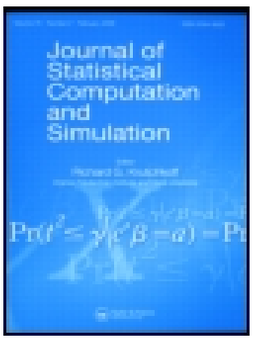

Journal of Statistical Computation and Simulation

ISSN: 0094-9655 (Print) 1563-5163 (Online) Journal homepage: http://www.tandfonline.com/loi/gscs20

Uncertainty estimation in heterogeneous capture–recapture count data

Orasa Anan, Dankmar Böhning & Antonello Maruotti

To cite this article: Orasa Anan, Dankmar Böhning & Antonello Maruotti (2017): Uncertainty estimation in heterogeneous capture–recapture count data, Journal of Statistical Computation and Simulation, DOI: 10.1080/00949655.2017.1315668

To link to this article: http://dx.doi.org/10.1080/00949655.2017.1315668

Published online: 20 Apr 2017.

JOURNAL OF STATISTICAL COMPUTATION AND SIMULATION, 2017
http://dx.doi.org/10.1080/00949655.2017.1315668

Taylor & Francis
Taylor & Francis Group

Uncertainty estimation in heterogeneous capture–recapture count data

Orasa Anan[a,b], Dankmar Böhning[a] and Antonello Maruotti[c,d]

[a]Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, UK; [b]Department of Mathematics and Statistics, Thaksin University, Phatthalung, Thailand; [c]Centre for Innovation and Leadership in Health Sciences, University of Southampton, Southampton, UK; [d]Dipartimento di Scienze Economiche, Politiche e delle Lingue Moderne, Libera Universitá Maria Ss. Assunta, Roma, Italy

ABSTRACT

The Conway–Maxwell–Poisson estimator is considered in this paper as the population size estimator. 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. Little emphasis is often placed on the variability associated with the population size estimate. This paper provides a deep and extensive comparison of bootstrap methods in the capture–recapture setting. It deals with the classical bootstrap approach using the true population size, the true bootstrap, and the classical bootstrap using the observed sample size, the reduced bootstrap. Furthermore, the imputed bootstrap, as well as approximating forms in terms of standard errors and confidence intervals for the population size, under the Conway–Maxwell–Poisson distribution, have been investigated and discussed. These methods are illustrated in a simulation study and in benchmark real data examples.

ARTICLE HISTORY
Received 10 April 2016
Accepted 1 April 2017

KEYWORDS
Conwway–Maxwell–Poisson distribution; capture–recapture methods; bootstrap; ratio-plot

AMS SUBJECT CLASSIFICATION
62F40; 62P12; 62-07

1. Introduction

It has been recognized that in capture–recapture (CR) experiments heterogeneity may influence capture probabilities, and failure to acknowledge this may lead to biased estimates of the unknown population size. References [1–4] show that the ignorance of heterogeneity effect yields a negatively biased estimation of the population size. Similarly, references [5–7] acknowledge the underestimation of sample size by the occurrence of unmodelled heterogeneity of capture probabilities.

In CR analyses, one of the most used approaches to deal with heterogeneity is the mixed model, that assumes that some components of the capture probabilities arise from a mixing distribution [8–10]. Other examples are given by finite mixture models [11,12].

In this paper we aim to investigate uncertainty in a simple and powerful CR estimator, namely the Conway–Maxwell–Poisson (CMP) estimator, able to capture different levels of heterogeneity adaptively. This is a challenge faced by existing CR estimators. The CMP estimator can capture power-law behaviour, excessive zeros or high skewness of the underlying distribution without the need for an additional mixture component.

The motivation behind considering the CMP estimator stems from the role recently played by the CMP distribution [13] and its extensions [14–17]. The estimation and inference for parameters of a CMP distribution have been investigated in a small number of studies [18,19]. The CMP distribution is a two-parameter generalized form of the Poisson distribution. It includes

CONTACT Antonello Maruotti a.maruotti@soton.ac.uk

as special submodels important distributions (i.e. the Poisson, the Bernoulli and geometric distributions) and generalizes the Poisson distribution allowing for overdispersion as well as underdispersion.

Model parameters are estimated through weighted least squares (WLS), based on a graphical device, namely the ratio-plot [20]. 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. Here, we go beyond its useful descriptive nature and use the ratio-plot to obtain the estimate of CMP parameters, and accordingly of the population size.

An overlooked issue in general CR analyses is the quantification of uncertainty surrounding the estimates of the unknown population size. An estimation of the population size can be accurate and precise, but if the associated estimation of variance is poor, then coverage by the 95% confidence interval may falsely indicate poor estimation by the point estimator, that is, the point estimator may result in a poor coverage rate. Focusing on the CMP estimator, we attempt here to investigate bootstrap methods as a robust and general approach to estimate variances and confidence intervals. Various bootstrap methods have been considered to estimate uncertainty in CR analyses with respect to other estimators [21–23]; however, bootstrap results in a variance estimate which is likely to be smaller than the true variance, because it conditions on being observed [24]. In addition to bootstrap methods, we examine a variance approximation method that could be of practical use. Indeed, although bootstraps are useful in CR analyses since they provide omnibus tools for variance and confidence interval estimation, for the CMP estimator, the approximated variance may give accurate estimates and reduce the required computational burden. Here, we attempt to compare the performance of different bootstrap methods (namely, the true, reduced and imputed bootstrap methods) and an approximation-based approach in terms of variance estimation and confidence intervals in CR count data where data are generated under a CMP distribution or from a model outside the CMP family. Together these methods encompass the various variance estimation proposals.

A large-scale simulation study is provided. Several data generation schemes are considered. We focus on comparing different methods to assess uncertainty about population size estimates. An approximated variance specification is compared with different bootstrap approaches in terms of recovering the true variability in the estimates as well as looking at 95% coverage probabilities. A complete investigation of these methods is provided on simulated data varying the sample size and the heterogeneity in the data. To provide evidence of what might happen in the analyses of real data, we provide further numerical examples based on well-known benchmark datasets. We discuss the implications of using different uncertainty assessment methods and the flexibility of the CMP-based estimator to address several data features.

The outline of the paper is as follows. In Section 2, we introduce the CMP estimator, along with the ratio-plot and the computational aspects of the adopted regression-based algorithm. In Section 3, we provide details on variance estimation. An approximated variance is analytically computed and bootstrap methods are introduced and implementation details are discussed. The performance of several model specifications under different data generation schemes by means of a simulation study is provided in Section 4. In Section 5, we present several real-data analyses. In Section 6, we point out some remarks, along with drawbacks that may arise by adopting the proposed methodology.

2. The CMP estimator for heterogeneous CRdata

2.1. Preliminaries

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, for a general introduction to CR

data modelling see e.g. [25]. In the following, we consider a closed population, that is, the unknown population size, say N, is assumed to be constant (with no births/deaths during sampling stages), misclassification is not allowed and all units act independently.

Formally, let X_i, $i = 1, \ldots, N$ 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, $x = 0, 1, \ldots, m$, where m is the largest observed count. 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, a count data model may represent a natural starting point.

2.2. The CMP distribution

In modelling count data, the CMP distribution has recently played an important role. The CMP probability distribution function, CMP(λ, ν), has the form [13]

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

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 contains some well-known discrete distributions:

- for $\nu = 1$, $z(\lambda, \nu) = e^\lambda$, and the CMP distribution simply reduces to the ordinary Poisson(λ);
- for $\nu \to \infty$, $z(\lambda, \nu) \to 1 + \lambda$, and the CMP distribution approaches the Bernoulli with parameter $\lambda(1 + \lambda)^{-1}$;
- for $\nu = 0$ and $0 < \lambda < 1$, $z(\lambda, \nu)$ is a geometric sum

$$z(\lambda, \nu) = \sum_{j=0}^{\infty} \lambda^j = \frac{1}{1 - \lambda},$$

  and, accordingly, the CMP distribution reduces to the geometric distribution $p_x = \lambda^x(1 - \lambda)$;
- for $\nu = 0$ and $\lambda \geq 1$, $z(\lambda, \nu)$ does not converge, leading to an undefined distribution.

In general, of course, the normalizing constant $z(\lambda, \nu)$ does not permit such a neat, closed-form expression. Asymptotic results are, however, available. Gillispie and Green [19] prove that, for fixed ν,

$$z(\lambda, \nu) \sim \frac{\exp(\nu\lambda^{1/\nu})}{\lambda^{(\nu-1)/2\nu}(2\pi)^{(\nu-1)/2}\sqrt{\nu}}(1 + O(\lambda^{-1/\nu})),$$

as $\lambda \to \infty$, confirming the conjecture made by [13].

To complete the description on the CMP distribution, let us specify CMP moments. There are no simple closed form linking the parameters to moments, but some recurrence relations and

approximations are provided by using an asymptotic approximation of $z(\lambda, \nu)$ (see e.g. [13,26])

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

It is evident that both $E(X)$ and $Var(X)$ are increasing functions of λ but decrease with respect to ν.

2.3. Estimating the CMP parameters

The likelihood function for a set of n independent and identically distributed observations is

$$L(\cdot \mid \lambda, \nu) = \lambda^{\sum_{i=1}^{n} x_i} \exp\left\{-\nu \sum_{i=1}^{n} \log x_i!\right\} z^{-n}(\lambda, \nu).$$

As widely discussed in [13], parameters estimates can be obtained by maximizing the likelihood function performing constrained numerical maximization techniques. Nevertheless, computational issues may arise as the maximization procedure involves the infinite sum $z(\lambda, \nu)$. Furthermore, in CR studies, the zero counts are truncated and, hence, the sample frequencies arise from a zero-truncated distribution. Thus, a zero-truncated CMP distribution should be considered and this may further complicate model inference.

To avoid numerical issues, we estimate model parameters by combining a simple graphical technique, that is, the ratio-plot [20], with a computationally efficient least squares method. The proposed method is based on ratios of successive probability counts

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

which is a function of the observed count x. Bearing in mind that the ratio r_x for the truncated and the untruncated distribution is identical as

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

the ratio for the CMP distribution is

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

and does not depend on the complex normalizing constant $z(\lambda, \nu)$. Let us consider the ratio on the log scale, we achieve a linear model

$$\begin{aligned}\log\{r_x\} &= \log\left\{(x+1)\frac{p_{x+1}}{p_x}\right\} = \log\{\lambda(x+1)^{1-\nu}\} \\ &= \log\lambda + (1-\nu)\log(x+1) = \beta_0 + \beta_1 \log(x+1).\end{aligned} \tag{2}$$

From (2), 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 $\beta_1 \leq 1$. Similarly, $\lambda > 0$ implies $\beta_0 \in (-\infty, +\infty)$. Furthermore, two basic assumptions of ordinary regression models are violated here. First, the variance of the dependent variable is not constant. The second deviation from ordinary regression assumptions is the fact that the 'observations' are not independent.

All these issues are relevant and should be accounted for. Thus, we address them by using WLS techniques to estimate the regression parameters β_0 and β_1. These are obtained as

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{Y}, \tag{3}$$

where $\mathbf{Y} = (\log r_1, \log r_2, \ldots, \log r_{m-1})'$, $\mathbf{X}$ is he design matrix containing the regression functions of the model and $\mathbf{W}$ is a diagonal matrix containing the estimated inverse variances of $Y_1, \ldots, Y_{m-1}$, i.e.

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

Here, m is the largest count observed.

Accordingly, the estimator based on the CMP distribution of target population size can be readily achieved as (see e.g. [27])

$$\hat{N}_{\text{CMP}} = n + \hat{f}_0 = n + f_1 \exp(-\hat{\beta}_0),$$

where $\hat{\beta}_0$ is got from the weighted least square method.

3. Variance estimation

While the WLS algorithm provides an efficient means of parameter estimation in the CR modelling context, the default output does not provide estimates of the uncertainty associated with the parameter estimates. Several approaches have been considered to facilitate the provision of standard errors within this context [21–23].

3.1. Approaches based upon resample techniques

Examples of variance estimation based on bootstrap methods have been proposed in the literature [22,23,28], but never fully investigated. However, all the existing works look at resampling techniques as a promising and useful alternatives to provide standard errors in the CR framework. In the following, bootstrap methods to obtain an estimate of the variance associated with the population size estimate are described. Bootstrap methods are straightforward to implement, regardless of the model under consideration. Here, we consider the True Bootstrap (TB), the Imputed Boostrap (IB) and the RB. In our setting, the algorithm for TB, IB and RB variance estimation techniques proceeds as follows

(i) Estimate $\hat{N}_{\text{CMP}}$ as described in Section 2. This provides an estimate of $f_0, \hat{f}_0$.
(ii) Form R samples comprising of observations from the original data as follows:
- Let $\hat{p}_{\text{TB}} = \{f_0/N, f_1/N, \ldots, f_m/N\}$. The TB can be applied for estimating variance of interested population size estimator only if the population size is known. Accordingly, under the TB approach, each of the R_{TB} samples contains N observations drawn from a Multinomial distributions with parameters N and $\hat{p}_{\text{TB}}$.
- Let $\hat{p}_{\text{IB}} = \{\hat{f}_0/\hat{N}_{\text{CMP}}, f_1/\hat{N}_{\text{CMP}}, \ldots, f_m/\hat{N}_{\text{CMP}}\}$. Under the IB approach, each of the R_{IB} samples contains $\hat{N}_{\text{CMP}}$ observations drawn from a Multinomial distributions with parameters $\hat{N}_{\text{CMP}}$ and $\hat{p}_{\text{IB}}$.

- Let $\hat{p}_{\mathrm{RB}} = \{f_1/n, f_2/n, \ldots, f_m/n\}$. Under the RB approach, each of the R_{RB} samples contains n observations, where the observations are sampled with replacement from the observed data.

(iii) For each sample, estimate $\hat{N}_{\mathrm{CMP}}$, under the CMP model. In the case of the TB, it means that the true f_0 is ignored.

(iv) Estimate the variance of $\hat{N}_{\mathrm{CMP}}$ on R bootstrapped samples

- The TB estimate of the variance of $\hat{N}_{\mathrm{CMP}}$ is equal to

$$\sigma^2_{\mathrm{TB}} = \frac{1}{R-1}\sum_{r=1}^{R}(\hat{N}_{\mathrm{CMP},r} - \bar{\hat{N}}_{\mathrm{CMP,TB}}))^2,$$

  where $\bar{\hat{N}}_{\mathrm{CMP,TB}}$ is the TB sample mean.
- The IB estimate of the variance of $\hat{N}_{\mathrm{CMP}}$ is equal to

$$\sigma^2_{\mathrm{IB}} = \frac{1}{R-1}\sum_{r=1}^{R}(\hat{N}_{\mathrm{CMP},r} - \bar{\hat{N}}_{\mathrm{CMP,IB}})^2,$$

  where $\bar{\hat{N}}_{\mathrm{CMP,IB}}$ is the IB sample mean.
- The RB estimate of the variance of $\hat{N}_{\mathrm{CMP}}$ is equal to

$$\sigma^2_{\mathrm{RB}} = \frac{1}{R-1}\sum_{r=1}^{R}(\hat{N}_{\mathrm{CMP},r} - \bar{\hat{N}}_{\mathrm{CMP,RB}})^2,$$

  where $\bar{\hat{N}}_{\mathrm{CMP,RB}}$ is the RB sample mean.

3.2. An approximation-based approach

Another benefit of the ratio regression approach is that variance estimators for f_0 can easily be developed as variance estimators for the estimated regression coefficients are easily available. Let $\hat{N}$ be the population size estimator, the variance of $\hat{N}_{\mathrm{CMP}} = n + f_1\,\mathrm{e}^{-\hat{\beta}_0}$ arise from two sources; these are influenced by the random variable n and the estimator $\hat{f}_0$. Therefore a simple formula for the variance of the population size estimator is given as

$$\mathrm{Var}(\hat{N}) = \mathrm{Var}_n\{E(\hat{N}\,|\,n\} + E_n\{\mathrm{Var}(\hat{N}\,|\,n)\}.$$

We apply a technique for computing moments usually referred to as conditioning (see e.g. [29]) to population size estimation. The technique provides a simple formula for variance computation of population size which can be applied to a general estimator. According to the conditional technique, we have

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

and thus

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

Using the delta method, we achieve that $\mathrm{Var}(\mathrm{e}^{-\hat{\beta}_0}) = (\mathrm{e}^{-\hat{\beta}_0})^2\mathrm{Var}(\hat{\beta}_0)$. Hence $E_{f_1}\{\mathrm{Var}(f_1\,\mathrm{e}^{-\hat{\beta}_0})\,|\,f_1\} \approx f_1^2(\mathrm{e}^{-\hat{\beta}_0})^2\mathrm{Var}(\hat{\beta}_0)$, where $\mathrm{Var}(\hat{\beta}_0)$ comes from the linear regression process. The approximated

expression for the variance of the CMP estimator $\hat{N}_{\text{CMP}}$ is given as

$$\widehat{\text{Var}}(\hat{N}_{\text{CMP}}) = \frac{nf_1 \, \text{e}^{-\hat{\beta}_0}}{n + f_1 \, \text{e}^{-\hat{\beta}_0}} + (\text{e}^{-\hat{\beta}_0})^2 f_1 \left(1 - \frac{f_1}{N}\right) + f_1^2 (\text{e}^{-\hat{\beta}_0})^2 \text{Var}(\hat{\beta}_0).$$

As $1 - f_1/N \leq 1$, a conservative asymptotic variance estimate of $\hat{N}_{\text{CMP}}$ is obtained as

$$\hat{\sigma}^2_{\text{CMP}} = \widehat{\text{Var}}(\hat{N}_{\text{CMP}}) = \frac{nf_1 \, \text{e}^{-\hat{\beta}^0}}{n + f_1 \, \text{e}^{-\hat{\beta}^0}} + (\text{e}^{-\hat{\beta}_0})^2 f_1 [1 + f_1 \text{Var}(\hat{\beta}_0)]. \tag{4}$$

4. Simulation study

To better understand the properties of the methods described above, a simulation study was undertaken. This section provides a comprehensive assessment of population size variance estimators performance. We plan the simulation study to cover schemes with different underlying *null* models, with varying population size $N = 100;250;500;1000;5000;10{,}000$ and levels of heterogeneity. In detail, we considering the following data generation settings

(i) *The CMP distribution*: Counts are generated from CMP distribution with parameters

$$\lambda \in \{0.5, 0.8, 1.0\},$$
$$\nu \in \{0.1, 0.5, 0.8\}.$$

(ii) *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

$$k \in \{2, 3, 4\},$$

expected value and variance given respectively by

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

and

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

(iii) *The Generalized Poisson distribution*: Counts are generated from a Generalized Poisson distribution

$$p_x = \frac{\lambda(\lambda + kx)^{x-1}}{x!} \exp(-(\lambda + kx)),$$

where $\lambda > 0$ and $\max(-1, -\lambda/4) \leq k \leq 1$. The expected value and variance are given respectively by

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

and

$$\text{Var}(X) = \frac{\lambda}{(1-k)^3} = \mu \frac{1}{(1-k)^2}.$$

The following parameters are considered to generate counts

$$\lambda \in \{0.8, 1.0, 2.0\}$$

and

$$k \in \{-0.1, 0.1, 0.3\}.$$

The setting (i) covers situations where the data are generated from the CMP distribution, with different levels of heterogeneity/overdispersion, whilst the settings (ii) and (ii) are considered to investigate what happens if we leave the family. We draw 1000 samples from each *null* model. Please, the reader be aware that λ have different meanings for the CMP/GP and NB.

4.1. Bias and variance

Tables 1–3 show the estimator's behaviour under different settings. We evaluate the performance of the estimator in terms of relative bias

$$\frac{1}{N}[E(\hat{N}) - N]$$

and relative variance

$$\frac{1}{N^2}(\widehat{\text{Var}}(\hat{N})).$$

The estimator performs very well under the CMP(λ, ν) data generation process. In particular, its behaviour is satisfactory as well as the data show a high degree of heterogeneity, that is, for $\nu \to 0$. As expected, increasing N lead to better estimates in terms of both relative bias and variance of the estimates. In detail, the CMP estimator shows very good finite sample properties if the data generation process follows a CMP distribution. As long as the sample size increases, that is, $N \to \infty$, the bias reduces and the estimate of the population size tends to the true value. In finite samples, it performs very well in the presence of overdispersion and/or large populations sizes. As the level of overdispersion reduces, that is, by increasing ν values, the CMP estimator is less accurate and even precision could be poor for small population sizes. Under Negative Binomial data generation settings, the simulation results show that the CMP-based estimator has small relative biases and variances if data are generated from a model that does not belong to the CMP family. A similar behaviour arises under Generalized Poisson data generation process, that is, the proposed estimator is robust against model specification. However, a persistent overestimation occurs for small population sizes and λ values, that may become an issue as the variance approaches to zero.

Table 1. Setting (i): Data are generated according to a CMP(λ, ν) distribution.

λ	ν	$N = 100$	$N = 250$	$N = 500$	$N = 1000$	$N = 5000$	$N = 10{,}000$
				Relative bias			
0.8	0.10	0.0110	0.0056	0.0076	0.0054	0.0020	0.0007
	0.50	0.1057	0.0595	0.0501	0.0254	0.0066	0.0034
	0.80	0.1726	0.1276	0.0709	0.0357	0.0090	0.0016
	1.25	0.5005	0.3084	0.1444	0.0860	0.0163	0.0107
1.0	0.10	0.0095	0.0080	0.0050	0.0016	0.0009	0.0001
	0.50	0.0704	0.0459	0.0209	0.0137	0.0020	0.0016
	0.80	0.1236	0.0833	0.0467	0.0278	0.0048	0.0032
	1.25	0.3433	0.1643	0.0754	0.0402	0.0065	0.0075
1.2	0.10	0.0084	0.0037	0.0013	0.0011	0.0002	0.0001
	0.50	0.0442	0.0289	0.0140	0.0068	0.0014	0.0005
	0.80	0.1020	0.0461	0.0278	0.0170	0.0026	0.0015
	1.25	0.1961	0.1178	0.0573	0.0296	0.0079	0.0015
				Relative variance			
0.8	0.10	0.0145	0.0062	0.0040	0.0022	0.0005	0.0002
	0.50	0.0712	0.0366	0.0194	0.0092	0.0017	0.0009
	0.80	0.1450	0.0811	0.0431	0.0207	0.0037	0.0017
	1.25	0.3063	0.2198	0.1246	0.0578	0.0093	0.0040
1.0	0.10	0.0050	0.0023	0.0011	0.0006	0.0001	0.0001
	0.50	0.0347	0.0181	0.0079	0.0040	0.0008	0.0003
	0.80	0.0921	0.0424	0.0201	0.0091	0.0016	0.0008
	1.25	0.2572	0.1270	0.0616	0.0239	0.0039	0.0021
1.2	0.10	0.0009	0.0004	0.0002	0.0001	0.0000	0.0000
	0.50	0.0207	0.0086	0.0042	0.0017	0.0003	0.0002
	0.80	0.0533	0.0212	0.0093	0.0042	0.0007	0.0004
	1.25	0.1522	0.0722	0.0302	0.0117	0.0021	0.0009

Note: Relative bias and relative variance of the CMP-based estimator.

Table 2. Setting (ii): Data are generated according to a NB(λ,k) distribution.

k	λ	$N = 100$	$N = 250$	$N = 500$	$N = 1000$	$N = 5000$	$N = 10{,}000$
				Relative bias			
2	0.20	0.1329	0.1303	0.1067	0.1002	0.0997	0.0877
	0.40	0.0710	0.0742	0.0700	0.0652	0.0554	0.0534
	0.60	0.0481	0.0441	0.0408	0.0360	0.0310	0.0305
	0.80	0.0227	0.0185	0.0156	0.0128	0.0121	0.0121
3	0.20	0.1827	0.1399	0.1244	0.1116	0.0784	0.0734
	0.40	0.0799	0.0657	0.0546	0.0470	0.0397	0.0397
	0.60	0.0363	0.0267	0.0209	0.0212	0.0182	0.0183
	0.80	0.0097	0.0066	0.0059	0.0048	0.0044	0.0044
4	0.20	0.1598	0.1473	0.1340	0.1229	0.0842	0.0735
	0.40	0.0579	0.0460	0.0368	0.0291	0.0275	0.0269
	0.60	0.0172	0.0131	0.0113	0.0103	0.0095	0.0093
	0.80	0.0255	0.0031	0.0020	0.0017	0.0014	0.0013
				Relative variance			
2	0.20	0.1461	0.1065	0.0755	0.0441	0.0146	0.0087
	0.40	0.0408	0.0214	0.0127	0.0072	0.0015	0.0007
	0.60	0.0095	0.0044	0.0026	0.0011	0.0002	0.0001
	0.80	0.0014	0.0005	0.0002	0.0001	0.001	0.0000
3	0.20	0.1454	0.0759	0.0475	0.0333	0.0066	0.0033
	0.40	0.0214	0.0109	0.0047	0.0024	0.0004	0.0002
	0.60	0.0032	0.0011	0.0005	0.0002	0.001	0.0000
	0.80	0.0003	0.0001	0.0001	0.0001	0.0000	0.0000
4	0.20	0.1279	0.0775	0.0514	0.0305	0.0068	0.0033
	0.40	0.0102	0.0042	0.0017	0.0009	0.0001	0.0001
	0.60	0.0011	0.0003	0.0001	0.0001	0.0001	0.0000
	0.80	0.0110	0.0001	0.0001	0.0000	0.0000	0.0000

Note: Relative bias and relative variance of the CMP-based estimator.

Table 3. Setting (iii): Data are generated according to a GP(λ,k) distribution.

λ	k	$N = 100$	$N = 250$	$N = 500$	$N = 1000$	$N = 5000$	$N = 10{,}000$
				Relative bias			
0.8	−0.1	0.6228	0.2159	0.0168	−0.0746	−0.1314	−0.1358
	0.1	0.1698	0.1610	0.1225	0.1088	0.0767	0.0748
	0.3	0.0139	0.0301	0.0391	0.0472	0.0578	0.0614
1	−0.1	0.2701	0.0402	−0.0203	−0.0571	−0.0840	−0.0876
	0.1	0.1607	0.1160	0.0980	0.0696	0.0594	0.0583
	0.3	0.0639	0.0748	0.0732	0.0736	0.0749	0.0745
2	−0.1	0.0292	0.0022	−0.0076	−0.0141	−0.0193	−0.0196
	0.1	0.0587	0.0377	0.0276	0.0226	0.0177	0.0175
	0.3	0.0750	0.0506	0.0457	0.0435	0.0379	0.0388
				Relative variance			
0.8	−0.1	0.4131	0.2315	0.0989	0.0338	0.0050	0.0022
	0.1	0.1074	0.0712	0.0405	0.0224	0.0044	0.0020
	0.3	0.0369	0.0167	0.0092	0.0051	0.0010	0.0005
1	−0.1	0.2234	0.0835	0.0373	0.0160	0.0022	0.0010
	0.1	0.0876	0.0443	0.0254	0.0110	0.0022	0.0011
	0.3	0.0319	0.0161	0.0088	0.0051	0.0013	0.0007
2	−0.1	0.0151	0.0045	0.0016	0.0008	0.0002	0.0001
	0.1	0.0139	0.0046	0.0020	0.0009	0.0002	0.0001
	0.3	0.0100	0.0040	0.0021	0.0010	0.0002	0.0001

Note: Relative bias and relative variance of the CMP-based estimator.

4.2. Comparing estimates of uncertainty

The methods are firstly compared in terms of how well they estimate the true standard error, which is one of the aims of this research. Let us focus on the (i) setting (see Figure 1), the TB performs very well and represents the best possible choice for the estimation of sample size standard error. However, in practice, the TB can be used only if the population size is known. In real data analyses, it is unlikely that the population size is known and, thus, although the TB is a valid method to estimate uncertainty in CR data, an alternative should be considered. The RB approach tends to persistently underestimate the uncertainty in the population size estimates. In particular for higher levels of dispersion (i.e. for small values of ν), such an effect does not disappear neither for very large N. By reducing heterogeneity in the data, approaching the Poisson distribution, even the RB performs reasonably. Such a behaviour is somehow expected as the RB approach uses only observed frequencies, ignoring the zero-truncation in the data. The IB represents a valid and practical alternative to the TB approach. The IB approach performs similarly as the TB and is valuable as it relies on available observations and an estimate of the population size.

All bootstrap methods may require a considerable amount of time, as a huge number (1000 in our case) of bootstrapped samples is required to get reliable uncertainty estimates. Thus, a more straightforward and less computational-intensive approach could be pursued. The approximated formulation proposed in Section 3.1 can be an easily computable alternative to bootstrap approaches. As expected, the approximated variance performs very well as the population size increases. However, for small sample sizes, it performs poorly. This is in line with the idea that the approximation is asymptotically valid. A further drawback of the use of such an approximation is revealed by our simulation study, for weak overdispersion cases the approximation does not have a satisfactory behaviour and even the RB approach performs better as the data approaches the Poisson distribution. Similar conclusions can be drawn even under the Negative Binomial and Generalized Poisson data generation settings (see Figures 2–4). The formula-based approximation tends to overestimate the true standard error for small population sizes, whilst provides reasonable estimates as the population sizes and λ values increase. The TB approach is still the best one, followed by the IB that is confirmed to be the one to use in practice.

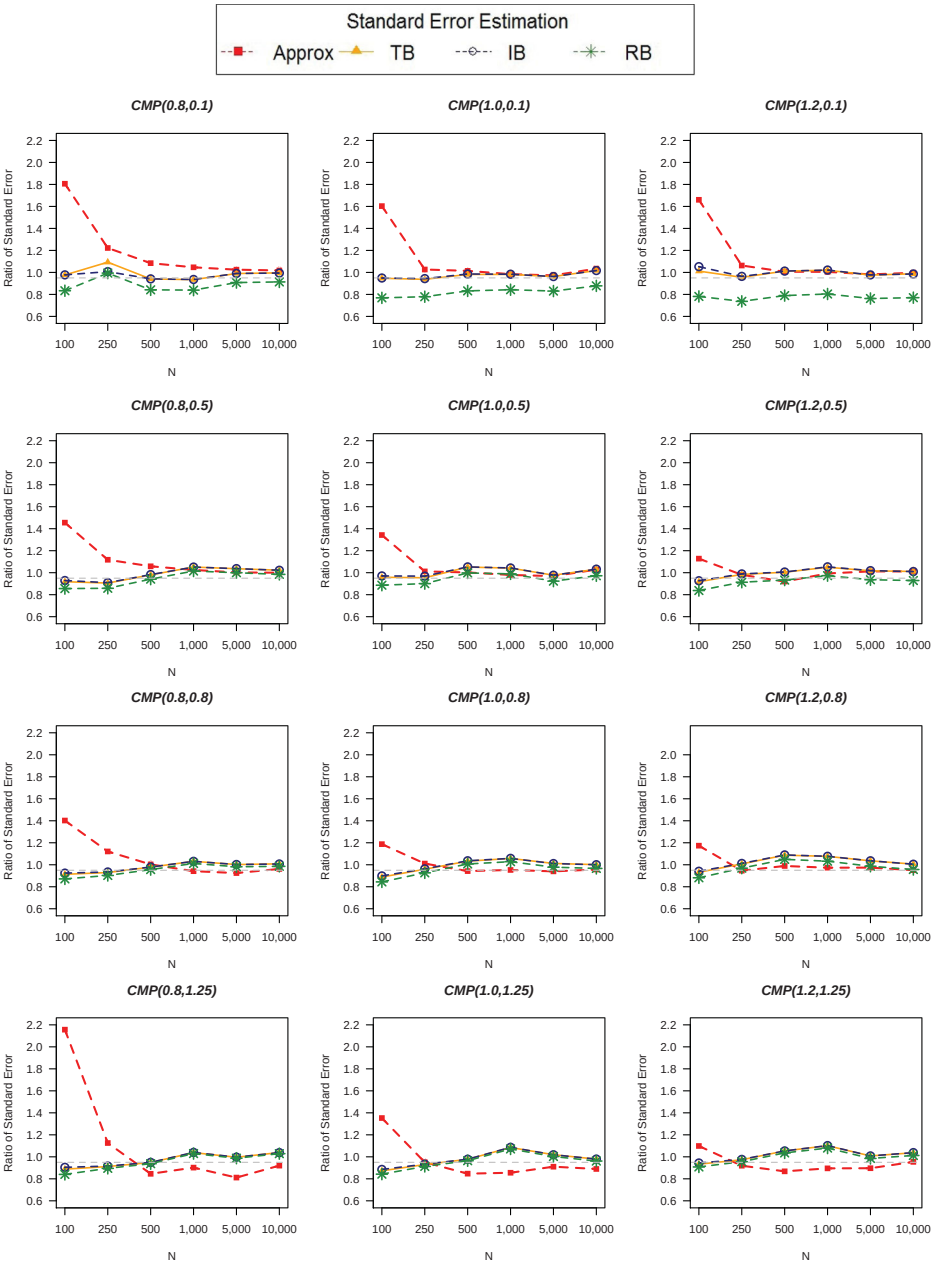

Figure 1. Setting (i): Data are generated according to a CMP(λ, ν) distribution. Ratio of bootstrapped/approximated standard errors over the Monte Carlo standard errors.

4.3. Comparing confidence intervals

We then used the approximation-based and bootstrap methods to derive 95% quantile confidence intervals for each data set. Using these intervals, we ascertained the coverage proportions for each of the methods. These results shed light on the confidence we can put on the obtained estimates and the related uncertainty. Confidence intervals are computed in different ways. A common procedure is to approximate 95% confidence interval for the true population size by the $\hat{N} \mp z_{0.975}\hat{\sigma}_{\text{CMP}}$, where $\hat{\sigma}_{\text{CMP}}$

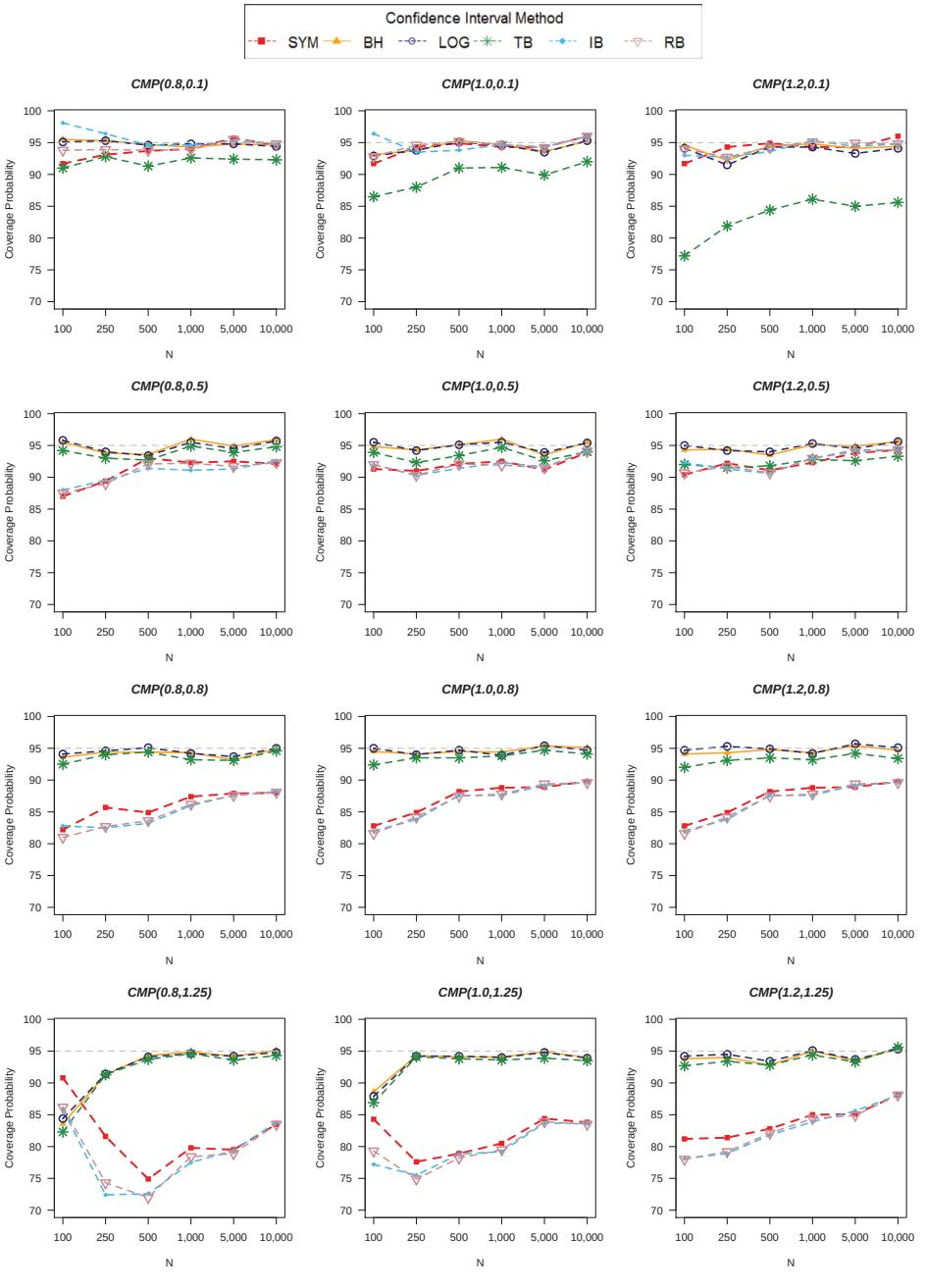

Figure 2. Setting (i): Data are generated according to a CMP($\lambda;\nu$) distribution Coverage probabilities.

is the estimated standard error in (4). This is referred to as a symmetric confidence interval (SYM). However, the construction of the symmetric confidence intervals is based on the large-sample normality for population size estimators. Several drawbacks for this method have been highlighted in [30]: the sampling distribution could be skewed, the lower bound of the resulting interval may be less than the number of units captured, the coverage probabilities may be unsatisfactory. To overcome these issues, coverage of the Burnham confidence interval (BH) $(n + (\hat{N} - n)/c; n + (\hat{N} - n)c)$,

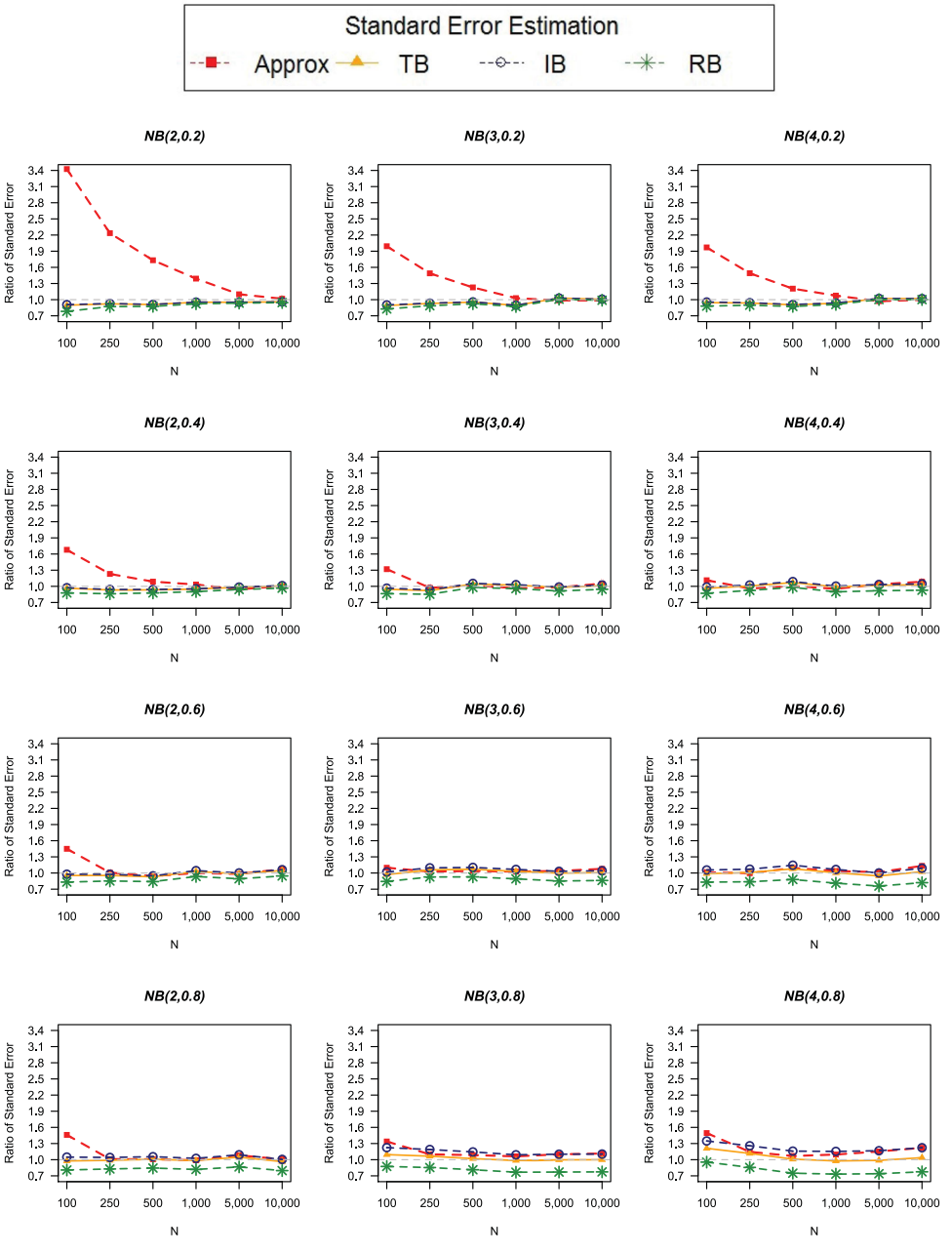

Figure 3. Setting (ii): Data are generated according to a NB(λ;k) distribution. Ratio of bootstrapped/approximated standard errors over the Monte Carlo standard errors.

where

$$c = \exp\left\{ z_{0.975} \left[\log\left(1 + \frac{\hat{\sigma}^2_{\mathrm{CMP}}}{(\hat{N} - n)^2} \right) \right]^{1/2} \right\},$$

is also evaluated [10,31]. We further suggest to look at intervals obtained by using a log-transformation of $\hat{N}$. From the log-normal distribution, it follows that $\log \hat{N}$ has mean $\log N - \frac{1}{2}$

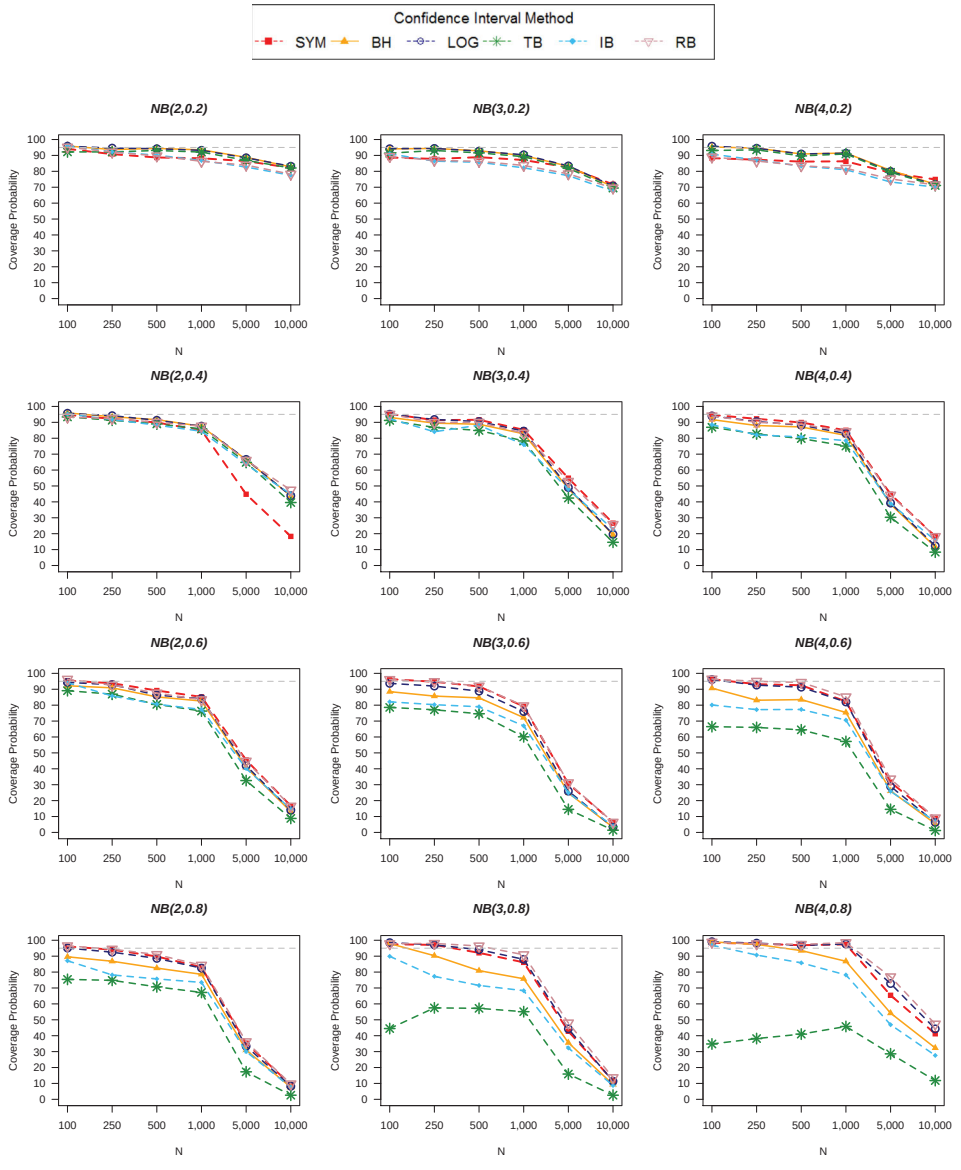

Figure 4. Setting (ii): Data are generated according to a NB(λ; k) distribution. Coverage probabilities.

$\log(1+\sigma^2_{\text{CMP}}/N^2)$ and variance $\log(1+\sigma^2_{\text{CMP}}/N^2)$. Plugging in estimates for σ^2_{CMP} and N leads to a confidence interval for $\log N$ (LOG) given by

$$\log\hat{N}+\frac{1}{2}\log(1+\hat{\sigma}^2_{\text{CMP}}/\hat{N}^2)\mp z_{0.975}\sqrt{\log(1+\hat{\sigma}^2_{\text{CMP}}/\hat{N}^2)}.$$

Taking the anti-logs provides the final form of the confidence interval for N [32]. Other approaches can be pursued to get confidence intervals [30].

For the bootstrap methods, considering all estimates $\hat{N}_b,\ b=1,\ldots,B$ from B bootstrapped samples results in an empirical distribution around the true value. From this distribution we can compute the standard error $\hat{\sigma}$ of the parameter by taking the sample standard deviation of the resulting distribution. The approximate 95% confidence interval of the population size $\hat{N}$ can be obtained using

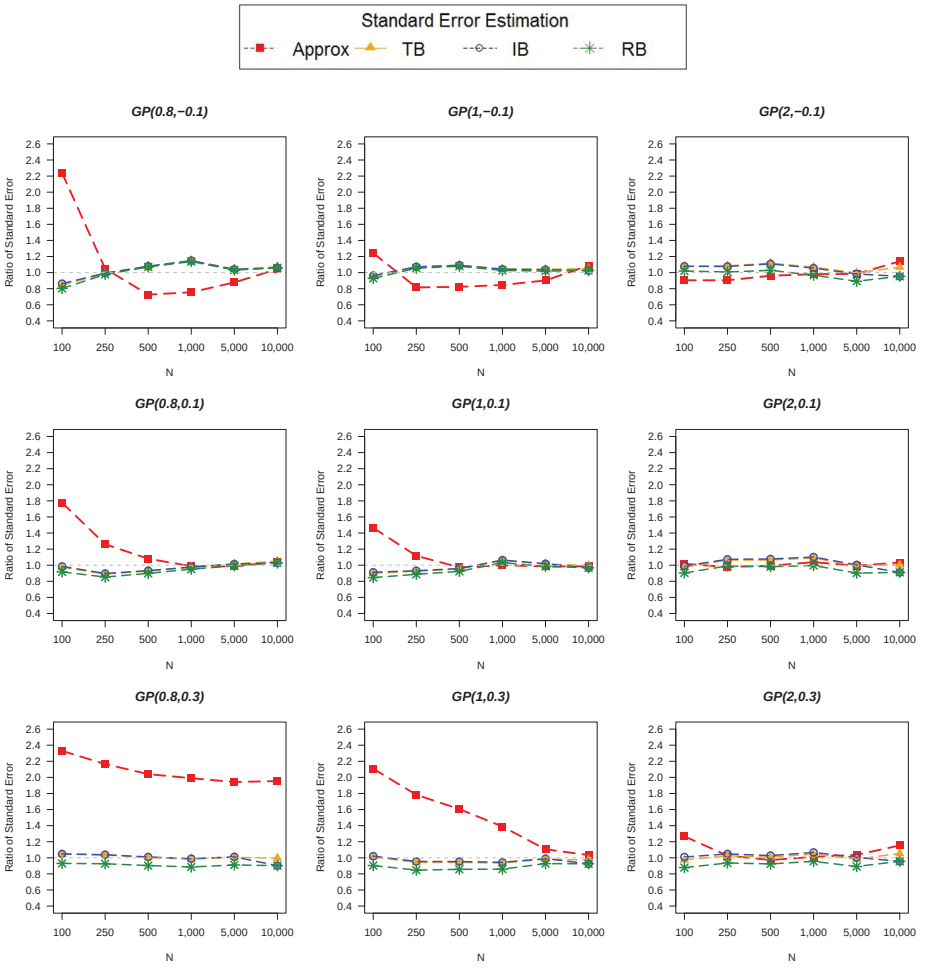

Figure 5. Setting (iii): Data are generated according to a GP(λ;k) distribution. Ratio of bootstrapped/approximated standard errors over the Monte Carlo standard errors.

the percentile method as follows: order $\hat{N}_b$ from the smallest to largest, and denote the ordered list by $\hat{N}_{(b)}$; the approximate 95% confidence limits are then given by $\hat{N}_{(B+1)*0.025}$ and $\hat{N}_{(B+1)*0.975}$, both rounded to the nearest integer value.

Overall, we compare six methods to get confidence intervals of the CMP estimator. Results are summarized in Figures 5, 3 and 6 for the three considered settings, respectively. Under the CMP data generation process, approximation methods provide coverages close to the nominal 95% value for $\nu = 0.1$, that is, high level of overdispersion; whilst the RB suffers in providing reasonable coverage results. Therefore, if the dispersion parameter approaches zero and population sizes are moderate or large, we suggest to construct confidence interval by any of the approximation-based approaches because they are easy to compute and do not require any computational-intensive methods. Interestingly, reducing the level of overdispersion, all approximation-based methods tend to be inappropriate, as under-coverage occurs. Thus, resampling approaches should be preferred in these situations, and the IB approach should be used in practice. Under the Negative Binomial and Generalized Poisson data generation processes, the IB, SYM and LOG methods to build confidence intervals perform similarly. The RB approach should be avoided in most situations, having the worst performance. All considered approaches provide poor coverage as the population size increases. This is mainly due to the very

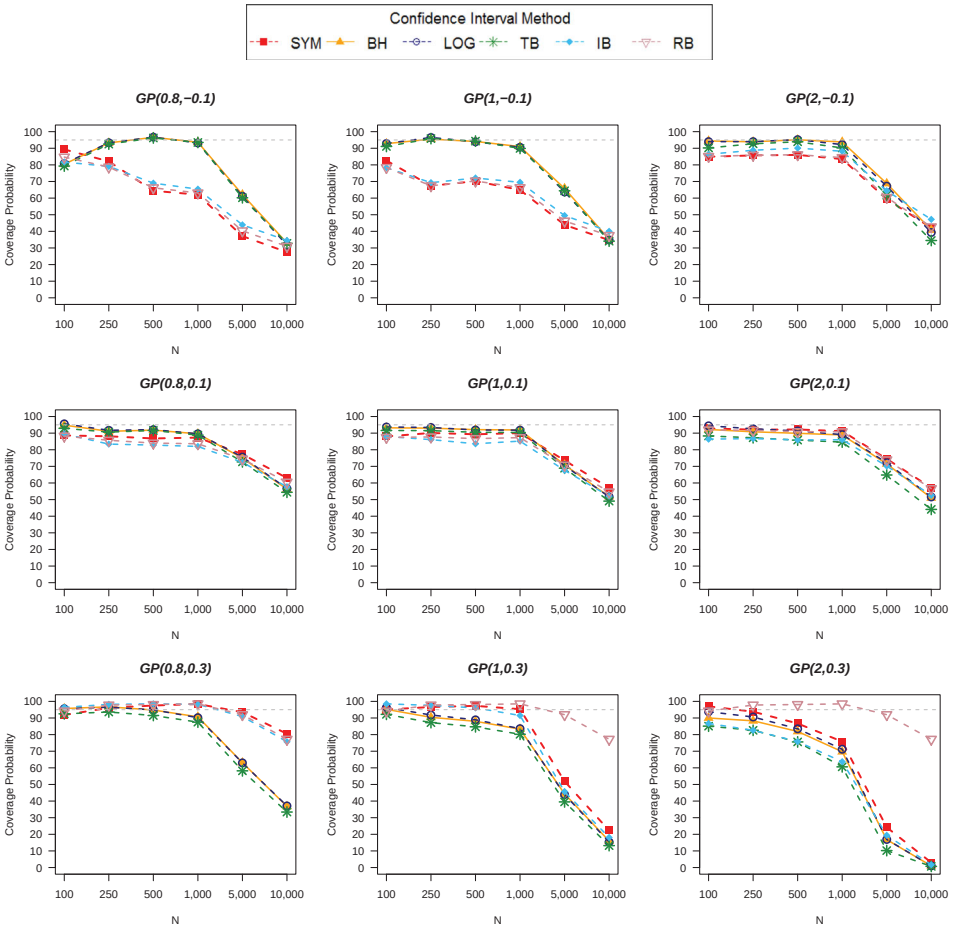

Figure 6. Setting (iii): Data are generated according to a GP distribution. Coverage probabilities.

small estimated variances under these settings that lead to narrow confidence intervals centred to a slightly biased estimate of the true sample size.

5. Real data examples

In the following we estimate population sizes through the CMP estimator so far considered in four well-known benchmark datasets. Data are provided in Table 4. Graphical data inspections through the (log) ratio-plot are provided in Figure 7. In two cases (the Golf-tees and the Taxicabs data) we know the true population size and, accordingly, the TB approach can be also considered. We would like to provide more insights on the uncertainty of the estimates in real data applications, focusing on implications of using different method to estimate such an uncertainty. Population size estimates and

Table 4. Data used in the empirical analyses.

Name	Source	f_0	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
Golf-tees	Borchers et al. [33]	88	46	28	21	13	23	14	6	11
Taxicabs A	Carothers [34]	137	142	81	49	7	3	1		
Hares	Otis et al. [5]	n.a.	25	22	13	5	1	2		
Cholera	Mao and Lindsay [37]	n.a	32	16	6	1				

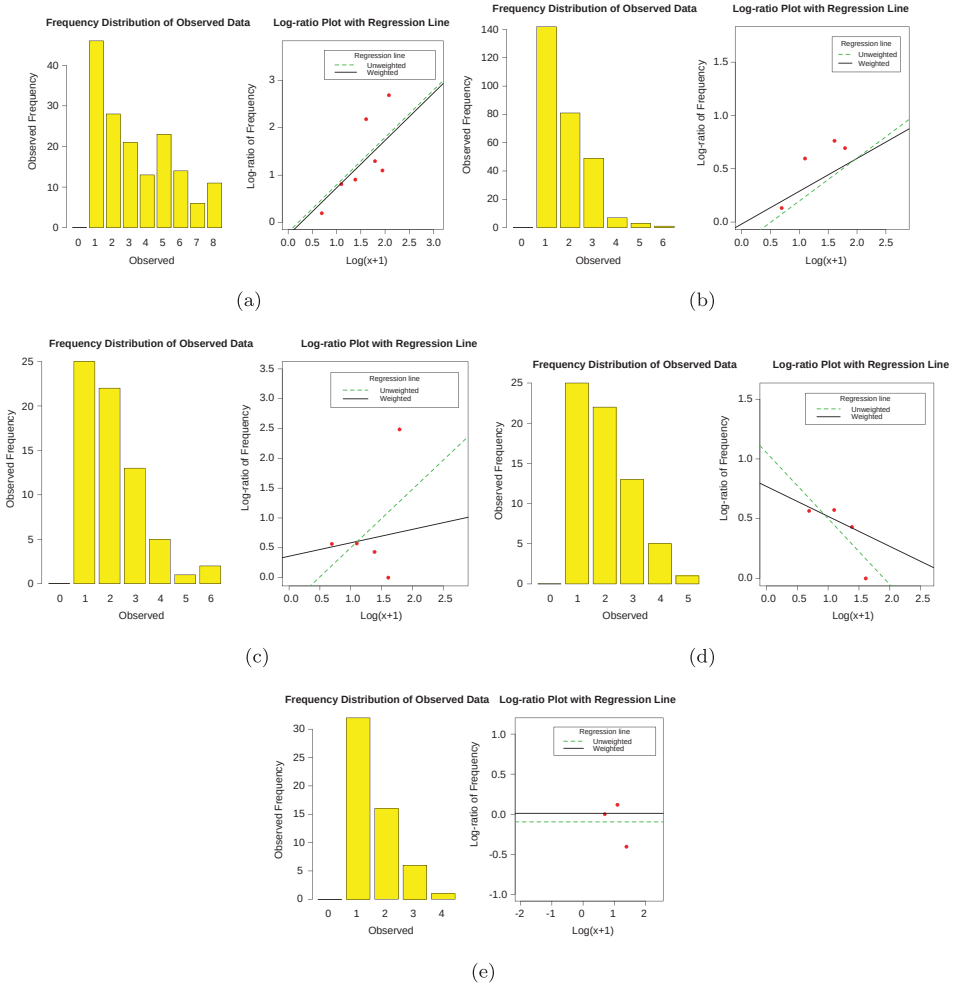

Figure 7. Real data analysis: Ratio-plots. (a) Golf-tees data, (b) Taxicabs data, (c) Hares data (all), (d) Hares data (reduced) and (e) Cholera data.

confidence intervals are reported in Table 5, along with CMP parameters estimate. Table 6 compares the observed and the estimated frequencies under different distributional assumptions.

5.1. Golf-tees data

We consider the dataset described by [33] involving golf tees. A total of eight individuals recorded the location of the golf tees that they observed in a survey region of 1680 m^2 , either exposed above the surrounding grass, or partly hidden by it, independent of each other. Each individual was essentially regarded as a capture event. The tees differed with respect to size, colour and visibility, so that there is some heterogeneity within the population being observed. A total of 162 groups of tees were found and $f_0 = 88$ group of tees were missed. The observed distribution refers to the count of times each group of tees has been found by eight independent observers.

The ratio-plot in Figure 7(a) suggests for heterogeneity, and, accordingly, we expect that the CMP estimator would perform well. The CMP estimator leads to $\hat{N} = 223$, with $\hat{\lambda} = 0.77$ and $\hat{\nu} = 0$, that is, a Geometric distribution results from parameters estimation. The formula-based variance approximation is larger than those from bootstrap methods, in line with the simulation results for small λ

Table 5. Population size estimation and uncertainty assessment in real data examples.

			Standard error estimation				95% Confidence Intervals			
Name	N	$\hat{N}$	Approx.	σ_{TB}	σ_{IB}	σ_{RB}	Approx.	TB	IB	RB
Golf-tees ($\lambda = 0.77, \nu = 0$)	250	223	33.09	15.11	14.41	11.16	(159–288)[(a)] (168–298)[(b)] (169–301)[(c)]	(193–252)	(196–253)	(203–247)
Taxicabs A ($\lambda = 0.98, \nu = 0.69$)	420	428	91.28	65.75	65.85	64.12	(250–607)[(a)] (284–648)[(b)] (290–662)[(c)]	(348–600)	(348–600)	(353–597)
Hares (all data) ($\lambda = 1.43, \nu = 0.77$)	n.a.	86	12.01	n.a.	15.10	14.43	(66–113)[(a)] (66–113)[(b)] (66–114)[(c)]	n.a.	(68–126)	(71–125)
Hares (reduced data) ($\lambda = 2.16, \nu = 1.25$)	n.a.	78	4.58	n.a.	14.08	13.50	(70–87)[(a)] (68–126)[(b)] (71–125)[(c)]	n.a.	(66–121)	(69–121)
Cholera ($\lambda = 1.01, \nu = 1$)	n.a.	87	7.59	n.a.	11.90	11.89	(73–102)[(a)] (73–103)[(b)] (74–104)[(c)]	n.a.	(67–114)	(67–114)

Note: (a) : Symmetric confidence interval , (b) : Burnham confidence interval and (c) : Logarithm transformation confidence interval

Table 6. Data used in the empirical analyses: observed and estimated recapture frequencies.

Name	Model	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	χ^2
Golf-tees	Observed	46	28	21	13	23	14	6	11	
	Fitted (Poisson)	22	35	37	30	120	11	5	2	86.10
	Fitted (CMP)	38	29	22	17	13	10	8	6	16.66
	Fitted (GP)	29	34	31	24	17	11	7	4	34.62
Taxicabs A	Observed	142	81	49	7	3	1			
	Fitted (Poisson)	140	89	38	12	3	1			6.02
	Fitted (CMP)	139	84	39	15	5	1			7.80
	Fitted (GP)	105	109	54	13	1	0			27.46
Hares (all data)	Observed	25	22	13	5	1	2			
	Fitted (Poisson)	25	22	13	6	2	1			1.67
	Fitted (CMP)	25	21	13	8	3	1			2.55
	Fitted (GP)	27	21	12	5	2	1			0.78
Hares (reduced data)	Observed	25	22	13	5	1				
	Fitted (Poisson)	26	21	12	5	2				0.67
	Fitted (CMP)	25	23	12	5	2				0.13
	Fitted (GP)	28	22	11	4	1				0.94
Cholera	Observed	32	16	6	1					
	Fitted (Poisson)	32	16	5	1					0.2
	Fitted (CMP)	32	16	5	1					0.2
	Fitted (GP)	36	13	4	1					2.14

and ν values. This leads to wider, and still plausible, confidence intervals; whilst the RB confidence interval does not cover the true sample size. Small differences in confidence interval computation, that is, symmetric, Burnham and log-transformed, are also observed.

5.2. Taxicabs data in Edinburgh

A further example of overdisperved count data is given by the Taxicabs data (see Figure 7(b)). Carothers [34] reported that 420 taxi cabs were registered in Edinburgh, Scotland during his mark-recapture study. This closed population was sampled for 10 consecutive days with observation points and times varied among days. Sighting a cab was considered a 'capture'. No taxis were observed on more than six occasions. These data have been analysed many times in the literature using different estimators (e.g. [5,30]). The performance of the CMP estimator is remarkably good

($\hat{N} = 428$), compared to other estimators. In all cases the true N is contained within the confidence intervals, no matter what procedure has been used to obtain them.

5.3. Snowshoe hare data

The snowshoe hare data have been analysed in [35,36]. From a graphical inspection through the ratio-plot (see Figure 7(c)), it is clear that the two animals caught on all occasions create some overdispersion with respect to the Poisson distribution. Therefore, the CMP estimator could be a good candidate to estimate the unknown population size. Parameter estimates are $\hat{\nu} = 0.77$, with $\hat{\lambda} = 1.43$ and the resulting estimated population size is $\hat{N} = 86$, slightly higher than the one estimated in [36]. If we remove the 2 hares caught 6 times (see Figure 7(c)), as [35], the situation changes considerably and underdispersion is estimated ($\hat{\lambda} = 2.16$; $\hat{\nu} = 1.25$), with $\hat{N} = 78$. However, the CMP estimator results to be flexible enough to capture even underdispersion.

Similarly, confidence intervals reflect the effect of the 2 hares caught at all occasions, which we have discussed above. They are very large if the complete data are considered, and much smaller if those two hares are left out of the analysis as unrepresentative of the unobserved part of the population. Bootstrap intervals are larger than those obtained by approximating the variance of the sample size estimator, in line with the simulation results. The Burnham- and the log-transformed-based intervals are more in line with the bootstrap ones, confirming that under the underdispersion case, assuming a symmetric confidence interval may lead to unreliable inference.

5.4. Cholera data

The example stems from [37] and has been discussed previously by others. A cholera epidemic affected a village in India. For the cholera epidemic data, evidence has been provided for the Poisson distribution, confirmed by looking at the ratio-plot in Figure 7(e) which displays a horizontal line. Indeed, the CMP estimator approaches the Poisson distribution, as $\hat{\nu} = 1$, that is, the proposed estimator can be used even if homogeneity is ensured. 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 .85. Under the homogeneity assumption, all formula-based confidence intervals behave similarly, whilst bootstrapped ones are wider and no differences are obtained between IB and RB confidence intervals.

6. Conclusions

Although CR methods are widely used approaches to estimate unknown population size, especially through the use of `Rcapture` or `mra` packages in R or of the `recap` module in STATA, little attention had previously been paid to the investigation of the uncertainty surrounding population size estimates.

Here, we provided several insights on the behaviour of bootstrap methods for variance estimation. Although some studies have already considered bootstrap methods to estimate uncertainty in the population size estimation, methods have not been compared through simulations and their behaviour was not study in depth. Here, three bootstrap methods have been considered: the TB, the RB, and the imputed bootstrap. What works and what does not? It is very clear that the RB does not work, in the sense that it does underestimate the true variance. This is independent of the fact that the model holds or not. This result indicates that current practice (using RB method in CR) should be discontinued.

The TB works, if the model holds or not, but it cannot be used in practice. This leaves the imputed bootstrap which seems to work like the TB but only if the model is valid. Hence it behaves similar to the parametric bootstrap. The results are encouraging to investigate the imputed bootstrap in further CR models and truncated data modelling.

Results are consistent with the work of [21] that shows confidence intervals obtained by bootstrap methods can be more accurate than those found analytically, using asymptotic approximations, for both Bailey's and Chapman's nearly unbiased estimators.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

A. Maruotti http://orcid.org/0000-0001-8377-9950

References

[1] Farcomeni A, Scacciatelli D. Heterogeneity and behavioral response in continuous time capture—recapture, with application to street cannabis use in Italy. Ann Appl Stat. 2013;7:2293–2314.
[2] Hwang W-H, Huggins R. An examination of the effect of heterogeneity on the estimation of population size using capture–recapture data. Biometrika. 2005;92:229–233.
[3] Böhning D, Schön D. Nonparametric maximum likelihood estimation of population size based on the counting distribution. J R Stat Soc Ser C (Appl Stat). 2005;54:721–737.
[4] van der Hejiden PG, Bustami R, Cruyff MJ, et al. Point and interval estimation of the population size using the truncated Poisson regression model. Stat Modell. 2003;3:305–322.
[5] Otis DL, Burnham K, White G, et al. Statistical inference from capture data on closed animal populations. Wildl Monogr. 1978;62:3–135.
[6] Pollock K, Nichols J, Brownie C, et al. Statistical inference for capture–recapture experiments. Wildl Monogr. 1990;107:3–97.
[7] Farcomeni A. A general class of recapture models based on the conditional capture probabilities. Biometrics. 2016;72:116–124.
[8] Sanathanan L. Estimating the size of a multinomial population. Ann Math Statist. 1972;43:142–152.
[9] Rocchetti I, Alfó M, Böhning D. A regression estimator for mixed binomial capture–recapture data. J Stat Plan Inf. 2014;145:165–178.
[10] Tounkara F, Rivest L-P. Mixture regression models for closed population capture–recapture data. Biometrics. 2015;71:721–730.
[11] Böhning D, Dietz E, Kuhnert R, et al. Mixture models for capture–recapture count data. Stat Meth Appl. 2005;14:29–43.
[12] Mao CX, Yang N, Zhong J. On population size estimators in the Poisson mixture model. Biometrics. 2013;69:758–765.
[13] Shmueli G, Minka TP, Kadane JB, et al. A useful distribution for fitting discrete data: revival of the Conway–Maxwell–Poisson distribution. J R Stat Soc: Ser C (Appl Stat). 2005;54:127–142.
[14] Cancho VG, Ortega EMM, Barriga GDC, et al. The Conway–Maxwell–Poisson-generalized gamma regression model with long-term survivors. J Stat Comput Simul. 2011;81:1461–1481.
[15] Borges P, Rodrigues J, Balakrishnan N, et al. A COM – Poisson type generalization of the binomial distribution and its properties and applications. Stat Probab Lett. 2014;87:158–166.
[16] Handcock MS, Gile KJ, Mar CM. Estimating the size of populations at high risk for HIV using respondent-driven sampling data. Biometrics. 2015;71:258–266.
[17] Chakraborty S, Ong SH. A COM–Poisson type generalization of the negative binomial distribution. Commun Stat Theory Methods. 2016;45:4117–4135.
[18] Sur P, Shmueli G, Bose S, et al. Modeling bimodal discrete data using Conway–Maxwell–Poisson mixture models. J Bus Econ Stat. 2015;33:352–365.
[19] Gillispie SB, Green CG. Approximating the Conway–Maxwell–Poisson distribution normalization constant. Statistics. 2015;49:1062–1073.
[20] Böhning D, Baksh MF, Lerdsuwansri R, et al. Use of the ratio plot in capture–recapture estimation. J Comput Graph Stat. 2013;22:135–155.
[21] Buckland ST, Garthwaite PH. Quantifying precision of mark – recapture estimates using the bootstrap and related methods. Biometrics. 1991;47:255–268.
[22] Norris JL, Pollock KH. Including model uncertainty in estimating variances in multiple capture studies. Environ Ecol Stat. 1996;3:235–244.
[23] Zwane EN, van der Heijden PGM. Implementing the parametric bootstrap in capture–recapture models with continuous covariates. Stat Probab Lett. 2003;65:121–125.
[24] Tilling K, Sterne JAC. Capture–recapture models including covariate effects. Am J Epidemiol. 1999;149:392–400.

[25] McCrea R, Morgan B. Analysis of Capture–Recapture Data. Boca Raton: Chapman & Hall, CRC Press; 2014.
[26] Nadarajah S. Useful moment and CDF formulations for the COM – Poisson distribution. Stat Pap. 2007;50:617–622.
[27] Anan O, Böhning D, Maruotti A. Population size estimation and heterogeneity in capture–recapture data: a linear regression estimator based on the Conway–Maxwell–Poisson distribution. Stat Methods Appl. 2017;26:49–79.
[28] Buckland ST, Garthwaite PH. Quantifying precision of mark-recapture estimates using the bootstrap and related methods. Biometrics. 1991;47:255–268.
[29] Böhning D. A simple variance formula for population size estimators by conditioning. Stat Methodol. 2008;5:410–423.
[30] Chao A. Estimating the population size for capture-recapture data with unequal catchability. Biometrics. 1987;43:783–791.
[31] Burnham KP, Overton WS. Estimation of the size of a closed population when capture probabilities vary among animals. Biometrika. 1978;65:625–633.
[32] Köse T, Orman M, Ikiz F, et al. Extending the Lincoln–Petersen estimator for multiple identifications in one source. Stat Med. 2014;33:4237–4249.
[33] Borchers D, Buckland S, Zucchini W. Estimating animal abundance. Closed populations. London: Springer; 2004.
[34] Carothers AD. Capture-recapture methods applied to a population with known parameters. J Anim Ecol. 1973;42:125–146.
[35] Cormack RM. Log-linear models for capture–recapture. Biometrics. 1989;45:395–413.
[36] Agresti A. Simple capture–recapture models permitting unequal catchability and variable sampling effort. Biometrics. 1994;50:494–500.
[37] Mao CX, Lindsay BG. Tests and diagnostics for heterogeneity in the species problem. Comput Stat Data Anal. 2003;41:389–398.