Advances and Applications in Statistics
Volume 34, Number 1, 2013, Pages 65-84

Published by Pushpa Publishing House, Allahabad, INDIA

A LATENT CLASS APPROACH WITH COVARIATES AND LOCAL DEPENDENCE IN CAPTURE-RECAPTURE MODELS

Joanne Thandrayen* **and Yan Wang**

School of Mathematical and Physical Sciences
University of Newcastle
Callaghan, NSW 2308, Australia
e-mail: joanne.thandrayen@newcastle.edu.au

School of Mathematical and Geospatial Sciences
RMIT University
Melbourne, VIC 3000, Australia

Abstract

Traditional capture-recapture methods assume that lists operate independently (local independence) and that capture probabilities are homogeneous. In studies involving human populations, these assumptions are often violated. This paper presents an approach where dependence between the lists and the effects due to the observable covariates are modelled directly in the capture probability. For this purpose, we employ a multinomial latent class model. Estimation of the model parameters is based on the maximum likelihood method via the EM algorithm. An approximation for the variance of the unknown population size is also formulated.

2010 Mathematics Subject Classification: 62F99.
Keywords and phrases: conditional likelihood, covariate, EM algorithm, heterogeneity, latent class model, multinomial logit.

*Corresponding author

Communicated by K. K. Azad
Received December 14, 2012

1. Introduction

Capture-recapture models have long been applied to estimate the population size of wild animals (see Schwarz and Seber [22], Pollock [21] for details). In practice, it is unlikely that an entire animal population will be surveyed, thus these models are used to determine an estimate of the population size. In a wildlife context, capture-recapture methods utilize the information available from the animals captured on a number of occasions. These methods are now widely used to count human populations in various settings such as in epidemiology (LaPorte et al. [15], Verlato and Muggeo [26]) and in social sciences (Davies et al. [9], Gurgel et al. [11]). In situations involving human populations, unlike animal studies, an estimate of the population size is obtained by using a number of existing lists. Each list can detect only a part of the population of interest, thus some individuals are unobserved and this is the parameter that we wish to estimate.

List dependence and heterogeneity are features of most capture-recapture data relating to human populations. If the effects of these features are not accounted for in the modelling, then a biased estimate of the population size will result (Wittes et al. [30], Hook and Regal [12], Chao et al. [7]). In the literature of capture-recapture analysis, various models are available which focus on controlling the effects of heterogeneity. When the capture probabilities are heterogeneous, modelling them in terms of observable covariates provides a way to account for observed heterogeneity (Huggins [13], Alho [1]). Zwane and van der Heijden [31] developed a model which simultaneously accounts for list dependence and observed heterogeneity by means of a multinomial logit model. In the absence of influential covariates latent class modelling (Lazarsfeld and Henry [16], McCutcheon [19]), which assumes that the whole population consists of a number of sub-populations or classes where the class membership for each individual is unknown, can be employed to control unobserved heterogeneity (Link [17], Böhning et al. [3]). New mixture models within the context of capture-recapture focus on estimation via the nonparametric maximum likelihood approach (Kuhnert et al. [14], Mao [18]); further methods are compiled in Böhning [5]. Recently, a few researchers have attempted to develop models which can simultaneously account for list dependence and heterogeneity (both observed and unobserved). Cruyff and van der Heijden [8] introduced a zero-truncated negative binomial regression model which explained both the observed and unobserved heterogeneity. Stanghellini and van der Heijden [23] and Bartolucci and Forcina [2] presented a hierarchical log-linear model and a class of latent marginal regression models, respectively, where the effects of list dependence, observed and unobserved heterogeneity are simultaneously modelled.

In this paper, we extend the method of Zwane and van der Heijden [31] by including the effects of unobserved heterogeneity. We use a multinomial latent class model in which the effects of the list interaction and the covariates directly affect the capture probability. In contrast, in our previous work (Thandrayen and Wang [25]), we treated the covariates as affecting the latent class probabilities. In practice, both ways in which the covariates can be incorporated in the modelling are important because they provide useful information about the population under study. Our proposed method can accommodate covariates which are both categorical and continuous in nature, unlike the model of Stanghellini and van der Heijden [23] which incorporates only categorical covariates. Our method also employs an approximation to the information matrix (McLachlan and Peel [20]) when evaluating the standard errors, which renders our model computationally less intensive than those of Stanghellini and van der Heijden [23] and Bartolucci and Forcina [2].

The structure of this paper is as follows: in Section 2, we formulate our proposed model and estimate it by the maximum likelihood method via the EM algorithm. In the present context, the available covariates affect only the capture probability. In Section 3, the mathematical formulae for the standard errors are derived. In Section 4, simulation studies are performed while in Section 5, the proposed method is illustrated by means of a study on the prevalence of diabetes in northern Italy. In Section 6, we draw our conclusions.

2. Method

Assume that there are *m* lists in the multiple list problem. Then there are $K = 2^m$ possible capture histories for each individual. Let $n_{ik} = 1$ if individual *i* has capture history k $(k = 1, ..., K)$ and 0 otherwise. We also assume that there exist influential covariates which may affect the capture probabilities of individuals. The information for these covariates is collected into a matrix *X* with dimension $n \times H,$ where *n* denotes the unknown population size and *H* represents the number of available covariates. The elements of ***X*** are denoted as $x_{ih}, i = 1, ..., n, h = 1, ..., H.$ List effects are collected into a design matrix ***Y*** with dimension $J \times K,$ where *J* represents the number of list effects. The elements of ***Y*** are $y_{jk},$ $j = 1, ..., J, k = 1, ..., K.$ The form of the design matrix depends on the existing interactions among the lists (Wang and Thandrayen [28]).

Let $f_i(k, \mathbf{\Lambda})$ denote a multinomial probability mass function for the capture history *k* $(k = 1, ..., K)$ of an individual i $(i = 1, ..., n),$ where $\mathbf{\Lambda}$ is a matrix of parameters associated with the covariate and list effects. The elements of $\mathbf{\Lambda}$ are $\lambda_{hj}, h = 1, ..., H, j = 1, ..., J.$ Thus, $f_i(k, \mathbf{\Lambda})$ is of the form (Zwane and van der Heijden [31])

$$f_i(k, \mathbf{\Lambda}) = \frac{\exp\left(\sum_{h=1}^{H}\sum_{j=1}^{J} x_{ih}\lambda_{hj}y_{jk}\right)}{\sum_{k=1}^{K}\exp\left(\sum_{h=1}^{H}\sum_{j=1}^{J} x_{ih}\lambda_{hj}y_{jk}\right)}, \; i = 1, ..., n, \; k = 1, ..., K.$$

In particular, $f_i(K, \mathbf{\Lambda})$ refers to the case where an individual *i* with capture history *K* is not captured by any list and

$$f_i(K, \mathbf{\Lambda}) = \frac{1}{\sum_{k=1}^{K}\exp\left(\sum_{h=1}^{H}\sum_{j=1}^{J} x_{ih}\lambda_{hj}y_{jk}\right)}.$$

Assume that the proposed model is a finite mixture model whose marginal form is

$$f_i(k) = \sum_{l=1}^{L} q_l \, f_i(k, \, \mathbf{\Lambda}^l),$$

where $0 \leq q_l \leq 1, \; l = 1, \ldots, L,$ are the mixing proportions with $\sum_{l=1}^{L} q_l = 1$; $\mathbf{\Lambda}^l$ is a matrix of parameters associated with q_l, and

$$f_i(k, \, \mathbf{\Lambda}^l) = \frac{\exp\left(\sum_{h=1}^{H}\sum_{j=1}^{J} x_{ih}\lambda_{hj}^{l} y_{jk}\right)}{\sum_{k=1}^{K}\exp\left(\sum_{h=1}^{H}\sum_{j=1}^{J} x_{ih}\lambda_{hj}^{l} y_{jk}\right)}.$$

We also assume that $\mathbf{\Lambda}^L = 0$ to ensure that the parameters are identifiable (Vermunt [27]).

Assume that a sample of n_{obs} individuals has been identified by the lists. The number n_K of uncaptured individuals is not known and we wish to estimate it. The Horvitz-Thompson estimator for the population size is $\sum_{i=1}^{n_{obs}} 1/\pi_i$, where $\pi_i = 1 - f_i(K)$, and

$$f_i(K) = \sum_{l=1}^{L} \frac{q_l}{\sum_{k=1}^{K}\exp\left(\sum_{h=1}^{H}\sum_{j=1}^{J} x_{ih}\lambda_{hj}^{l} y_{jk}\right)}.$$

Since the model parameters are unknown, we estimate *n* by

$$\hat{n} = \sum_{i=1}^{n_{obs}} \frac{1}{\hat{\pi}_i}. \tag{2.1}$$

An estimate of the unobserved number of individuals is thus given as $\hat{n}_K = \hat{n} - n_{obs}$. In the literature, various estimation methods are available for finite mixtures; the most popular method is the maximum likelihood method via the EM algorithm (Dempster et al. [10]) due to its interesting properties: it shows monotone convergence and provides reasonable estimates if the starting values are appropriate. However, the main drawbacks for the EM algorithm are: it converges slowly and the optimal solution depends on both the initial values used and the stopping rule employed to detect whether the maximum has been reached. We employ the conditional likelihood function as it accommodates the fact that the covariate information for the unobserved individuals is not available (Thandrayen and Wang [24]). To proceed with the EM algorithm, we assume the unobserved variable z_{ik}^l is 1 if individual *i* with capture history *k* belongs to class *l*; otherwise z_{ik}^l is 0 (Böhning and Kuhnert [4]). Wang and Thandrayen [28] provided the details of how to construct the conditional likelihood function, which is given by

$$L_c(\boldsymbol{\Lambda}, \boldsymbol{Q}) = \prod_{i=1}^{n_{obs}} \prod_{k=1}^{K-1} \left[\frac{\prod_{l=1}^{L} \{q_l f_i(k, \boldsymbol{\Lambda}^l)\}^{z_{ik}^l}}{\pi_i} \right]^{n_{ik}},$$

where $\boldsymbol{Q} = (q_1, ..., q_L)'$.

Let

$$D_1 = \sum_{k=1}^{K} \exp\left(\sum_{h=1}^{H} \sum_{j=1}^{J} x_{ih} \lambda_{hj}^l y_{jk} \right).$$

The log-likelihood of L_c is

$$l_c = \sum_{i=1}^{n_{obs}} \sum_{k=1}^{K-1} \sum_{l=1}^{L} n_{ik} z_{ik}^l \left(\sum_{h=1}^{H} \sum_{j=1}^{J} x_{ih} \lambda_{hj}^l y_{jk} - \log D_1 \right)$$

$$+ \sum_{i=1}^{n_{obs}} \sum_{k=1}^{K-1} \sum_{l=1}^{L} n_{ik} z_{ik}^l \log q_l - \sum_{i=1}^{n_{obs}} \sum_{k=1}^{K-1} n_{ik} \log \pi_i. \tag{2.2}$$

In the *E*-step, the algorithm substitutes z_{ik}^l by their expected values e_{ik}^l based on the observed data and current values of $\boldsymbol{\Lambda}$ and $\boldsymbol{Q}$ such that

$$e_{ik}^l = \frac{q_l f_i(k, \boldsymbol{\Lambda}^l)}{\sum_{l=1}^{L} q_l f_i(k, \boldsymbol{\Lambda}^l)}$$

(Böhning and Kuhnert [4]).

In the *M*-step, the expected version of the conditional log-likelihood (2.2) is maximized to generate new values $\hat{\mathbf{\Lambda}}$ and $\hat{\boldsymbol{Q}}$. This maximization procedure does not give a closed form solution for $\hat{\mathbf{\Lambda}}$ and $\hat{\boldsymbol{Q}}$. Instead it provides an iterative solution, which needs to be updated until convergence. Once the optimal values of $\mathbf{\Lambda}$ and $\boldsymbol{Q}$ are obtained, (2.1) is used to obtain an estimate of the population size.

3. Standard Error Approximation

The unknown parameter is $\boldsymbol{\theta} = (\lambda_{hj}^{l}, q_l)'$, $h = 1, ..., H$, $j = 1, ..., J$, $l = 1, ..., L$. Since the dimension of $\boldsymbol{\theta}$ is large, standard errors based on the observed information matrix are difficult to obtain. To evaluate standard errors, we apply the technique of McLachlan and Peel [20, pp. 64-66], which involves approximating the observed information matrix. Let $\hat{\boldsymbol{\theta}}$ denote the MLE of $\boldsymbol{\theta}$. It is convenient to approximate the observed information matrix $I(\hat{\boldsymbol{\theta}})$ by the so-called empirical observed information matrix $I_e(\hat{\boldsymbol{\theta}})$.

To proceed with the empirical observed information matrix, we need to decompose the conditional log-likelihood (2.2). Under the assumption that data are independently and identically distributed, this decomposition allows (2.2) to be formulated as $l_c(\boldsymbol{\theta}) = \sum_{i=1}^{n_{obs}} l_{c|i}(\boldsymbol{\theta})$, where $l_{c|i}(\boldsymbol{\theta})$ is the likelihood function for $\boldsymbol{\theta}$ formed in respect of each individual *i*. The corresponding mathematical formula for $l_{c|i}(\boldsymbol{\theta})$ is

$$l_{c|i}(\boldsymbol{\theta}) = \sum_{k=1}^{K-1}\sum_{l=1}^{L} n_{ik} z_{ik}^{l}\left(\sum_{h=1}^{H}\sum_{j=1}^{J} x_{ih}\lambda_{hj}^{l} y_{jk} - \log D_1\right)$$

$$+ \sum_{k=1}^{K-1}\sum_{l=1}^{L} n_{ik} z_{ik}^{l} \log q_l - \sum_{k=1}^{K-1} n_{ik} \log \pi_i. \tag{3.1}$$

In a similar manner, the score vector $s(\boldsymbol{\theta})$ can be formulated as $s(\boldsymbol{\theta}) = \sum_{i=1}^{n_{obs}} s(i, \boldsymbol{\theta})$, where $s(i, \boldsymbol{\theta}) = \partial l_{c|i}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$. The mathematical equations for the evaluation of the score vector are obtained by differentiating (3.1) with respect to λ_{hj}^{l} and q_l.

Let

$$D_2 = \sum_{k=1}^{K-1} \exp\left(\sum_{h=1}^{H}\sum_{j=1}^{J} x_{ih}\lambda_{hj}^{l} y_{jk}\right) y_{jk}.$$

The derivative of (3.1) with respect to λ_{hj}^{l} is

$$\frac{\partial l_{c|i}}{\partial \lambda_{hj}^{l}} = x_{ih}\sum_{k=1}^{K-1} n_{ik} z_{ik}^{l} y_{jk} - \frac{x_{ih}D_2}{D_1}\sum_{k=1}^{K-1} n_{ik} z_{ik}^{l} - \frac{1}{\pi_i} \times \frac{x_{ih}q_l D_2}{D_1^2}\sum_{k=1}^{K-1} n_{ik},$$

while that with respect to q_l is

$$\frac{\partial l_{c|i}}{\partial q_l} = \frac{1}{q_l}\sum_{k=1}^{K-1} n_{ik} z_{ik}^{l} + \frac{1}{\pi_i} \times \frac{1}{D_1}\sum_{k=1}^{K-1} n_{ik}.$$

The observed empirical information matrix is then evaluated by the relationship $I_e(\hat{\boldsymbol{\theta}}) = \sum_{i=1}^{n_{obs}} s(i, \hat{\boldsymbol{\theta}}) s'(i, \hat{\boldsymbol{\theta}})$. Under asymptotic conditions, the covariance matrix of $\hat{\boldsymbol{\theta}}$ can be approximated by the inverse of the empirical observed information matrix, that is, $cov(\hat{\boldsymbol{\theta}}) \approx I_e^{-1}(\hat{\boldsymbol{\theta}})$. The square root of the diagonal elements of $cov(\hat{\boldsymbol{\theta}})$ leads to the standard error of $\hat{\boldsymbol{\theta}}$.

We then employ the method of conditioning to derive a variance estimator of the population size and express (2.1) as

$$\hat{n} = \sum_{i=1}^{n_{obs}} \frac{1}{\hat{\pi}_i} = \sum_{i=1}^{n} \frac{\delta_i}{\hat{\pi}_i}, \tag{3.2}$$

where δ_i is an indicator which equals 1 (individual is observed in the sample) and 0 (individual is not observed in the sample). When we condition on the observed sample size, then

$$Var(\hat{n}) = E\{Var(\hat{n} \mid n_{obs})\} + Var\{E(\hat{n} \mid n_{obs})\}. \tag{3.3}$$

The first term of (3.3) is a measure of the variability in the multinomial distribution conditional on being observed whereas the second term measures the fluctuation in the observed sample size (Zwane and van der Heijden [31]). We follow the procedure in Wang and Yip [29] to evaluate the first and second terms of (3.3). The variance estimate of the population size is given by $\hat{s}^2 = \hat{K}'\hat{\mathbf{\Sigma}}\hat{K} + \sum_{i=1}^{n} \frac{\delta_i(1-\hat{\pi}_i)}{\hat{\pi}_i^2}$ (see details in Wang and Thandrayen [28]), where

$$\hat{K} = -\sum_{i=1}^{N} \delta_i \times \frac{1}{\hat{\pi}_i^2} \times \frac{\partial \pi_i}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}.$$

The evaluation of *K* employs the following mathematical formulae:

$$\frac{\partial \pi_i(\boldsymbol{\theta})}{\partial \lambda_{hj}^{l}} = \frac{x_{ih} q_l D_2}{D_1^2},$$

and

$$\frac{\partial \pi_i(\boldsymbol{\theta})}{\partial q_l} = \frac{-1}{D_1}.$$

4. Simulations

To better understand the performance of the estimators of the proposed method, a small simulation study, with a total number of 1000 simulations, was undertaken. It was not possible to carry out an extensive number of simulation studies due to the time factor associated with the running of the EM algorithm. A four list capture-recapture experiment was generated for a population size which was fixed to be 500. The heterogeneity effects (both observed and unobserved heterogeneity) were modelled by a covariate matrix and a multinomial mixture model with two latent classes. The covariate matrix consists of a standard normal random variable and an intercept. The class proportions were $q_1 = 0.2$ and $q_2 = 0.8.$ The capture probabilities of each individual were generated by setting

$$\mathbf{\Lambda}^1 = \begin{pmatrix} 0.1 & -0.5 & -0.5 & -0.5 & -0.5 \\ -0.5 & -0.1 & -0.1 & -0.1 & -0.5 \end{pmatrix},$$

where lists 1 and 3 are assumed to be negatively dependent.

The standard deviation of the 1000 population size estimates was calculated and this value was then compared with the standard error $\hat{s}$ derived as in Section 3. Based on the simulation results, the mean and standard deviation of $\hat{n}$ were 502.91 and 14.35, respectively. The mean of the estimated standard errors of the population size was $\hat{s} = 13.79.$ As such, the mean of $\hat{n}$ is close to the true population size of 500. Likewise, the mean of the standard errors estimates $\hat{s}$ is close to the empirical standard deviation of $\hat{n}.$ We also noted that the divergence rate for this simulation study was 2.9%. It is thus reasonable to conclude that the estimators of the proposed method have performed reasonably well.

5. Results

The data are taken from Stanghellini and van der Heijden [23, p. 513, Table 1] and refer to 2047 cases of diabetes in a town in northern Italy (Bruno et al. [6]). The data were recorded on the basis of four different lists: diabetic clinic and/or family physicians data list (list 1), hospital discharges data list (list 2), insulin and oral hypoglycemic data list (list 3), and reagent strips and insulin syringes data list (list 4). There also exists a categorical covariate, namely treatment, by which the diabetes patients can be split into three separate groups: diet (205 cases), hypoglycemic agents (1514 cases), and insulin (328 cases). We wish to estimate the number of cases which were missed by the four lists.

The methodology is illustrated with models involving only two latent classes to facilitate the discussion. Similarly, only models including a single first order interaction between the lists were examined. The preliminary analyses of the interactions between the lists have led us to conclude that the best model should include the interaction between list 1 and list 3.

A series of models can be regarded as competitors and they were thus analyzed to identify the best fitting one. We started by fitting a multinomial mixture model with no list interactions and covariates (NLNC) and one with list interactions but with no covariates (LNC). In this particular context, since we are modelling the effects of the covariates within the capture probability, we then fitted a multinomial mixture model with no list interactions but with covariates in the capture probability (NLCC) and the proposed model with list interactions and covariates in the capture probability (LCC). Table 1 shows the results for all these models and Table 2 sets up a comparison between various sets of models by means of the likelihood ratio test statistics. The comparisons showed that the inclusion of the list interaction significantly improved the model fit (see model comparisons 1 and 5, each with *p*-value $< 10^{-4}$). Likewise, the same effect was observed when the covariate was included in the modelling (comparisons 2 and 4, *p*-values $< 10^{-4}$). Furthermore, the results of Table 2 indicated that the addition of both the list interaction and the covariate leads to the best fitting model (comparisons 3, 4 and 5, *p*-values $< 10^{-4}$), which is LCC (the proposed model). This conclusion is also supported by the values of the Bayesian Information Criterion (see Table 1) whereby LCC is the model with the smallest BIC value. We then observed how the value of the population size estimator changes across the various models. The population estimates obtained from LNC and NLCC, respectively, are quite similar and exhibit only a small increase over the population estimate obtained from NLNC though the addition of either the list interaction or the covariate significantly improved the fit (see Table 2). On the other hand, the population estimate undergoes a considerable increase when LCC is used. Thus, this confirms the importance of adding both the covariate and the list interaction in the modelling since ignoring one of these effects might result in the estimate being biased (see Section 1).

The results from the chosen LCC are shown in Table 3. We also determined that the probabilities of class 1 and class 2 are 0.978 and 0.022, respectively, with corresponding standard errors of 0.019 and 0.007. The values of the standard errors reflect positively on the accuracy of the estimated sizes of class 1 and class 2. Using the coefficients in Table 3, we can derive estimates of probabilities (see Table 4) as illustrated in the following example: For diabetes patients following a diet and being identified by list 1, the probability of being in class 1 is $\frac{\exp(-0.086)}{[\exp(-0.086)+\exp(0)]} = 0.479,$ and the probability of being in class 2 is $\frac{\exp(0)}{[\exp(-0.086)+\exp(0)]} = 0.521.$ For diabetes patients following a diet and being identified by both lists 1 and 3, the probability of being in class 1 is $\frac{\exp(-0.086 + -0.088)}{[\exp(-0.086 + -0.088)+\exp(0)]} = 0.457,$ and the probability of being in class 2 is $\frac{\exp(0)}{[\exp(-0.086 + -0.088)+\exp(0)]} = 0.543.$ For diabetes patients receiving hypoglycemic agents and being identified by list 1, the probability of being in class 1 is $\frac{\exp(0.260)}{[\exp(0.260)+\exp(0)]} = 0.565,$ and the probability of being in class 2 is $\frac{\exp(0)}{[\exp(0.260)+\exp(0)]} = 0.435.$ For diabetes patients receiving hypoglycemic agents and being identified by both lists 1 and 3, the probability of being in class 1 is $\frac{\exp(0.260+1.095)}{[\exp(0.260+1.095)+\exp(0)]} = 0.795,$ and the probability of being in class 2 is $\frac{\exp(0)}{[\exp(0.260+1.095)+\exp(0)]} = 0.205.$ In a similar manner, the other probabilities in Table 4 can be calculated.

We noted that some of the estimated standard errors are larger than expected (see Table 3). The very large standard error estimate associated with parameter γ_{14}^{1} might be due to the fact that list 4 did not identify individuals who were following a diet. In addition, the number of individuals who were identified by list 4 (irrespective of which treatment they had received) was very low (see Stanghellini and van der Heijden [23, p. 513, Table 1]). Consequently, the amount of available information might have been insufficient for estimation purposes. It is also possible that the conditional likelihood function was rather at and this might explain why two other standard error estimates $(\lambda_{11}^{1}$ and $\lambda_{15}^{1})$ were slightly larger than we would expect.

Within class 1, for diabetes patients receiving hypoglycemic agents (as compared to those being administered insulin), the probability of being identified by list 1 (diabetic clinic and/or family physicians) is relatively high. This probability increases when the patients are also identified by list 3 (insulin and oral hypoglycemic). This seems reasonable in practice as we expect diabetes patients who are receiving hypoglycemic agents (list 3) to be primarily seen at diabetic clinics and/or by family physicians (list 1), leading to a high probability of being in both lists. This increase in probability was also observed when the capture probabilities for list 3 were computed. Within class 2, for both diabetes patients following a diet or receiving hypoglycemic agents (as compared to those being administered insulin), the probabilities of being identified by list 2 (hospital discharges) and list 4 (reagent strips and insulin syringes), respectively, are quite high. A similar pattern in the probability was observed when diabetes patients following a diet were identified by list 3. Based on the above, we can conclude that the probability of identification within class 2 is generally higher than in class 1.

6. Discussion

The model presented here for estimating the population size presented combines three related problems. As human behaviour is expected to be less homogeneous than that of animals, their capture probabilities are heterogeneous and require modelling in terms of individual characteristics such as the covariates observed during a particular study. Though covariate modelling controls to a great extent the existing heterogeneity within a dataset, it cannot account for the whole of the heterogeneity. There is still some heterogeneity which is left unexplained. Latent class modelling can help with controlling the effects of unobserved heterogeneity. The lists for identification are usually constructed from various sources so that some dependence is likely to arise between them. In this paper, we have introduced a multinomial latent class model where the effects of both observed and unobserved heterogeneity are accounted for and the assumption of list independence is also relaxed. Our method thus minimizes the risk of obtaining an inaccurate estimate of the population size when these three problems are not accounted for in the modelling.

As discussed in Section 2, the estimation technique employed via the conditional likelihood approach and the EM algorithm reduces the computational effort required. However, we were faced with the common problems associated with the use of the EM algorithm: we encountered slow convergence and selecting an appropriate set of initial values proved to be quite a tedious task. We also had to test multiple sets of initial values so as to avoid the problem of local maxima which frequently occurs in latent class analysis. The choice of the conditional likelihood approach over the full likelihood approach is explained by the fact that our modelling involves covariates. If the latter approach is adopted, then the EM algorithm is quite complicated as it would involve the handling of missing data, i.e., the missing covariate information of those uncaptured individuals.

Future research will involve developing a more unified framework where the available covariates will affect both the capture probability and the latent class membership. This future work will be an extension of the model proposed in this paper, with the latent class membership being modelled in terms of the available covariates through a multinomial logit model.

Acknowledgement

The first author wishes to thank Prof. P. G. M van der Heijden (Utrecht University, Utrecht, The Netherlands) for his suggestion to use the diabetes dataset in Section 5.

References

[1] J. M. Alho, Logistic regression in capture-recapture models, Biometrics 46 (1990), 623-635.

[2] F. Bartolucci and A. Forcina, A class of latent marginal models for capture-recapture data with continuous covariates, J. Amer. Statist. Assoc. 101 (2006), 786-794.

[3] D. Böhning, E. Dietz, R. Kuhnert and D. Schön, Mixture models for capture-recapture count data, Stat. Methods Appl. 14 (2005), 29-43.

[4] D. Böhning and R. Kuhnert, Equivalence of truncated count mixture distributions and mixtures of truncated count distributions, Biometrics 62 (2006), 1207-1215.

[5] D. Böhning, Editorial - Recent developments in capture-recapture methods and their applications, Biom. J. 50 (2008), 954-956.

[6] G. Bruno, A. Biggeri, R. E. LaPorte, D. McCarty, F. Merletti and G. Pagano, Application of capture-recapture to count diabetes?, Diabetes Care 17 (1994), 548-556.

[7] A. Chao, P. K. Tsay, S. H. Lin, W. Y. Shau, and D. Y. Chao, The applications of capture-recapture models to epidemiological data, Stat. Med. 20 (2001), 3123-3157.

[8] M. J. L. F. Cruyff and P. G. M. van der Heijden, Point and interval estimation of the population size using a zero-truncated negative binomial regression model, Biom. J. 50 (2008), 1035-1050.

[9] A. G. Davies, R. M. Cormack and A. M. Richardson, Estimation of injecting drug users in the city of Edinburgh, Scotland, and number infected with human immunodeficiency virus, Int. J. Epidemiol. 28 (1999), 117-121.

[10] A. P. Dempster, N. M. Laird and D. B. Rubin, Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion), J. Roy. Statist. Soc. Ser. B 39 (1977), 1-38.

[11] R. Q. Gurgel, J. D. C. da Fonseca, D. Neyra-Castañeda, G. V. Gill and L. E. Cuevas, Capture-recapture to estimate the number of street children in a city in Brazil, Arch. Dis. Child. 89 (2004), 222-224.

[12] E. B. Hook and R. R. Regal, Effect of variation in probability of ascertainment by sources (“variable catchability”) upon “capture-recapture” estimates of prevalence, Am. J. Epidemiol. 137 (1993), 1148-1166.

[13] R. M. Huggins, On the statistical analysis of capture-recapture experiments, Biometrika 76 (1989), 133-140.

[14] R. Kuhnert, V. J. Del Rio Vilas, J. Gallagher and D. Böhning, A bagging-based correction for the mixture model estimator of population size, Biom. J. 50 (2008), 993-1005.

[15] R. E. LaPorte, S. R. Dearwater, Y. F. Chang, T. J. Songer, D. J. Aaron, R. L. Anderson and T. Olsen, Efficiency and accuracy of disease monitoring systems: application of capture-recapture methods to injury monitoring, Am. J. Epidemiol. 142 (1995), 1069-1077.

[16] P. F. Lazarsfeld and N. W. Henry, Latent Structure Analysis, Houghton Mifflin, Boston, 1968.

[17] W. A. Link, Nonidentifiability of population size from capture-recapture data with heterogeneous detection populations, Biometrics 59 (2003), 1123-1130.

[18] C. X. Mao, Computing an NPLME for a mixing distribution in two closed heterogeneous population size models, Biom. J. 50 (2008), 983-992.

[19] A. L. McCutcheon, Latent Class Analysis, Sage Publications, Beverly Hills, 2000.

[20] G. McLachlan and D. Peel, Finite Mixture Models, Wiley, New York, 2000.

[21] K. H. Pollock, Capture-recapture models, J. Amer. Statist. Assoc. 95 (2000), 293-296.

[22] C. J. Schwarz and G. A. F. Seber, Estimating animal abundance: review III, Statist. Sci. 14 (1999), 427-456.

[23] E. Stanghellini and P. G. M. van der Heijden, A multiple-record systems estimation method that takes observed and unobserved heterogeneity into account, Biometrics 60 (2004), 501-516.

[24] J. Thandrayen and Y. Wang, A latent variable regression model for capture-recapture data, Comput. Statist. Data Anal. 53 (2009), 2740-2746.

[25] J. Thandrayen and Y. Wang, Capture-recapture analysis with a latent class model allowing for local dependence and observed heterogeneity, Biom. J. 52 (2010), 552-561.

[26] G. Verlato and M. Muggeo, Capture-recapture method in the epidemiology of Type 2 diabetes: a contribution from the Verona diabetes study, Diabetes Care 23 (2000), 759-764.

[27] J. K. Vermunt, Latent class regression analysis, Workshop at the 24th Biennial Conference of SMABS, Jena University, 17-22 July 2004.

[28] Y. Wang and J. Thandrayen, Multiple-record systems estimation using latent class models, Aust. N. Z. J. Stat. 51 (2009), 101-111.

[29] Y. Wang and P. S. F. Yip, A semiparametric model for recapture experiments, Scand. J. Statist. 30 (2003), 667-676.

[30] J. T. Wittes, T. Colton and V. W Sidel, Capture-recapture methods for assessing the completeness of case ascertainment when using multiple information sources, J. Chronic. Dis. 2 (1974), 25-36.

[31] E. N. Zwane and P. G. M. van der Heijden, Population estimation using the multiple system estimator in the presence of continuous covariates, Stat. Model. 5 (2005), 39-52.

Appendix

Table 1. Estimates from different models

Model	No. of parameters	–2log-likelihood	BIC	Population estimate
NLNC	5	7556.212	7594.333	2228.910(14.357)
LNC	6	7542.806	7588.551	2380.171(54.090)
NLCC	9	7075.457	7144.074	2396.126(58.092)
LCC†	11	7043.595	7127.460	2619.750(95.769)

†: the proposed model; BIC: Bayesian Information Criterion standard errors are given in brackets.

Table 2. Likelihood ratio test statistics among different models

Model comparison	LRT	df	*p*-value
1. NLNC v/s LNC	13.406	1	2.508×10^{-4}
2. NLNC v/s NLCC	480.755	4	$< 10^{-4}$
3. NLNC v/s LCC	512.617	6	$< 10^{-4}$
4. LNC v/s LCC	499.211	5	$< 10^{-4}$
5. NLCC v/s LCC	31.862	2	$< 10^{-4}$

LRT: likelihood ratio test; df: degrees of freedom.

Table 3. Estimates from chosen LCC

$\boldsymbol{\Lambda}$ entry	Parameter	Estimate	s.e.	*z*-value	*p*-value
Class 1					
λ_{11}^{1}	List 1: *D*	–0.086	0.410	–0.210	0.834
λ_{12}^{1}	List 2: *D*	–2.535	0.308	–8.231	$< 10^{-4}$
λ_{13}^{1}	List 3: *D*	–2.579	0.497	–5.189	$< 10^{-4}$
λ_{14}^{1}	List 4: *D*	–11.244‡			
λ_{15}^{1}	$(\text{List } 1 \times \text{List } 3) : D$	–0.088	0.579	–0.152	0.879
λ_{21}^{1}	List 1: *H*	0.260	0.160	1.625	0.104
λ_{22}^{1}	List 2: *H*	–1.963	0.084	–23.369	$< 10^{-4}$
λ_{23}^{1}	List 3: *H*	–0.943	0.174	–5.420	$< 10^{-4}$
λ_{24}^{1}	List 4: *H*	–4.942	0.499	–9.904	$< 10^{-4}$
λ_{25}^{1}	$(\text{List } 1 \times \text{List } 3) : H$	1.095	0.185	5.919	$< 10^{-4}$
Class 2					
λ_{11}^{2}	List 1: *D*	0			
λ_{12}^{2}	List 2: *D*	0			
λ_{13}^{2}	List 3: *D*	0			
λ_{14}^{2}	List 4: *D*	0			
λ_{15}^{2}	$(\text{List } 1 \times \text{List } 3) : D$	0			
λ_{21}^{2}	List 1: *H*	0			
λ_{22}^{2}	List 2: *H*	0			
λ_{23}^{2}	List 3: *H*	0			
λ_{24}^{2}	List 4: *H*	0			
λ_{25}^{2}	$(\text{List } 1 \times \text{List } 3) : H$	0			

s.e.: standard error; *D*: diet (with insulin as baseline); *H*: hypoglycemic agents (with insulin as baseline).

List 1: diabetic clinic and/or family physicians; List 2: hospital discharges; List 3: insulin and oral hypoglycemic; List 4: reagent strips and insulin syringes (lists identifying diabetes patients). List 1 × List 3: interaction between lists 1 and 3. The $\boldsymbol{\Lambda}$ entries for class 2 are all zero due to the constraint $\boldsymbol{\Lambda}^{L} = 0,$ which is employed for identifiability purposes.

‡: the s.e. is very large rendering the estimate useless.

Table 4. Estimated probabilities from chosen LCC

		Latent class	
List	Covariate	1	2
1	D	$\begin{cases} 0.479 \text{ if not in list 3;} \\ 0.457 \text{ if also in list 3.} \end{cases}$	$\begin{cases} 0.521 \text{ if not in list 3;} \\ 0.543 \text{ if also in list 3.} \end{cases}$
	H	$\begin{cases} 0.565 \text{ if not in list 3;} \\ 0.795 \text{ if also in list 3.} \end{cases}$	$\begin{cases} 0.435 \text{ if not in list 3;} \\ 0.205 \text{ if also in list 3.} \end{cases}$
2	D	0.073	0.927
	H	0.123	0.877
3	D	$\begin{cases} 0.071 \text{ if not in list 1;} \\ 0.065 \text{ if also in list 1.} \end{cases}$	$\begin{cases} 0.929 \text{ if not in list 1;} \\ 0.935 \text{ if also in list 1.} \end{cases}$
	H	$\begin{cases} 0.280 \text{ if not in list 1;} \\ 0.538 \text{ if also in list 1.} \end{cases}$	$\begin{cases} 0.720 \text{ if not in list 1;} \\ 0.462 \text{ if also in list 1.} \end{cases}$
4	D	⋆	⋆
	H	0.007	0.993

D: diet (with insulin as baseline); H: hypoglycemic agents (with insulin as baseline). 1: diabetic clinic and/or family physicians; 2: hospital discharges; 3: insulin and oral hypoglycemic; 4: reagent strips and insulin syringes (lists identifying diabetes patients).

⋆: calculations are not valid because the s.e. is very large (see Table 3).