Biometrics 69, 1033–1042
December 2013
DOI: 10.1111/biom.12082

A Generalization of Chao's Estimator for Covariate Information

Dankmar Böhning,[1,*] Alberto Vidal-Diez,[1] Rattana Lerdsuwansri,[2]
Chukiat Viwatwongkasem,[3] Mark Arnold[4]

[1]Southampton Statistical Sciences Research Institute & School of Mathematics, University of Southampton, Southampton SO17 1BJ, U.K.
[2]Department of Mathematics and Statistics, Thammasat University, Bangkok, Thailand
[3]Faculty of Public Health, Department of Biostatistics, Mahidol University, Bangkok, Thailand
[4]Biomathematics and Statistics Unit, Animal Health and Veterinary Laboratories Agency (AHVLA) Loughborough, Leicestershire LE12 5RB, U.K.
[*]*email:* d.a.bohning@soton.ac.uk

Summary. This note generalizes Chao's estimator of population size for closed capture–recapture studies if covariates are available. Chao's estimator was developed under unobserved heterogeneity in which case it represents a lower bound of the population size. If observed heterogeneity is available in form of covariates we show how this information can be used to reduce the bias of Chao's estimator. The key element in this development is the understanding and placement of Chao's estimator in a truncated Poisson likelihood. It is shown that a truncated Poisson likelihood (with log-link) with all counts truncated besides ones and twos is equivalent to a binomial likelihood (with logit-link). This enables the development of a generalized Chao estimator as the estimated, expected value of the frequency of zero counts under a truncated (all counts truncated except ones and twos) Poisson regression model. If the regression model accounts for the heterogeneity entirely, the generalized Chao estimator is asymptotically unbiased. A simulation study illustrates the potential in gain of bias reduction. Comparisons of the generalized Chao estimator with the homogeneous zero-truncated Poisson regression approach are supplied as well. The method is applied to a surveillance study on the completeness of farm submissions in Great Britain.

Key words: Bias reduction; Chao's estimator; Closed capture–recapture; Covariate modelling.

1. Introduction

For integer N, we consider a sample of counts $Y_1, Y_2, \ldots, Y_N \in \{0, 1, 2, \ldots\}$ arising with a mixture probability density function

$$p_y = \int_0^\infty p(y|\lambda)q(\lambda)\,\mathrm{d}\lambda, \tag{1}$$

where the mixing density $q(\lambda)$ is unspecified and the mixture kernel $p(x|\lambda)$ comes from the Poisson family $p(y|\lambda) = \mathrm{Po}(y|\lambda) = \exp(-\lambda)\lambda_y/y!$. Whenever $Y_i = 0$ unit i remains unobserved, so that only a zero-truncated sample of size $n = \sum_{y=1}^m f_y$ is observed, where f_y is the frequency of counts with value $Y = y$ and m is the largest observed count. Hence, f_0 and consequently N are unknown. The purpose is to find an estimate of the size N. Since frequently the count variable Y represents repeated identifications of an individual in an observational period, the problem at hand is a special form of the capture–recapture problem (see Bunge and Fitzpatrick, 1993, Wilson and Collins, 1992, or Chao et al., 2001, for a review on the topic).

The sample of counts $Y_1, Y_2, \ldots, Y_N$ can occur in several ways. A target population which might be difficult to count consists of N units. This population might be a wildlife population, a population of homeless people or drug addicts, software errors or animals with a specific disease (see also Hay and Smit, 2003; van Hest et al., 2008; Roberts and Brewer, 2006). Furthermore, let an identification device (a trap, a register, a screening test) be available that identifies unit i at occasion t where $t = 1, \ldots, T$ and T potentially being random and/or unknown itself. Let the binary result be y_{it} where $y_{it} = 1$ means that unit i has been identified at occasion t and $y_{it} = 0$ means that unit i has not been identified at occasion t. The indicators y_{it} might be observed or not, but it is assumed that $y_i = \sum_{t=1}^T y_{it}$ is observed if at least one $y_{it} > 0$ for $t = 1, \ldots, T$. Only if $y_{i1} = y_{i2} = \cdots = y_{iT} = 0$ and, consequently $y_i = 0$, the unit i remains *unobserved*. In this kind of situation the *clustering* occurs by repeated identifications of the same unit, the latter being the cluster.

In another setting, which is also the basis for this work, the clustering occurs by means of a grouping variable such as herds, holdings, farms, households, or villages. In this case, y_{it} denotes if the tth element in cluster i is identified ($y_{it} = 1$) or not ($y_{it} = 0$). In the example given in this article the clusters are holdings of cattle and y_{it} informs about the disease status of tth animal in holding i. Only $y_i = \sum_t y_{it}$ is observed if it is positive. In other examples the cluster corresponds to villages or households, one of the earliest applications of this kind is the cholera-outbreak in a community in India studied by McKendrick (1926) in which the cluster corresponds to households in a village. A more recent example involves cholera occurrence in rural East Pakistan where the cluster

structure consists of villages (see also Mosley, Bart, and Sommer, 1972). Böhning and Del Rio Vilas (2008) used the clustering approach to estimate hidden scrapie in the sheep holding population in Great Britain.

The article is organized as follows. In the next section the estimator of Chao (1987, 1989) is reviewed and positioned into a truncated likelihood approach. In Section 3, the major result is provided which delivers a generalisation of Chao's estimator if covariate information is available. Section 4 provides a small simulation study illustrating the benefits of the approach also in comparison to existing alternatives such as the zero-truncated Poisson and zero-truncated negative-binomial model. Section 5 presents a case study on estimating the completeness of surveillance of farm submissions of material from dead cattle in England. The article closes with a brief discussion of results.

2. Chao's Estimator Revisited

The importance of the mixture $p_y = \int_0^\infty p(y|\lambda)q(\lambda)\,\mathrm{d}\lambda$ can be seen in the fact that it is a natural model for modeling population heterogeneity. There appears to be consensus (see, e.g., Pledger, 2005, for the discrete mixture model approach and Dorazio and Royle, 2005, for the continuous mixture model approach) that a simple model $p(x|\lambda)$ is not flexible enough to capture the variation in the re-capture probability for the different members of most real life populations. Every item might be different, as might be every animal or human being. However, there has recently been also a debate on the identifiability of the binomial mixture model (see Link, 2003, 2006; Holzmann, Munk, and Zucchini, 2006). Furthermore, using the nonparametric maximum likelihood estimate (NPMLE) $\hat{q}(\lambda)$ of the mixing density $q(\lambda)$ in constructing an estimate of the population size $\hat{N} = n/[1 - \int_0^\infty \exp(-\lambda)\hat{q}(\lambda)\,\mathrm{d}\lambda]$ leads to the *boundary problem* implying often unrealistically high values for the estimate of the population size (Wang and Lindsay, 2005, 2008). Hence, a renewed interest has re-occurred in the lower bound approach for population size estimation suggested by Chao (1987). For further developments see also Mao (2008). In the lower bound approach there is neither need to specify a mixing distribution, nor is there need to estimate it. In this sense it is completely non-parametric. To give some details of the lower bound approach consider the Poisson mixture kernel $\exp(-\lambda)\lambda^x/x!$. It follows from the Cauchy–Schwarz inequality that

$$\left(\int_0^\infty \exp(-\lambda)\lambda q(\lambda)\,\mathrm{d}\lambda\right)^2 \le \int_0^\infty \exp(-\lambda)q(\lambda)\,\mathrm{d}\lambda \times \int_0^\infty \exp(-\lambda)\lambda^2 q(\lambda)\,\mathrm{d}\lambda,$$

or equivalently, $p_1^2 \le p_0(2p_2)$. Replacing the theoretical probabilities p_j by their sample estimates f_j/N for $j = 0, 1, 2$, the Chao lower bound estimate $f_1^2/(2f_2)$ for f_0 follows (see Chao, 1987, 1989) since the unknown denominator N cancels out. The estimate for the population size N is $\hat{N}_C = n + f_1^2/(2f_2)$. In the following we develop a likelihood framework in which the estimator of Chao develops. Since the Chao estimator uses only frequencies with counts of 1 and 2, a truncated sample consisting only out of counts of ones and twos might be considered. The associated truncated Poisson probabilities are

$$q_1 = \frac{1}{1+\lambda/2} \text{ and } q_2 = \frac{\lambda/2}{1+\lambda/2}.$$

This truncated sample leads to a binomial log-likelihood $f_1\log(q_1) + f_2\log(q_2)$ which is uniquely maximized for $\hat{q}_2 = 1 - \hat{q}_1 = f_2/(f_1+f_2)$. Since $q_2 = \lambda/(\lambda+2)$ and $q_1 = 2/(\lambda+2)$, the estimate $\hat{\lambda} = 2f_2/f_1$ for the Poisson parameter λ suggested by Zelterman (1988) arises. In the approach of Zelterman the homogeneous Poisson serves only as a working model and it was suggested by Zelterman that the estimate $\hat{N}_Z = \frac{n}{1-\hat{p}_0} = \frac{n}{1-\exp(-\hat{\lambda})}$ is more robust against misspecifications of the Poisson model than the usual maximum likelihood estimate.

THEOREM 1.

(a) *Let* $\log L(\lambda) = f_1\log(q_1) + f_2\log(q_2)$ *with* $q_1 = e^{-\lambda}\lambda/(e^{-\lambda}\lambda + e^{-\lambda}\lambda^2/2) = 2/(\lambda+2)$ *and* $q_2 = e^{-\lambda}\lambda^2/2/(e^{-\lambda}\lambda + e^{-\lambda}\lambda^2/2) = \lambda/(\lambda+2)$ *being the Poisson probabilities truncated to counts of ones and twos. Then* $\log L(\lambda)$ *is maximized for*

$$\hat{\lambda} = 2f_2/f_1.$$

(b)

$$E(f_0|f_1, f_2; \hat{\lambda}) = f_1^2/(2f_2), \text{ for } \hat{\lambda} = 2f_2/f_1.$$

Proof. For the first part, it is clear that $f_1\log(q_1) + f_2\log(q_2)$ is maximal for $\hat{q}_1 = f_1/(f_1+f_2)$, which is attained for $\hat{\lambda} = 2f_2/f_1$. For the second part, we see that with $e_y = E(f_y|f_1, f_2; \lambda) = \mathrm{Po}(y|\lambda)N$ we have the following:

$$\begin{aligned} e_y &= \mathrm{Po}(y|\lambda)N \\ &= \mathrm{Po}(y|\lambda)\left(e_0 + f_1 + f_2 + \sum_{j=3}^\infty e_j\right), \end{aligned}$$

so that

$$\begin{aligned} e_0 + e_3^+ = {} & [1 - \mathrm{Po}(1|\lambda) - \mathrm{Po}(2|\lambda)](e_0 + e_3^+) \\ & + [1 - \mathrm{Po}(1|\lambda) - \mathrm{Po}(2|\lambda)](f_1 + f_2) \end{aligned}$$

with $e_3^+ = \sum_{j=3}^\infty e_j$. Hence

$$e_0 + e_3^+ = \frac{1 - \mathrm{Po}(1|\lambda) - \mathrm{Po}(2|\lambda)}{\mathrm{Po}(1|\lambda) + \mathrm{Po}(2|\lambda)}(f_1 + f_2)$$

and

$$\begin{aligned} e_0 &= \mathrm{Po}(0|\lambda)(f_1 + f_2 + e_0 + e_3^+) \\ &= \mathrm{Po}(0|\lambda)(f_1 + f_2)\left[1 + \frac{1 - \mathrm{Po}(1|\lambda) - \mathrm{Po}(2|\lambda)}{\mathrm{Po}(1|\lambda) + \mathrm{Po}(2|\lambda)}\right] \\ &= \frac{\mathrm{Po}(0|\lambda)}{\mathrm{Po}(1|\lambda) + \mathrm{Po}(2|\lambda)}(f_1 + f_2) = \frac{f_1 + f_2}{\lambda + \lambda^2/2}. \end{aligned}$$

Plugging in the maximum likelihood estimate $\hat{\lambda} = 2f_2/f_1$ for λ yields the desired result. □

Theorem 1 establishes a close connection between the approach by Zelterman and Chao's estimator. It shows that Zelterman's estimator of the Poisson parameter λ arises when all counts are truncated except counts of ones and twos and when the resulting likelihood is maximized. If the correct prediction for f_0 is used, namely the conditional expectation for the truncated Poisson model, the Chao estimator arises. Hence the strong overestimation of the original Zelterman estimator which is occasionally observed in practice (van der Heijden et al., pers. comm., 2006) stems from using a *wrong* conditional expectation. If λ becomes small then Chao's and Zelterman's estimator become identical (Böhning, 2010).

3. Chao's Estimator with Covariates

3.1. *The Generalized Chao Estimator*

Here we will develop a generalization of Chao's estimator for covariate information. This is a generalization in the sense that if there is only an intercept (hence no covariates) the generalization is identical to the original or simple Chao estimator. Consider a sample with covariate information $(Y_1, \mathbf{z}_1), \ldots, (Y_N, \mathbf{z}_N)$ where $\mathbf{z}_i$ is a p-dimensional vector additional information on unit i. We assume that the heterogeneity expressed in the mixture (1) can be captured by means of a Poisson regression model with log-link function

$$\lambda_i = \exp(\alpha + \boldsymbol{\beta}^T \mathbf{z}_i), \tag{2}$$

where $\lambda_i = E(Y_i|\mathbf{z}_i)$ is the conditional Poisson mean with $P(Y_i = y) = \mathrm{Po}(y|\lambda_i)$. The associated truncated Poisson model with all counts truncated besides $Y_i = 1$ and $Y_i = 2$ is

$$P(Y_i = 1) = (1 - q_i) = \frac{1}{1 + \lambda_i/2} \text{ and } P(Y_i = 2) = q_i = \frac{\lambda_i/2}{1 + \lambda_i/2}.$$

Suppose there are M different observed covariate combinations with $n_1 + \cdots + n_M = n$, where n_i is the frequency of covariate combination i. Then the truncated Poisson likelihood is given as

$$\prod_{i=1}^{M} \left(\frac{1}{1+\lambda_i/2}\right)^{f_{i1}} \times \left(\frac{\lambda_i/2}{1+\lambda_i/2}\right)^{f_{i2}} = \prod_{i=1}^{M} \left(\frac{1}{1+\exp(\alpha+\boldsymbol{\beta}^T\mathbf{z}_i)/2}\right)^{f_{i1}} \times \left(\frac{\exp(\alpha+\boldsymbol{\beta}^T\mathbf{z}_i)/2}{1+\exp(\alpha+\boldsymbol{\beta}^T\mathbf{z}_i)/2}\right)^{f_{i2}}, \tag{3}$$

where f_{ij} is the frequencies of counts of j in the ith covariate combination where $j = 1$ or $j = 2$. Note that $n_i = f_{i1} + f_{i2} + \cdots + f_{im}$. Clearly, the likelihood (3) is identical to a conventional binomial logistic likelihood

$$\prod_{i=1}^{M} (1 - q_i)^{f_{i1}} q_i^{f_{i2}} = \prod_{i=1}^{M} \left(\frac{1}{\exp(\alpha' + \boldsymbol{\beta}^T\mathbf{z}_i)}\right)^{f_{i1}} \times \left(\frac{\exp(\alpha' + \boldsymbol{\beta}^T\mathbf{z}_i)}{1 + \exp(\alpha' + \boldsymbol{\beta}^T\mathbf{z}_i)}\right)^{f_{i2}} \tag{4}$$

with $\alpha' = \log(1/2) + \alpha$. Hence maximum likelihood estimates for the truncated Poisson model can be found by means of a logistic regression analysis. Having found estimates $\hat{\alpha}'$ and $\hat{\boldsymbol{\beta}}$ by maximizing the binomial likelihood (4) the estimate for λ_i is provided as

$$\hat{\lambda}_i = 2\frac{\hat{q}_i}{1 - \hat{q}_i} = 2\exp(\hat{\alpha}' + \hat{\boldsymbol{\beta}}^T\mathbf{z}_i) \tag{5}$$

for $i = 1, \ldots, M$.

The next step is to construct an estimate for f_0. For the ith covariate combination an estimate for f_{i0} is found according to Theorem 1 as the expected value of f_{i0} estimated using $\hat{\lambda}_i$:

$$\hat{f}_{i0} = \frac{\mathrm{Po}(0|\hat{\lambda}_i)}{\mathrm{Po}(1|\hat{\lambda}_i) + \mathrm{Po}(2|\hat{\lambda}_i)}(f_{i1} + f_{i2}) = \frac{f_{i1} + f_{i2}}{\hat{\lambda}_i + \hat{\lambda}_i^2/2}.$$

The final estimator arises by summing up over all M different covariate combinations to yield the *generalized Chao estimator*

$$\hat{N}_{\mathrm{GC}} = n + \sum_{i=1}^{M} \frac{\mathrm{Po}(0|\hat{\lambda}_i)}{\mathrm{Po}(1|\hat{\lambda}_i) + \mathrm{Po}(2|\hat{\lambda}_i)}(f_{i1} + f_{i2}) = n + \sum_{i=1}^{M} \frac{f_{i1} + f_{i2}}{\hat{\lambda}_i + \hat{\lambda}_i^2/2}.$$

The estimator achieves a particular simple and attractive format if written in case data format ($n_i = 1$):

$$\hat{N}_{\mathrm{GC}} = n + \sum_{i=1}^{N} \frac{\Delta_i}{\hat{\lambda}_i + \hat{\lambda}_i^2/2} = n + \sum_{i=1}^{f_1+f_2} \frac{1}{\hat{\lambda}_i + \hat{\lambda}_i^2/2}, \tag{6}$$

where $\Delta_i = 1$ if $y_i \in \{1, 2\}$ and $\Delta_i = 0$ otherwise, for $i = 1, \ldots, N$. Here we have assumed the conventional ordering of the sample such that $1, 2, \ldots, f_1 + f_2$ are the first $f_1 + f_2$ observed units of counts of ones and twos, followed by $n - (f_1 + f_2)$ units of observed counts larger than 2, and $n + 1, \ldots, N$ are the remaining $N - n$ unobserved units. Note that $\hat{\lambda}_i$ is provided in (5). We have the following theorem for the generalized Chao estimator.

THEOREM 2. *Let the Poisson regression model (2) hold. Then the generalized Chao estimator is asymptotically unbiased:*

$$\lim_{N\to\infty} \frac{E(\hat{N}_{\mathrm{GC}})}{N} = 1.$$

Proof. We note that $E(n|\hat{\lambda}_1,\ldots,\hat{\lambda}_N)=\sum_{i=1}^{N}[1-\text{Po}(0|\hat{\lambda}_i)]$ and $E(\Delta_i|\hat{\lambda}_i)=\text{Po}(1|\hat{\lambda}_i)+\text{Po}(2|\hat{\lambda}_i)$. Hence,

$$E(\hat{N}_{\text{GC}}|\hat{\lambda}_1,\ldots,\hat{\lambda}_N)=\sum_{i=1}^{N}[1-\text{Po}(0|\hat{\lambda}_i)]+\sum_{i=1}^{N}\frac{\text{Po}(1|\hat{\lambda}_i)+\text{Po}(2|\hat{\lambda}_i)}{\hat{\lambda}_i+\hat{\lambda}_i^2/2},$$

which becomes

$$\sum_{i=1}^{N}[1-\text{Po}(0|\hat{\lambda}_i)]+\sum_{i=1}^{N}\text{Po}(0|\hat{\lambda}_i)\frac{\hat{\lambda}_i+\hat{\lambda}_i^2/2}{\hat{\lambda}_i+\hat{\lambda}_i^2/2}=N.$$

Then argument is completed by observing that $\lim_{N\to\infty}E(\hat{\lambda}_i)=\lambda_i$. □

Note that the generalized Chao estimator reduces to the conventional Chao estimator if there are no covariates. In this case, $\hat{\lambda}_i=\hat{\lambda}=2f_2/f_1$ and

$$\hat{N}_{\text{GC}}=n+\sum_{i=1}^{f_1+f_2}\frac{1}{\hat{\lambda}_i+\hat{\lambda}_i^2/2}=n+\frac{f_1+f_2}{2f_2/f_1+2f_2^2/f_1^2}$$
$$=n+\frac{f_1^2}{2f_2}\frac{f_1+f_2}{f_1+f_2}=\hat{N}_C.$$

This result is easily extended to the case of *stratified* heterogeneity. Suppose the population consists of M strata and let z_{ij} denote the membership indicator for unit i to belong to stratum j ($z_{ij}=1$ if i belongs to stratum j and 0 otherwise) for $j=1,\ldots,M-1$ (here the stratum M serves as reference). Then the generalized Chao estimator coincides with the stratified Chao estimator $\sum_{i=1}^{M}(n_i+\frac{f_{1i}^2}{2f_{2i}})$ assuming that $f_{2i}>0$ for $i=1,\ldots,M$. Here n_i is the observed size of stratum i, f_{1i} and f_{2i} are the frequencies of ones and twos in stratum i, respectively.

THEOREM 3. *Consider the stratified situation as above with $f_{2i}>0$ for $i=1,\ldots,M$. Then the generalized Chao estimator is at least as large as the conventional Chao estimator:*

$$\hat{N}_{\text{GC}}\geq\hat{N}_C.$$

Proof. We know that in this situation the generalized Chao estimator coincides with the stratified Chao estimator

$$\hat{N}_{\text{GC}}=\sum_{i=1}^{M}\left(n_i+\frac{f_{1i}^2}{2f_{2i}}\right).$$

Note that the conventional (unstratified) Chao estimator can be written as

$$\hat{N}_C=\sum_{i=1}^{M}n_i+\frac{(\sum_{i=1}^{M}f_{1i})^2}{2\sum_{i=1}^{M}f_{2i}}.$$

We show $\sum_i\frac{f_{1i}^2}{f_{2i}}\times\sum_i f_{2i}\geq(\sum_i f_{1i})^2$ where the summation goes from 1 to M as above. Recall the Cauchy–Schwarz inequality in the Euclidean space as

$$\left(\sum_i x_iy_i\right)^2\leq\left(\sum_i x_i^2\right)\left(\sum_i y_i^2\right),$$

where x_i and y_i are arbitrary real numbers. The result follows by choosing $x_i=f_{1i}/\sqrt{f_{2i}}$ and $y_i=\sqrt{f_{2i}}$ where $\sqrt{x}$ refers to the positive root of a non-negative number x. □

The result implies that the generalized Chao estimator is at least as large the conventional Chao estimator whether the observed heterogeneity explains all heterogeneity (including the unobserved heterogeneity) or not. This explains some of its potential for bias reduction. The result was proved for the stratified situation, but the conjecture is that it holds in more generality.

3.2. *Standard Errors of the Generalized Chao Estimator*

Standard errors of the generalized Chao estimator are derived following the conditioning techniques used in van der Heijden, Cruyff, and van Houwelingen (2003a) and Böhning (2008). We use the result

$$\text{Var}(\hat{N}_{\text{GC}})=\text{Var}[E(\hat{N}_{\text{GC}}|\Delta_i,i=1,\ldots,N)]+E[\text{Var}(\hat{N}_{\text{GC}}|\Delta_i,i=1,\ldots,N)].$$

For the *first* term, we have

$$E(\hat{N}_{\text{GC}}|\Delta_i,i=1,\ldots,N)=E\left(n+\sum_{i=1}^{N}\frac{\Delta_i}{\hat{\lambda}_i+\hat{\lambda}_i^2/2}\middle|\Delta_i,i=1,\ldots,N\right)\approx\sum_{i=1}^{N}\Delta_iw_i,$$

where $w_i=1+\exp(-\lambda_i)/p_i$ and

$$p_i=P(Y_i=1\text{ or }Y_i=2)=\text{Po}(1|\lambda_i)+\text{Po}(2|\lambda_i).$$

Note that Δ_i is binary with

$$E(\Delta_i)=P(Y_i=1\text{ or }Y_i=2)=\text{Po}(1|\lambda_i)+\text{Po}(2|\lambda_i)=p_i$$

and $\text{Var}(\Delta_i)=p_i(1-p_i)$. Hence, we have

$$\text{Var}\left(\sum_{i=1}^{N}\Delta_iw_i\right)=\sum_{i=1}^{N}p_i(1-p_i)w_i^2,$$

which we estimate as

$$\widehat{\text{Var}}[E(\hat{N}_{\text{GC}}|\Delta_i,i=1,\ldots,N)]=\sum_{i=1}^{N}\frac{\Delta_i}{\hat{p}_i}\hat{p}_i(1-\hat{p}_i)\hat{w}_i^2=\sum_{i=1}^{f_1+f_2}(1-\hat{p}_i)\left[1+\frac{\exp(-\hat{\lambda}_i)}{\hat{p}_i}\right]^2,\tag{7}$$

where $\hat{p}_i=\text{Po}(1|\hat{\lambda}_i)+\text{Po}(2|\hat{\lambda}_i)$ and $\hat{\lambda}_i$ is provided in (5).

For the *second* term $E[\mathrm{Var}(\hat{N}_{\mathrm{GC}}|\Delta_i, i = 1, \ldots, N)]$, we focus on developing an estimator for $\mathrm{Var}(\hat{N}_{\mathrm{GC}}|\Delta_i, i = 1, \ldots, N)$ which we then take as a moment estimator for the expected value. We use the multivariate δ-method for deriving a variance-estimate of $g(\hat{\alpha}', \hat{\boldsymbol{\beta}}) = \sum_{i=1}^{f_1+f_2} 1/[\hat{\lambda}_i + \hat{\lambda}_i^2/2]$, where again $\hat{\lambda}_i = 2\exp(\hat{\alpha}' + \hat{\boldsymbol{\beta}}^T \mathbf{z}_i)$, so that

$$\widehat{\mathrm{Var}}\left(\sum_{i=1}^{f_1+f_2} \frac{1}{[\hat{\lambda}_i + \hat{\lambda}_i^2/2]}\right) = \nabla g(\hat{\alpha}', \hat{\boldsymbol{\beta}})^T \widehat{\mathbf{cov}}(\hat{\alpha}', \hat{\boldsymbol{\beta}}) \nabla g(\hat{\alpha}', \hat{\boldsymbol{\beta}}), \tag{8}$$

where the estimated covariance matrix $\widehat{\mathbf{cov}}(\hat{\alpha}', \hat{\boldsymbol{\beta}})$ of the regression parameter estimates $\hat{\alpha}'$ and $\hat{\boldsymbol{\beta}}$ is readily available from the logistic regression in (4) as the inverse of the Fisher information matrix. Here $\nabla g(\hat{\alpha}', \hat{\boldsymbol{\beta}})$ denotes the vector of partial derivatives $(\frac{\partial g}{\partial \alpha'}, \frac{\partial g}{\partial \beta_1}, \ldots, \frac{\partial g}{\partial \beta_p})^{\mathrm{T}}$ evaluated at $(\hat{\alpha}', \hat{\boldsymbol{\beta}})^{\mathrm{T}}$. The partial derivatives are easily obtained as

$$\frac{\partial g}{\partial \alpha'} = \sum_i \frac{\hat{\lambda}_i + \hat{\lambda}_i^2}{(\hat{\lambda}_i + \hat{\lambda}_i^2/2)^2},$$

$$\frac{\partial g}{\partial \beta_j} = \sum_i \frac{\hat{\lambda}_i + \hat{\lambda}_i^2}{(\hat{\lambda}_i + \hat{\lambda}_i^2/2)^2} z_{ij},$$

where $\lambda_i(\alpha', \boldsymbol{\beta}) = 2\exp(\alpha' + \boldsymbol{\beta}^T \mathbf{z}_i)$, and $\hat{\lambda}_i = 2\exp(\hat{\alpha}' + \hat{\boldsymbol{\beta}}^T \mathbf{z}_i)$ is $\lambda_i(\alpha', \boldsymbol{\beta})$ evaluated at $(\hat{\alpha}', \hat{\boldsymbol{\beta}})^T)$. The *final variance estimate* of $\mathrm{Var}(\hat{N}_{\mathrm{GC}})$ is obtained as the sum of (7) and (8). The performance of this variance estimate is investigated in the next section.

4. Simulation

To illustrate the performance of the generalized Chao estimator four simulation experiments were conducted. In the first experiment count data were generated according to

$$Y_i \sim \mathrm{Po}(\exp(\alpha + \beta z_i)),$$

where $z_i \sim N(20, 15^2)$ and $\alpha = 0$, $\beta = 0.04$. We looked at $N = 2000$, $N = 1000$, $N = 500$, and $N = 200$. Zeros were truncated and the population size estimators of interest computed. Besides $\hat{N}_{\mathrm{GC}}$ and $\hat{N}_C$ we have also included the Turing estimator $\hat{N}_T = n/(1 - f_1/S)$ as an estimator under homogeneity for comparison. Here $S = f_1 + 2f_2 + \cdots + mf_m$. The background of the Turing estimator as discussed (for example) in Chao and Lee (1992) is based on the sample coverage estimator $1 - f_1/S$ which was first introduced by Turing (Good, 1953). In the equally likely case the sample coverage is n/N which, if equated to $1 - f_1/S$, leads to $\hat{N}_T$ which was suggested by Darroch and Ratcliffe (1980). Yet another way of approaching Turing estimation is as follows. Write N as $Np_0 + (1 - p_0)N$, the latter can be estimated by n, so that the estimator $n/(1 - p_0)$ of N arises. Under Poisson homogeneity we have that $p_0 = \exp(-\lambda) = \frac{\exp(-\lambda)\lambda}{\lambda} = p_1/E(Y)$ which can be estimated by $\frac{f_1/N}{S/N} = f_1/S$, and the Turing estimator arises once again. The results are presented in Table 1. Clearly, there is the known underestimation of Turing estimator if heterogeneity is ignored. The conventional Chao estimator shows an improved performance, although as a lower bound estimator it still shows an underestimation bias. Only the generalized Chao estimator is asymptotically unbiased. This becomes clear if we look at $E(\hat{N}_{\mathrm{GC}})/N$ which is $1.0250, 1.0086, 1.0044, 1.0026$ for $N = 200, 500, 1000, 2000$, respectively. The corresponding values for the Turing estimator are $0.9275, 0.9288, 0.9275, 0.9275$ and for the conventional Chao $0.9775, 0.9748, 0.9713, 0.9718$. Nevertheless, for small sample sizes (200 and 500) the mean squared error of the conventional Chao estimator is smaller than the one of the generalized Chao estimator. This is due to the fact that its variance is increased, a consequence of including variability through covariate information. As with any asymptotically biased estimator, with increasing size N the persisting bias of Chao and Turing becomes dominant and bounds their relative mean squared error $E(\hat{N} - N)^2/N^2$ away from zero. Hence the new estimator will have greater benefit for larger studies.

Table 1
Performance measures for the generalized Chao, Chao, and Turing estimator for four populations with true sizes $N = 200$, $N = 500$, $N = 1000$, and $N = 2000$, respectively; $Y_i|z_i \sim \mathrm{Po}(\exp(\alpha + \beta z_i))$ with $\alpha = 0$, $\beta = 0.04$, and $z_i \sim N(20, 15^2)$ (experiment 1)

Estimator	Mean	SD	Minimum	Median	Maximum
			$N = 200$		
G-Chao	205.0	19.0	174.5	201.4	398.0
Chao	195.5	10.8	170.1	194.6	243.5
Turing	185.5	6.3	167.1	185.6	206.0
			$N = 500$		
G-Chao	504.3	21.7	455.4	503.6	593.8
Chao	487.4	15.9	444.7	487.1	567.3
Turing	464.4	9.6	434.5	464.3	504.6
			$N = 1000$		
G-Chao	1004.4	30.0	933.0	1001.7	1115.6
Chao	971.3	20.4	913.8	970.9	1049.0
Turing	927.5	12.8	890.1	927.3	974.4
			$N = 2000$		
G-Chao	2005.1	42.3	1894.3	2002.7	2168.3
Chao	1943.7	31.9	1853.7	1942.2	2038.1
Turing	1854.9	20.0	1794.0	1854.8	1910.6

A second experiment uses a binary covariate indicating sampling Poisson counts from two different populations of fixed size. This experiment realizes the idea of a Poisson distribution with contaminations: $Y_i \sim \mathrm{Po}(\exp(\alpha + \beta z_i))$ where $z_i = 1$ if count Y_i is sampled from the first (uncontaminated) component with mean $\exp(\alpha) = 0.5$ for $i = 1, \ldots, p$ and $z_i = 0$ if count Y_i is sampled from a second (contaminating) component with mean $\exp(\alpha + \beta) = 3.0$ for $i = p + 1, \ldots, N$. Two values of $N = 1000; 2000$ and $p = 0.5 \times N; 0.1 \times N$ where chosen with results very similar to the first experiment. We present in Figure 1 the case with $N = 1000$ and $p = 0.5 \times N$.

We have also compared the estimated standard error based upon (7) and (8) with the estimated true standard errors in the two simulation experiments described above. To be pre-

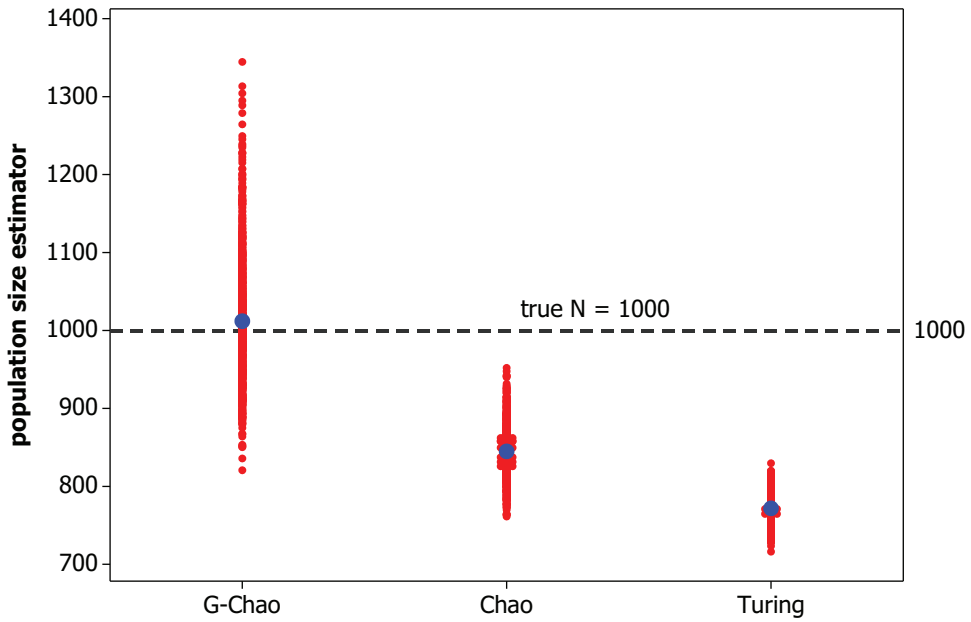

Figure 1. Individual value plot for the generalized Chao, the Chao and the Turing estimator using the contamination model $Y_i \sim \text{Po}(\exp(\alpha + \beta z_i))$ where z_i is a dummy representing membership to the first (uncontaminated) component with mean $\exp(\alpha) = 0.5$ for $i = 1, \ldots, p$ or second (contaminating) component with mean $\exp(\alpha + \beta) = 3.0$ for $i = p + 1, \ldots, N$; the true population size is $N = 1000$ and $p = 500$; the bullet indicates the mean of the distribution of the simulated estimates; replication size is 1000

cise, if $\hat{N}^r_{\text{GC}}$ is the estimate in the rth simulation run, then the estimated true variance is given as $\frac{1}{R}\sum_r(\hat{N}^r_{\text{GC}} - \bar{N}_{\text{GC}})^2$ and the estimated true standard error is its root. This is compared with the estimated variance according to (7) and (8), averaged over the R replications. The results are provided in Table 2. The ratio in the last column of Table 2 provides a comprehensive performance measure for the variance estimator. The target value of this ratio is 1 and approximations become quite acceptable for sizes of 500 and above. Note that for smaller population sizes the approximation appears to be conservative.

Table 2

Comparison of estimated standard error (Estimated SE) and true standard error (True SE) of the generalized Chao estimator for five populations with true sizes $N = 200$, $N = 500$, $N = 1000$, $N = 2000$, and $N = 5000$, respectively; ratio refers to the ratio of Estimated SE to True SE

N	True SE	Estimated SE	Ratio
Experiment 1			
200	19.765	24.266	1.228
500	23.210	24.787	1.068
1000	30.644	32.139	1.049
2000	43.492	44.283	1.018
5000	67.140	68.900	1.026
Experiment 2			
200	51.585	59.220	1.148
500	60.110	60.929	1.014
1000	76.354	78.044	1.022
2000	106.034	105.951	0.999
5000	165.092	164.629	0.997

In a third experiment with $N = 1000$, the contamination model $Y_i \sim \text{Po}(\exp(\alpha + \beta z_i))$ was modified as follows. Count Y_i was generated from a Poisson with mean $\exp(\alpha + \beta z_i)$, where z_i is a dummy representing membership of the first component with mean $\exp(\alpha) = 0.5$ for $i = 1, \ldots, 450$ or of the second component with mean $\exp(\alpha + \beta) = 3.0$ for $i = 451, \ldots, 1000$, respectively; however, for the fitting the model it is assumed that $\exp(\alpha) = 0.5$ for $i = 1, \ldots, 550$ and $\exp(\alpha + \beta) = 3.0$ for $i = 551, \ldots, 1000$ so that 100 observations are misclassified (first component has additional, unobserved heterogeneity). In this case, the generalized Chao estimator becomes downwards biased, although the bias of the conventional Chao estimator is clearly larger (Figure 2). Hence, in these situations with additional heterogeneity it can be expected that the generalized Chao estimator is experiencing slight to moderate bias.

Finally, we have compared the generalized Chao estimator with existing modeling work based upon the zero-truncated Poisson (ZTP) model. Population size estimation with covariate information has been previously considered by van der Heijden et al. (2003a,b). Their approach assumes the validity of a Poisson regression model with only zero counts being truncated. The estimated population size is here

$$\hat{N} = \sum_{i=1}^{n} \frac{1}{1 - \exp[-\exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \mathbf{z}_i)]},$$

where $\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \mathbf{z}_i$ is the fitted linear predictor with covariate vector $\mathbf{z}_i$ under the ZTP model. We expect—if the Poisson

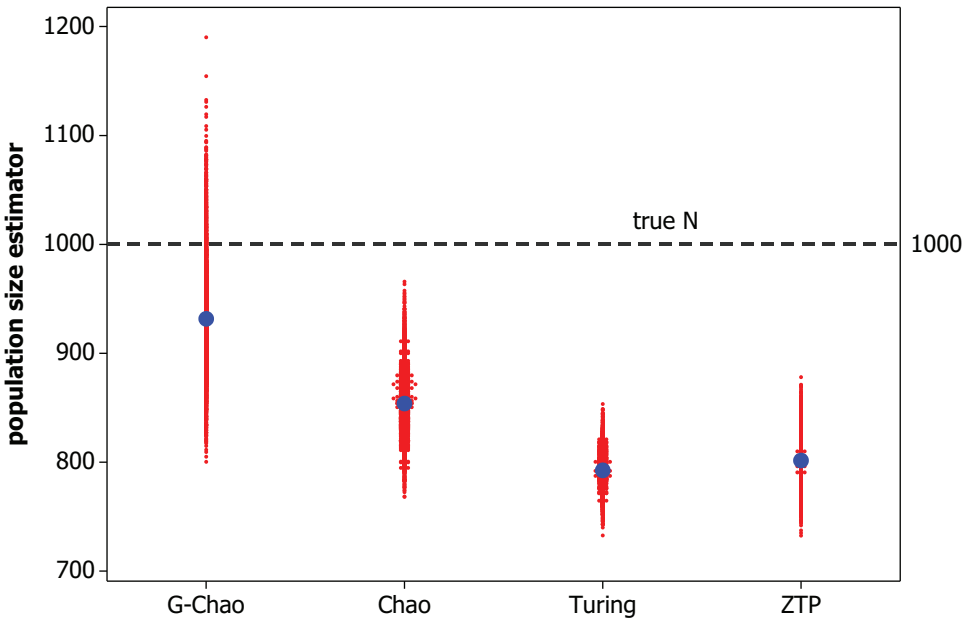

Figure 2. Individual value plot for the generalized Chao, the Chao, the Turing estimator and the ZTP-model based estimator using the contamination model $Y_i \sim \text{Po}(\exp(\alpha + \beta z_i))$ where z_i is a dummy representing membership to the first component with mean $\exp(\alpha) = 0.5$ for $i = 1, \ldots, 450$ or second component with mean $\exp(\alpha + \beta) = 3.0$ for $i = 451, \ldots, 1000$; however, for the fitting it is assumed that $\exp(\alpha) = 0.5$ for $i = 1, \ldots, 550$ and $\exp(\alpha + \beta) = 3.0$ for $i = 551, \ldots, 1000$ so that 100 observations are misclassified (first component has additional, unobserved heterogeneity); the bullet indicates the mean of the distribution of the simulated estimates; replication size is 1000

model holds—that the population size estimate based upon the zero-truncated Poisson model is asymptotically unbiased and efficient. Indeed, Table 3 shows that the generalized Chao estimator as well as the population size estimator based upon the ZTP model are asymptotically unbiased. However, the ZTP-based estimator provides the smaller standard error. It appears that the generalized Chao estimator has 1.5-times higher standard error than the ZTP-based estimator. This finding might lead to a strategy which first investigates the ZTP model for validity and if the validity check fails proceeds with the generalized Chao estimator. Indeed, if the ZTP model fails to hold or if the considered covariates account only for part of the heterogeneity, the population size is likely to be underestimated. This is expressed in the bottom part of Table 3 where now both estimators experience underestimation bias, as expected, but the generalized Chao estimator is experiencing less bias in comparison to the ZTP-based estimator. This is also quite apparent in Figure 2 where the display is supplemented by the ZTP-model based population size estimator. Here the ZTP-model based estimators can only slightly adjust for the unobserved heterogeneity. Note also the higher value of the ratio in the last column of Table 3, indicating a deflated standard error of the ZTP-model based population size estimator.

Table 3

Comparison of the generalized Chao estimator and the population size estimator based upon the zero-truncated Poisson (ZTP) regression model; ratio refers to the ratio of the estimated SE of the generalized Chao estimator and the estimated SE of the population size estimator under the ZTP

N	G-Chao (SE)	ZTP (SE)	Ratio
Experiment 1			
200	205.08 (20.36)	200.61 (8.53)	2.39
500	503.79 (23.58)	500.29 (13.44)	1.75
1000	1005.01 (30.80)	999.93 (19.00)	1.62
2000	2004.19 (42.24)	2000.31 (26.96)	1.57
5000	5003.76 (66.46)	5001.14 (42.48)	1.56
Experiment 3			
1000	930.48 (48.10)	801.42 (19.81)	2.43

More recently Cruyff and van der Heijden (2008) suggested utilizing the truncated negative-binomial model for population size estimation. Adapting the notation in Cruyff and van der Heijden (2008) it is assumed that $Y_i|z_i \sim NB(\mu_i, \theta)$ with $\mu_i = \exp(\alpha + \boldsymbol{\beta}^T \mathbf{z}_i)$, where NB stands for the negative-binomial distribution. Zero-counts are truncated as before and parameter estimates are found from the zero-truncated negative-binomial (ZNB) likelihood. We have also compared the generalized Chao estimator with this approach. Results are provided in Table 4. Whereas the ZNB-based population size estimator is unbiased, the ZTP-based estimator underestimates considerably. The generalized Chao estimator is moderately underestimating as well, but has reasonable standard errors in comparison to the ZNB-based estimator. The problem with the ZNB approach appears to be that the fitting (with any maximum likelihood algorithm) only works when the count data follow the negative-binomial model. Outside this class severe convergence problems due to boundary con-

Table 4
Comparison of the generalized Chao estimator with the population size estimator based upon the zero-truncated Poisson (ZTP) and the zero-truncated negative-binomial (ZNB) regression model for $N = 1000$; in experiment 4 simulated counts are $Y_i|z_i \sim NB(\mu_i, \theta)$ with $\mu_i = e^{0.02z_i}$, $z_i \sim N(8, 5)$, and $\theta = 3$; experiment 3 is as described previously

N	G-Chao (SE)	ZTP (SE)	ZNB (SE)
Experiment 4			
200	187.02 (22.29)	167.96 (12.47)	208.80 (46.45)
500	459.40 (31.80)	417.73 (19.22)	504.98 (58.02)
1000	913.06 (43.84)	833.78 (27.09)	1005.80 (78.72)
2000	1819.01 (60.35)	1664.40 (37.55)	2003.66 (107.31)
5000	4542.05 (94.85)	4159.92 (58.87)	5005.08 (168.75)
Experiment 3			
1000	930.48 (48.10)	801.42 (19.81)	829.09 (23.92)

ditions can occur which does not really allow complete comparisons for the simulation experiments done in section 4.1. However, fitting of the negative-binomial was possible in experiment 3 and the results are provided in the bottom part of Table 4. Here we see that the ZNB-based estimator experiences strong underestimation bias, whereas the generalized Chao estimator remains with its less severe underestimation bias. Note also in Table 4 that in all cases the generalized Chao estimator has the smallest mean squared error among all three model-based estimators considered.

5. Case Study on Completeness of Carcass Submissions from Farms in Great Britain

Farm animal submissions in England and Wales are made to Animal Health and Veterinary Laboratories Agency (AHVLA) regional laboratories from private veterinary surgeons (PVS) wanting a post-mortem on a carcass for which the reason for death cannot be determined, or a sample from an animal requiring further diagnostic tests. Unless there is a notifiable disease suspected, it is up to the PVS to decide whether to submit a sample. The cost of tests and post-mortems is subsidized by the Department of Food and Rural Affairs (DEFRA) to encourage submissions so that any new or emerging disease threats will be identified. Clearly, since not every farm submits samples to AHVLA, it is of great interest to determine the completeness of farm submissions. For cattle, in the period 1 January–31 December 2009, there were an estimated 60,571 farms of which 48,535 had no submissions. The frequency f_y of farms with exactly y submissions (carcass submissions) to AHVLA regional laboratories is 48,535 (58,713), 6,340 (1,532), 2,520 (231) , 1,149 (51), 709 (27), 380 (6), 249 (5), 173 (2), 135 (1), 94 (3), 80 (0), 207 (0), for $y = 0, 1, \ldots, 11$ where f_{11} refers to the frequency of exactly 11 submissions or more. Note that there are $f_0 = 48{,}535$ farms with no submissions at all (58,713 with no carcass submissions). It is known that many farms do not necessarily make submissions to AHVLA, even where they may have unknown diseases in their farm, and it is of interest to estimate the frequency of farms that made no submissions but had unknown disease on their farm. In other words, there are $n = 12{,}036$ farms with submission of samples to AHVLA and we are interested in estimating $N = n + \tilde{f}_0$, the total number of farms with unknown disease, namely those which submit and those which do not.

Logistic regression has previously been applied to the holdings that have submitted samples to AHVLA and a dependence has been found between holding size, holding type and the distance from an AHVLA regional laboratory on the likelihood of a submission (only done on the total submissions not on any particular disease or syndrome). Distance is likely to be particularly important for carcass submissions since the farmer is responsible for delivering the carcass to the regional laboratory and so the cost and time involved may influence the decision of whether to submit the carcass. Note that in the previous analysis the dependent variable is the decision to submit or not to submit. Here the dependent variable is the (truncated) number of submissions, given that there is a submission at all.

In Table 5 a completeness analysis is provided for two count variables. One is the total number of submissions including the carcass itself but also other material such as blood samples. The other count variable is the number of submitted carcasses. It can be seen that in both cases the size of the farm (on log-scale) is an important covariate as is the type of farm (dairy or not). The distance of the farm to the regional laboratory appears to provide important covariate information only for the count of submitted carcasses. The generalized Chao estimator is considerably higher in both situations in comparison to the conventional Chao estimator. For comparison we have included the Turing estimator and the population size estimator based upon the zero-truncated Poisson regression model as well. The latter two underestimate strongly with the ZTP-based estimate being able to adjust for some of the observed heterogeneity. Note that the completeness of surveillance for the total number of submissions (12,036/21,657×100%=56%) is more than double than for the submission of carcasses only (18,58/7,688×100%=24%) which is in line with expectation since there is a much greater distance dependent cost for carcass submissions compared with other types of sample.

6. Discussion

Chao's estimator is widely accepted in the community because of its robustness with respect to misspecifications of the Poisson model. Nevertheless, if heterogeneity is present, Chao's estimator is only a lower bound. This lower bound provides a less biased estimator in comparison to estimators under homogeneity such as the Turing estimator, but it is still biased and this bias persists asymptotically. Hence, bias reduction methods such as suggested here by means of utilizing covariate information are useful. A similar approach was suggested in Böhning and van der Heijden (2009) for the estimator of Zelterman (1988). However, the estimator of Zelterman can experience extreme forms of overestimation and has less favorable properties if compared to Chao's estimator (see also Böhning, 2010). However, whereas it is relatively easy to generalize Zelterman's estimator to covariate information, it was for a considerable time unclear how this could be accomplished for Chao's estimator. This article has now filled this gap.

Table 5
Chao, generalized Chao, Turing, and ZTP estimator for total number of submissions and number of Carcass submissions to the farm file

Logistic regression analysis total number of submissions				
Covariate	coef	SE-coef.	Z	*p*-Value
Log-size	0.33	0.03	12.5	0.00
Type (1=dairy 0=beef)	0.29	0.05	5.55	0.00
Log-distance	−0.01	0.04	−0.10	0.92
Estimated farms with carcasses (95% CI): [based on total number of submissions]				
n	G-Chao	Chao	Turing	ZTP
12,036	21,657 (20,885, 22,429)	20,011 (19,993, 20,029)	15,532 (15,348, 15,716)	18,346 (17,932, 18,760)
Logistic regression analysis number of carcass submissions				
Covariate	Coef.	SE-coef.	Z	*p*-Value
Log-size	0.32	0.08	4.10	0.00
Type (1=dairy, 0=beef)	0.05	0.16	0.38	0.71
Log-distance	−0.15	0.09	−1.66	0.09
Estimated farms with carcasses (95% CI): [based on number of carcass submissions only]				
n	G-Chao	Chao	Turing	ZTP
1858	7688 (6523, 8853)	6938 (6868, 7009)	5279 (4645, 5913)	6008 (5293, 6723)

Population size estimation with covariate information has been previously considered by van der Heijden et al. (2003a,b) using Poisson regression model with only zero counts being truncated, as was mentioned above. If the model fails to hold or if the considered covariates account only for part of the heterogeneity, the population size is likely to be underestimated. This has been illustrated in the simulation study but also the case study on completeness of farm submissions has provided evidence that the ZTP-model based estimate is too low.

Incorporating covariate information is an important step in reducing bias in population size estimation of elusive target populations using truncated count distributions of repeated identifications. The suggested extension of the Chao estimator for covariate modeling appears to allow bias reduction of the conventional Chao lower bound estimator to the extent that it becomes unbiased if all heterogeneity is included as covariate information. It was also shown for the situation of stratified heterogeneity (Theorem 3) that the generalized Chao estimator is always at least as large as the conventional Chao estimator. This is correct whether the observed heterogeneity captures all of the heterogeneity or not. It also seems to compare favorable to alternative approaches including the negative-binomial model.

The question arises in which way the proposed approach generalizes to other capture–recapture situations. For example, if the number of sampling occasions T is known it might be more reasonable to model X_i with a binomial distribution and include covariate information by means of a generalized linear model using the logit-link function. However, the associated binomially truncated binomial likelihood with involved logit-link does not lead to a simple or well-known likelihood (though in principle this could be handled numerically as well). It is a rather unique feature (and the main result of this article) that the binomially truncated Poisson likelihood with involved log-link becomes identical (up to the intercept) to a binomial (Bernoulli) likelihood with involved logit-link. Of course, the posed question becomes more burning when the number of sampling occasions T_i varies with the unit i. A practical way to handle this could be to include $\log T_i$ as an offset in the Poisson regression model, or ultimately, in the logistic regression model.

7. Supplementary Materials

A Web Appendix, which contains the data used in Section 5, the associated R-code for producing the results in this section as well as the R-code for running the simulations in Section 4, are available with this paper at the Biometrics website on the Wiley Online Library.

Acknowledgements

The authors are grateful to the Department of Environment, Food and Rural Affairs (DEFRA) in the UK for contributing to the funding of this work in project SE4302. The authors would also thank two anonymous referees, the associate editor and the editor for very useful comments that improved the presentation of the paper.

References

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

Böhning, D. (2010). Some general comparative points on Chao's and Zelterman's estimator of population size. *Scandinavian Journal of Statistics* **37**, 221–236.

Böhning, D. and Del Rio Vilas, V. J. (2008). Estimating the hidden number of scrapie affected holdings in Great Britain using a simple, truncated count model allowing for heterogeneity. *Journal of Agricultural, Biological and Environmental Statistics* **13**, 1–22.

Böhning, D. and Kuhnert, R. (2006). The equivalence of truncated count mixture distributions and mixtures of truncated count distributions. *Biometrics* **62**, 1207–1215.

Böhning, D. and van der Heijden, P. (2009). A covariate adjustment for zero-truncated approaches to estimating the size of hidden and elusive populations. *Annals of Applied Statistics* **3**, **595**–610.

Bunge, J. and Fitzpatrick, M. (1993). Estimating the number of species: A review. *Journal of the American Statistical Association* **88**, 364–373.

Chao, A. (1987). Estimating the population size for capture–recapture data with unequal catchability. *Biometrics* **43**, 783–791.

Chao, A. (1989). Estimating population size for sparse data in capture–recapture experiments. *Biometrics* **45**, 427–438.

Chao, A. and Lee, S.-M. (1992). Estimating the number of classes via sample coverage. *Journal of the American Statistical Association* **87**, 210–217.

Chao A., Tsay P. K., Lin S. H, Shau W. Y, and Chao D. Y. (2001). Tutorial in Biostatistics: The applications of capture–recapture models to epidemiological data. *Statistics in Medicine* **20**, 3123–3157.

Cruyff, M. J. L. and Van der Heijden, P. G. M. (2008). Point and interval estimation of the population size using a zero-truncated negative binomial regression model. *Biometrical Journal* **50**, 1035–1050.

Darroch, J. N., and Ratcliff, D. (1980). A note on capture–recapture estimation. *Biometrics* **36**, 149–153.

Dorazio, R. M. and Royle, J. A. (2005). Mixture models for estimating the size of a closed population when capture rates vary among individuals. *Biometrics* **59**, 351–364.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika* **40**, 237–264.

Hay, G. and Smit, F. (2003). Estimating the number of drug injectors from needle exchange data. *Addiction Research and Theory* **11**, 235–243.

Holzmann, H., Munk, A., and Zucchini, W. (2006). On identifiability in capture–recapture models. *Biometrics* **62**, 934–939.

Link, W. A. (2003). Nonidentifiability of population size from capture–recapture data with heterogeneous detection probabilities. *Biometrics* **59**, 1123–1130.

Link, W. A. (2006). Response to a paper by Holzmann, Munk and Zucchini. *Biometrics* **62**, 936–939.

Mao, C. X. (2008). Lower bounds to the population size when capture probabilities vary over individuals. *Australian & New Zealand Journal of Statistics* **50**, 125–134.

McKendrick, A. G. (1926). Application of Mathematics to Medical Problems. *Proceedings of the Edinburgh Mathematical Society* **44**, 98–130.

Mosley, W. H., Bart, K. J., and Sommer, A. (1972). An epidemiological assessment of cholera control programs in rural East Pakistan. *International Journal of Epidemiology* **1**, 5–11.

Pledger, S. A. (2005). The performance of mixture models in heterogeneous closed population capture–recapture. *Biometrics* **61**, 868–876.

Roberts, J. M. and Brewer, D. D. (2006). Estimating the prevalence of male clients of prostitute women in Vancouver with a simple capture–recapture method. *Journal of the Royal Statistical Society, Series A* **169**, 745–756.

Van der Heijden, P. G. M., Cruyff, M., and van Houwelingen, H. C. (2003a). Estimating the size of a criminal population from police records using the truncated Poisson regression model. *Statistica Neerlandica* **57**, 1–16.

Van der Heijden, P. G. M., Bustami, R., Cruyff, M., Engbersen, G., and van Houwelingen, H. C. (2003b). Point and interval estimation of the population size using the truncated Poisson regression model. *Statistical Modelling* **3**, 305–322.

Van Hest, N. A. H., De Vries, G., Smit, F., Grant, A. D., and Richardus, J. H. (2008). Estimating the coverage of Tuberculosis screening among drug users and homeless persons with truncated models. *Epidemiology and Infection* **136**, 14–22.

Wang, J.-P. and Lindsay, B. G. (2005). A penalized nonparametric maximum likelihood approach to species richness estimation. *Journal of the American Statistical Association* **100**, 942–959.

Wang, J.-P. and Lindsay, B. G. (2008). An exponential partial prior for improving nonparametric maximum likelihood estimation in mixture models. *Statistical Methodology* **5**, 30–45.

Wilson, R. M. and Collins, M. F. (1992). Capture–recapture estimation with samples of size one using frequency data. *Biometrika* **79**, 543–553.

Zelterman, D. (1988). Robust estimation in truncated discrete distributions with applications to capture–recapture experiments. *Journal of Statistical Planning and Inference* **18**, 225–237.

Received November 2011. *Revised June* 2013.
Accepted June 2013.