat{v} = 1)$			

ators, but the LCMP one, underestimate the population size. Indeed, the long tail of the count variable may lead to biased estimates even for the Zelterman’s estimator. Population size estimates are displayed in Table 6 and provide contradictory inference

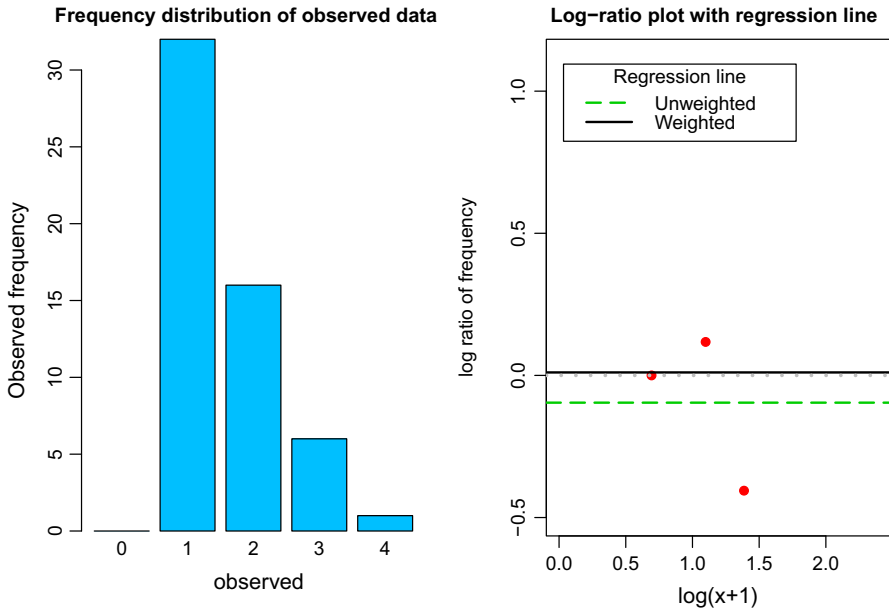


Fig. 14 Cholera data: the log-ratio plot of $\log \left\{ (x + 1) \frac{f_{x+1}}{f_x} \right\}$ versus $\log(x + 1)$

Table 3 Frequency distribution of golf-tees groups detected by eight observers

x	0	1	2	3	4	5	6	7	8	N
f_x	88	46	28	21	13	23	14	6	11	250

about N . Homogeneous population-based estimators shows very low estimates for \hat{N} with small standard errors (a similar behavior was found in the simulation study), and the corresponding 95% confidence intervals do not overlap with those obtained accounting for heterogeneity. Chao’s estimator provides a lower bound for N in presence of heterogeneity, and Zelterman’s estimator does not differ too much, but shows a very large standard error (as expected). The LCMP-based estimator seems to fit well the data, and provides an estimate of N in line with the values obtained in Link (2003) under other parametric distributions accounting for heterogeneity. Figure 18 shows the inability of the homogeneous truncated Poisson distribution to fit the data and the close performance of the LCMP and the Geometric MLE. This is not surprising as we estimate $\nu = 0.0856$, and, as ν approaches zero, the LCMP approaches the MLE under the geometric distribution.

4.4 Heroin drug users in Bangkok

The study used all data on drug use from 61 health treatment centers in the Bangkok metropolitan region collected by the Office of the Narcotics Control Board (ONCB), Ministry of the Prime Minister, which occurred from 1, October to 31, December in

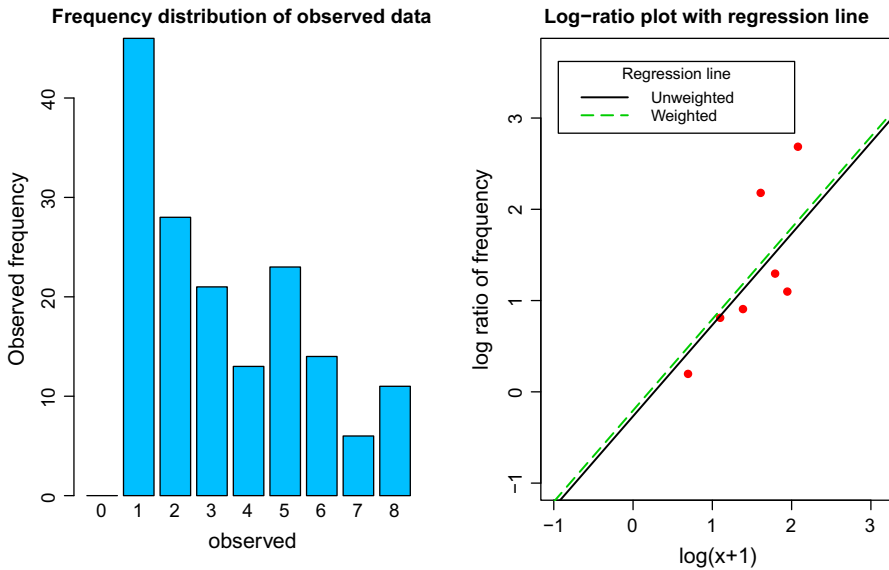


Fig. 15 Golf tees data: the log-ratio plot of $\log \left\{ (x + 1) \frac{f_{x+1}}{f_x} \right\}$ versus $\log(x + 1)$

Table 4 Golf tees data: population size estimates (N = 250)

Model	\hat{N}	Bias	$\widehat{SE}(\hat{N})$	95% CI
Homogeneous Poisson				
Turing	177	-73	4.59	169–186
MLEPoi	169	-81	2.83	164–175
Heterogeneous structure				
Chao	200	-50	13.0	175–226
Zelterman	231	-19	29.83	173–290
LCMP	223	-27	33.09	159–288
$(\hat{\lambda} = 0.765 \text{ and } \hat{\nu} = 0)$				

Table 5 Frequency distribution of Link (2003) data

x	1	2	3	4	5	6	7	8	9	10
f_x	679	531	379	272	198	143	99	67	46	32
x	11	12	13	14	15	16	n			
f_x	22	14	9	5	3	1	2500			

2001. Data are presented in Table 7. From Fig. 17, the log-ratio plot suggests for the use of a heterogeneous model and the LCMP approach seems to fit the data well. A formal test reject the null hypothesis of homogeneity at with a p value < 0.001 . Accordingly, we look at the estimates of the population size under different assumptions. Results are displayed in Table 8. As in the Golf-tees data, the geometric distribution is obtained as a special case of the LCMP, i.e $\nu = 0$. The Zelterman’s estimator does differ too

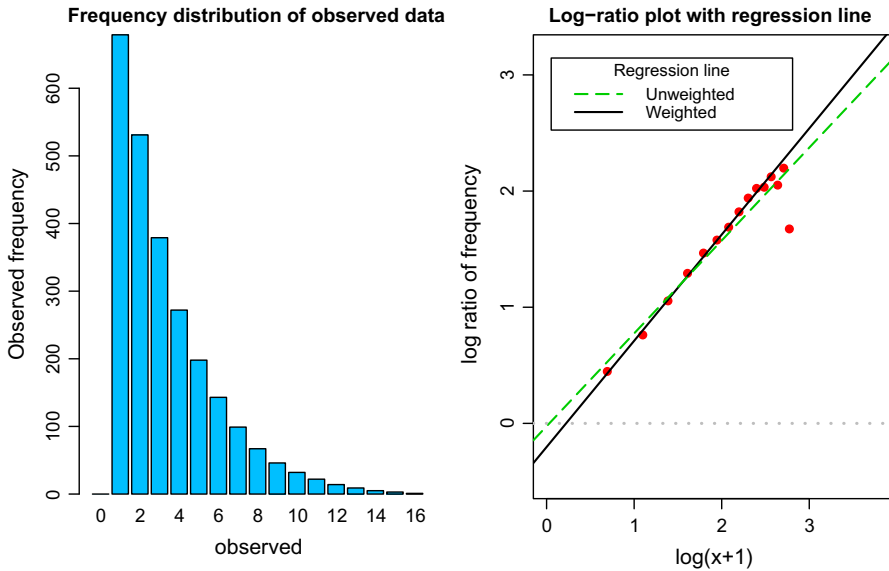


Fig. 16 Link (2003) data: the log-ratio plot of $\log \left\{ (x + 1) \frac{f_{x+1}}{f_x} \right\}$ versus $\log(x + 1)$

Table 6 Link (2003) data: population size estimates

Model	\hat{N}	$\widehat{SE}(\hat{N})$	95 % CI
Homogeneous Poisson			
Turing	2719	17.53	2685–2754
MLEPoi	2602	11.04	2602–2581
Heterogeneous structure			
Chao	2935	37.34	2935–2862
Zelterman	3162	81.15	3003–3322
LCMP	3333	45.21	3245–3422
$(\hat{\lambda} = 0.816 \text{ and } \hat{v} = 0.086)$			

Table 7 Frequency distribution of heroin users in Bangkok

x	1	2	3	4	5	6	7	8	9	10	
f_x	2176	1600	1278	976	748	570	455	368	281	254	
x	11	12	13	14	15	16	17	18	19	20	21
f_x	188	138	99	67	44	34	17	3	3	2	1

much from the LCMP estimator (with almost overlapping confidence intervals), whilst all the homogeneous estimators provide smaller sample sizes. Again, the LCMP and the MLE under the geometric distributions are equivalent in terms of model fit (see Fig. 18).

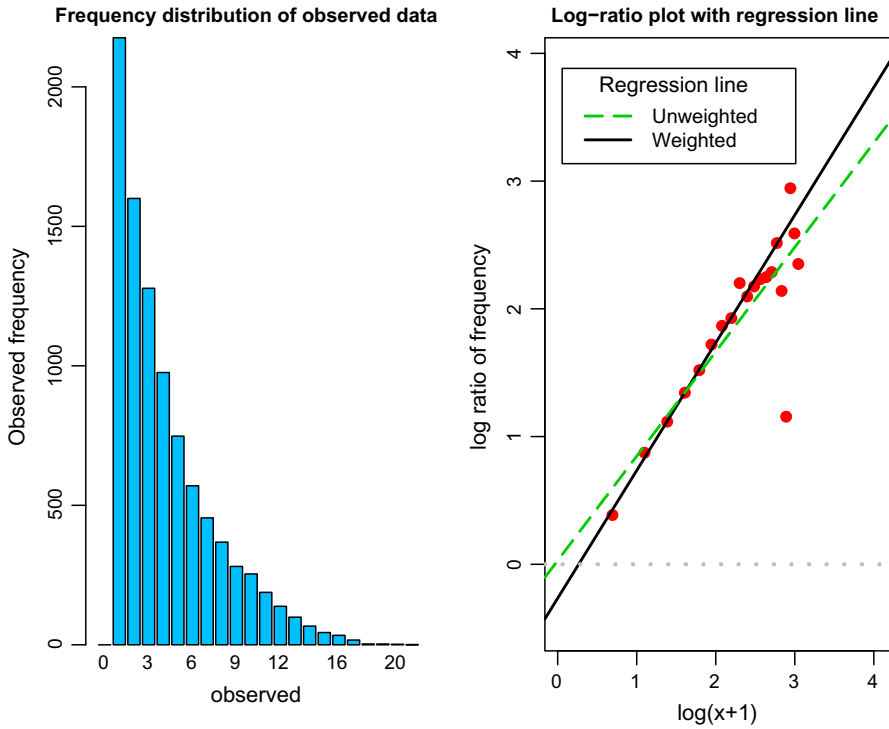


Fig. 17 Heroin users in Bangkok: the log-ratio plot of $\log \left\{ (x + 1) \frac{f_{x+1}}{f_x} \right\}$ versus $\log(x + 1)$

Table 8 Heroin users in Bangkok data: population size estimates

Model	\hat{N}	$SE(\hat{N})$	95% CI
Homogeneous Poisson			
Turing	9850	26.65	9798–9850
MLEPoi	9454	12.84	9429–9479
Heterogeneous structure			
Chao	10,782	71.86	10,641–10,923
Zelnerman	12,077	184.54	11,715–12,439
LCMP	12,141	210.24	11,729–12,554
$(\hat{\lambda} = 0.77 \text{ and } \hat{\nu} = 0)$			

5 Conclusion

A diversity of estimators in the capture–recapture field exists, being widely applied in many areas of interest. Here, we have introduced a new method of estimating the population size under a specific form of heterogeneity based on the Conway–Maxwell–Poisson distribution. We have also been able to see how accurate and precise the method

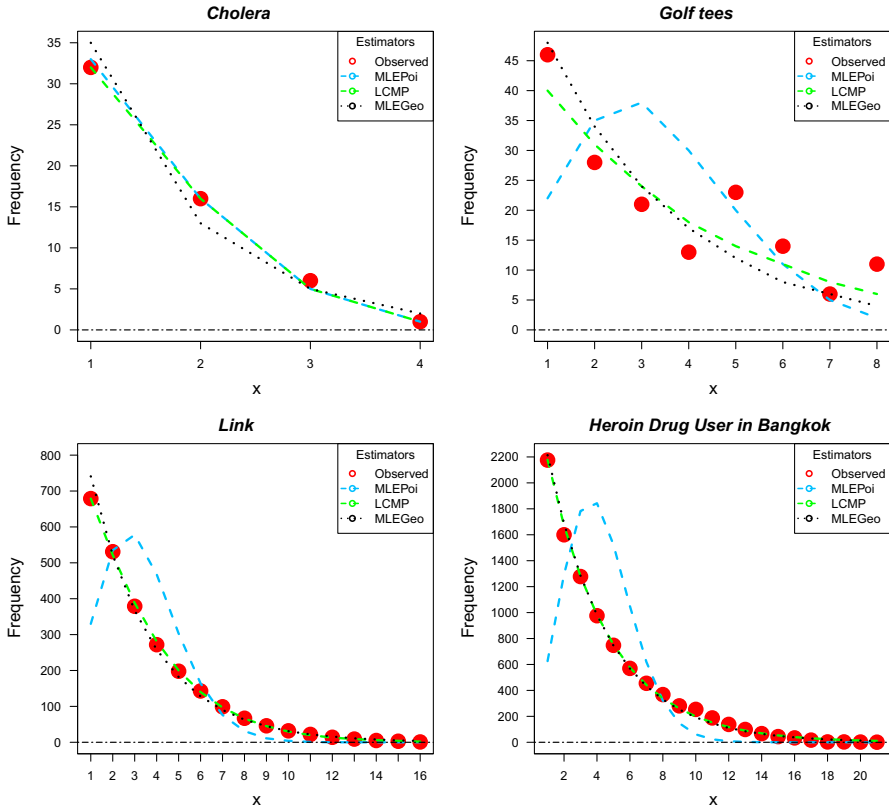


Fig. 18 Real data examples: estimated versus observed frequencies

is performing when it is compared to other frequently used estimators. Overall, the proposed estimator is more accurate as well as providing small bias in the homogeneous Poisson case which asymptotically disappears. It is also found that the new estimator performs well under different heterogeneous data generation processes (i.e. Geometric, Negative Binomial); hence, it improves existing heterogeneous estimators (e.g. Chao’s and Zelterman’s estimators). Although the proposed estimator showed a better performance in terms of accuracy, it evidently gave also the largest variation; nonetheless, the variation of the new estimator considerably decreases for large population size (1000 and more), as often in real-world applications. We also provided a formula of variance approximation of the new estimator. This variance formula is not only useful to determine the efficiency of estimating, but it can be also used to construct confidence intervals. In short, the new estimator can be an alternative form of population size estimation especially for large populations and heterogeneous capturing probabilities.

The use of the ratio plot allows us to avoid computational issues related to CMP distribution. Furthermore, by using the ratio plot, formal tests can be conducted on null hypotheses of zero-truncated Poisson, i.e. $H_0 : \beta_1 = 0$, or geometric, i.e. $H_0 : \beta_1 = 1$, data. The proposed LCMP estimator performs equivalently as the MLE under the

Poisson and the geometric distribution, supporting that the use of the ratio plot, instead of computing the MLE under the CMP distribution, does not affect estimates. We have not reported these results, but they are available upon request.

Acknowledgements Authors would like to thank the reviewers for the very helpful comments which lead to considerable improvement of the paper as well as to interesting and promising modifications of the proposed estimator. Authors are grateful to the Ministry of Science and Technology, the Royal Thai Government for providing Ph.D. funding for the first author.

References

- Alunni-Fegatelli D, Tardella L (2013) Improved inference on capture-recapture models with behavioural effects. *Stat Methods Appl* 22:45–66
- Baksh MF, Böhning D, Lerdsuwansri R (2011) An extension of an over-dispersion test for count data. *Comput Stat Data Anal* 55:466–474
- Bartolucci F, Forcina A (2006) A class of latent marginal models for capture-recapture data with continuous covariates. *J Am Stat Assoc* 101:786–794
- Blumenthal JA, Williams RB, Kong Y, Schanberg SM, Thompson LW (1978) Type A behavior pattern and coronary atherosclerosis. *Circulation* 58:634–639
- Böhning D, Dietz E, Kuhnert R, Schön D (2005) Mixture models for capture–recapture count data. *Stat Methods Appl* 14:29–43
- Böhning D, Schön D (2005) Nonparametric maximum likelihood estimation of population size based on the counting distribution. *J R Stat Soc Ser C* 54:721–737
- Böhning D (2008) A simple variance formula for population size estimators by conditioning. *Stat Methodol* 5:410–423
- Böhning D, Baksh MF, Lerdsuwansri R, Gallagher J (2013) Use of the ratio plot in capture-recapture estimation. *J Comput Gr Stat* 22:135–155
- Borchers DL, Buckland ST (2002) Estimating animal abundance: closed populations. Springer, New York
- Bunge J, Barger K (2008) Parametric models for estimating the number of classes. *Biom J* 50:971–982
- 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
- Chiu CH, Wang YT, Walther BA, Chao A (2014) An improved non-parametric lower bound of species richness via a modified good-turing frequency formula. *Biometrics* 70:671–682
- Dorazio RM, Royle AJ (2003) Mixture models for estimating the size of a closed population when capture rates vary among individuals. *Biometrics* 59:351–364
- Farcomeni A (2011) Recapture models under equality constraint. *Biometrika* 98:237–242
- Farcomeni A, Scacciatelli D (2013) Heterogeneity and behavioural response in continuous time capture–recapture, with application to street cannabis use in Italy. *Ann Appl Stat* 7:2293–2314
- Gerritse S, van der Heijden PGM, Bakker B (2015) Sensitivity of population size estimation for violating parametric assumptions in loglinear models. *J Off Stat* 31:357–379
- Guikema SD, Coffelt JP (2008) A flexible count data regression model for risk analysis. *Risk Anal* 28:213–223
- Kuhnert R, Böhning D (2009) CAMCR: computer-assisted mixture model analysis for CaptureRecapture count data. *AStA Adv Stat Anal* 93:61–71
- Lanumteang K (2011) Estimating of size of a target population using capture–recapture methods based upon multi-source and continuous time experiments. Ph.D. thesis, University of Reading
- Lanumteang K, Böhning D (2011) An extension of Chao’s estimator of population size based on the first three capture frequency counts. *Comput Stat Data Anal* 55:2302–2311
- Lerdsuwansri R (2012) Generalisation of the Lincoln-Petersen approach to non-binary source variable. Ph.D. thesis, University of Reading
- Lindsay BG, Roeder K (1987) A unified treatment of integer parameter models. *J Am Stat Assoc* 82:758–764
- Link WA (2003) Nonidentifiability of population size from capture-recapture data with heterogeneous detection probabilities. *Biometrics* 59:1123–1130

- Mao CX, Lindsay BG (2003) Tests and diagnostics for heterogeneity in the species problem. *Comput Stat Data Anal* 41:389–398
- McCrea RS, Morgan BJT (2014) *Analysis of capture–recapture data*. CRC Press, Boca Raton
- McKendrick AG (1926) Application of mathematics to medical problems. *Proc Edinb Math Soc* 44:98–130
- Meurant G (1992) A review on the inverse of symmetric tridiagonal and block matrices. *SIAM J Matrix Anal Appl* 13:707–728
- Morgan BJT, Ridout MS (2008) A new mixture model for capture heterogeneity. *J R Stat Soc Ser C* 57:433–446
- Niwitpong SA, Böhning D, van der Heijden PG, Holling H (2013) Capture-recapture estimation based upon the geometric distribution allowing for heterogeneity. *Metrika* 76:495–519
- Pledger S (2005) The performance of mixture models in heterogeneous closed population capture-recapture. *Biometrics* 61:868–873
- Rocchetti I, Bunge J, Böhning D (2011) Population size estimation based upon ratios of recapture probabilities. *Ann Appl Stat* 5:1512–1533
- Rocchetti I, Alfó M, Böhning D (2014) A regression estimator for mixed binomial capture–recapture data. *J Stat Plan Inference* 145:165–178
- Scollnik DP (1997) Inference concerning the size of the zero class from an incomplete Poisson sample. *Commun Stat Theory Methods* 26:221–236
- Shmueli G, Minka TP, Kadane JB, Borle S, Boatwright P (2005) A useful distribution for fitting discrete data: revival of the ConwayMaxwellPoisson distribution. *J R Stat Soc Ser C* 54:127–142
- van Der Heijden PG, Bustami R, Cruyff MJ, Engbersen G, Van Houwelingen HC (2003) Point and interval estimation of the population size using the truncated Poisson regression model. *Stat Model* 3:305–322
- Viwatwongkasem C, Kunhert R, Sativipawee P (2008) A Comparison of population size estimators under the truncated count model with and without allowance for contaminations. *Biom J* 50:1006–1021
- Zelterman D (1988) Robust estimation in truncated discrete distributions with application to capture-recapture experiments. *J Stat Plan Inference* 18:225–237