Latent Class Analysis for Estimating an Unknown Population Size - with application to censuses

Bernard Baffour[1], James J. Brown[2,3] and Peter W. F. Smith[4,5]

1. School of Demography,

Australian National University, Canberra, Australia.

Email: bernard.baffour@anu.edu.au

2. School of Mathematical and Physical Sciences,

University of Technology Sydney (UTS), Sydney, Australia.

Email: james.brown@uts.edu.au

3. Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS),

University of Technology Sydney (UTS), Sydney, Australia.

4. Department of Social Statistics and Demography,

University of Southampton, Southampton, United Kingdom.

Email: p.w.smith@soton.ac.uk

5. Southampton Statistical Sciences Research Institute (S3RI),

University of Southampton, Southampton, United Kingdom.

Abstract

Estimation of the unknown population size using capture-recapture techniques relies on the key assumption that the capture probabilities are homogeneous across individuals in the population. This is usually accomplished via post-stratification by some key covariates believed to influence individual catchability. Another issue that arises in population estimation from data collected from multiple sources is list dependence, where an individual's catchability on one list is related to that of another list. The earlier models for population estimation heavily relied upon list independence. However, there are methods available that can adjust the population estimates to account for dependence amongst lists. In this paper, we propose the use of latent class analysis through log-linear modelling to estimate the population size in the presence of both heterogeneity and list dependence. The proposed approach is illustrated using data from the 1988 US census dress rehearsal.

Keywords: capture-recapture, latent class analysis, log-linear models

1 Introduction

In the original development of capture-recapture methods in application to wildlife population measurement (see Seber (1986)), animals were captured, marked and recaptured resulting in two incomplete lists. Estimation of the unknown population size then relied on a set of assumptions. Firstly, there is no change in the population between captures (i.e. the population is closed). Secondly, individuals can be matched from capture to recapture (without error). Thirdly, there is homogeneity of capture or recapture (i.e. on each sampling occasion all individuals have the same capture probability). Fourthly, there is independence between the capture and recapture processes. In fact, the third and fourth assumptions are connected since independence implies capture does not affect recapture. However, it is convenient to state them separately, and it will be shown that, in particular for human populations, the homogeneity and list independence assumptions are different. These assumptions are all intertwined and a failure of any one can invalidate the others leading to biased estimates of the population (International Working Group for Disease Monitoring and Forecasting (1995); Zhang (2019)). The earliest paper that applied capture-recapture for the measurement of human populations, Chandrasekar and Deming (1949), discussed the practical problems of ensuring (list) independence and homogeneity. Both heterogeneity and list dependence result in biased population estimates. This bias is termed 'correlation bias' (Alho et al. (1993); Brown, Abbott and Diamond (2006)) and can be due to two types of dependence:

(a) List dependence: the act of being included in the first list makes an individual more or less likely to be included in the second list, i.e. inclusion in the first sample has a causal effect on inclusion in the second sample. This is sometimes referred to as *causal dependence*.

(b) Heterogneity: even if the two lists are independent within individuals, the lists may become dependent if the capture probabilities are not the same (i.e. not homogenous, or are heterogenous) amongst individuals. This is similar to the Simpson paradox which shows that an aggregation of two independent 2×2 tables may result in a dependent table. This is sometimes referred to as *apparent dependence* (see, for example, Cormack (1972) and Coull and Agresti (1999)).

Although possible for animal populations, where some degree of control can be exercised by the experimenter to ensure list (in)dependence and homogeneity (heteregeneity) of capture, this can be difficult to ensure in human populations. Specifically, the difficulty is that these two types of dependence are confounded and cannot be separated unless additional information is provided. In the case where there are three capture occasions, the third list allows for the possibility of examining the list dependence between pairs of lists. Subsequently the independence assumption that underpins the two list capture-recapture problem is no longer necessary. Nonetheless, the homogeneity assumption is still needed.

In practice, heterogeneity (lack of homogeneity) can be accounted for by dividing the population into homogeneous sub-groups through post-stratification (Chandrasekar and Deming (1949)). This is often undertaken in population censuses on the basis of geography, race/ethnicity, housing characteristics, age and sex (for example, Hogan (1992) for the US and Brown et al. (1999) for the UK). When

the covariates that account for the heterogeneity of capture are continuous, instead of categorical, so that in effect there are as many categories as individuals, Rasch-type models can be used (Agresti (1994) and Fienberg et al. (1999)). The choice of these covariates to ensure that capture is homogeneous across individuals requires a great deal of effort, and it is inevitable that in some applications, there is a failure to account for all the heterogeneity leading to inaccurate estimates of the population (Chao (2001)). Also for post-stratification to properly work, the independence assumption needs to hold within each strata.

An alternative approach to estimating the population size is to assume that individuals cluster into latent classes, such that individuals within the same class have the same chance of being captured. Under latent class analysis, the assumption is that the whole population can be subdivided into L subgroups but the choice of these classes is unknown. For population estimation using latent class analysis, however, there needs to be at least four lists, or three lists with some constraints to ensure model identifiability (Goodman (1974)). Several authors have discussed latent class analysis within a capture-recapture framework. For instance, Agresti (1994) fits various latent class models to estimate the population of snowshoe hares. In an application to human populations, Bruno et al. (1994) estimate the incidence of diabetes in the northern Italian town of Casale Monteferrato, while Wang and Thandrayen (2009) use a similar approach to estimate the number of homeless people in the Australian city of Adelaide's central business district. The current paper concentrates on an application to population censuses as proposed by Biemer et al. (2001). In official statistics, the Dutch have been using log-linear modelling to provide population estimates when linking information from multiple registers (Gerritse et al. (2015)). More recently, statisticians from the Italian National Statistical Insitute applied this approach to estimate the number of active local enterprises for the production of business statistics (Di Cecco et al. (2018)). In our application, we present a model that can provide estimates of the unknown population from three lists when the assumptions of homogeneity of capture probabilities of individuals, and marginal independence of the lists are violated.

When estimating the unknown population size within a capture-recapture framework, many authors have considered latent class modelling to account for both list dependence and heterogeneity using additional covariate information; see for example Stanghellini and van der Heijden (2004) and Bartolucci and Forcina (2006). However, for the case where there are three data sources, the latent model cannot cope with list dependence due to problems with model identifiability unless there are some restrictions (usually equality constraints are placed on the conditional probabilities) or a continuous covariate relating to the capture probabilities is available (resulting in a logistic regression model). As an alternative, we propose an identifiable latent class model that can cope with heterogeneity of individual capture probabilities and dependence between the lists using a categorical covariate. As stated, the basic methodology described has been used in many areas. However, for population estimation (i.e. census measurement) the costs associated with multiple capture-recapture methods have meant that most national statistical institutes have to make a trade-off between the number of sources and overall quality of data, and as such constrained the number of lists to two - a census and a post-enumeration survey. Our main contribution is to extend the literature and investigate the use of latent class models in population estimation to account for both list dependence

and heterogeneity when data has been collected from three lists (here a census, a post-enumeration survey and an administrative register).

The outline of the paper is as follows. In Section 2 we lay out the general framework for the estimation of the unknown population size within a capture-recapture context. In Section 3 we apply various population estimation approaches to data gathered as part of the US 1988 census dress rehearsal. 3.1 provides a description of the data and the post-stratification scheme used to ensure that the homogeneity assumption holds, while 3.2 presents the population estimation using log-linear models fitted separately to the different post-strata. In 3.3 we propose a log-linear model fitted simultaneously to the post-strata as an alternative and more efficient parameterization for population estimation. In Section 4 we demonstrate how log-linear modelling can be extended to fit latent class models allowing the population size to be estimated when there is both observed and unobserved heterogeneity. 4.1 introduces the latent class model and embeds it within a log-linear modelling framework. In 4.2 we extend this log-linear model to account for both heterogeneity and missingness (i.e unobserved cells) through effectively using the post-stratification information. This proposed modelling approach is applied in 4.3 to the US dress rehearsal data, and the results compared to the previous modelling results in Section 3. We find that the conventional approach in Section 3 does not work, but our proposed approach seems to work. Finally, we conclude with a brief discussion.

2 Population size estimation for incomplete contingency tables

In the simplest version of the capture-recapture model, there is an initial capture followed by a subsequent recapture of the closed population of interest. The individuals that are found or missed by the two lists can be placed in a 2×2 contingency table (see Table 1). The estimate of the individuals missed by both lists, $\hat{n}_{00}$, is found by assuming that, firstly, there is independence between the two lists and, secondly, that individuals have a constant probability of capture or recapture. These assumptions are ultimately untestable unless additional information can be provided (Seber (1986)). Mathematically, the estimate of the missing cell count, $\hat{n}_{00}$, is given by $\rho\frac{n_{01}n_{10}}{n_{11}}$, where ρ is the cross product ratio or dependence parameter. Since the depedence parameter cannot be estimated, we make the assumption that $\rho = 1$, i.e. that the samples are independent of each other.

Isaki and Schultz (1986) suggested several alternative dual list estimates to incorporate dependence (i.e. $\rho \neq 1$). Wolter (1990) and Brown, Abbott and Diamond (2006) suggest using demographic or other information to assess dependence, while a number of authors, including Alho (1990) and Darroch et al. (1993), propose using explanatory variables. It is important to note that the dependence being discussed here is due to individuals not having the same probability of capture, i.e. heterogeniety.

The obvious extension to overcome the restrictive assumptions imposed by the two-capture model is to increase the number of capture occasions. This has been the preferred approach in ecological literature (for example, Cormack (1972)). The added advantage when there are three lists is that the inter-relationships between the various captures can now be explored, in particular the (in)dependence between the first and second captures can be investigated. Having three or more capture occasions

Table 1: Two sample capture-recapture

		Second Sample	
		Counted	Missed
First Sample	Counted	n_{11}	n_{10}
	Missed	n_{01}	n_{00}

allows both list dependence and heterogeneity of capture to be investigated during population estimation.

For the situation when there is information collected about individuals on three separate occasions, the capture history can be represented in a 2×2×2 contingency table (see Table 2), with the missing cell denoted by n_{000}.

Table 2: Three list general capture-recapture problem

		Third List			
		Counted		Missed	
		Second List		Second List	
		Counted	Missed	Counted	Missed
First List	Counted	n_{111}	n_{101}	n_{110}	n_{100}
	Missed	n_{011}	n_{001}	n_{010}	n_{000}

The incomplete 2×2×2 table of counts can be divided into one complete 2×2 sub-table and one incomplete 2×2 sub-table (with the missing n_{000}-cell). If we assume that the cross-product ratio ρ is the same in both sub-tables, we can use the information from the complete sub-table to estimate the missing cell in the incomplete sub-table. Mathematically, for the complete sub-table, the cross-product ratio can be written as $\rho = \frac{n_{111}n_{001}}{n_{011}n_{101}}$. For the incomplete sub-table, we have $\rho = \frac{n_{110}n_{000}}{n_{100}n_{010}}$.

We can then apply the dual-system estimate to the incomplete sub-table, to obtain

$$\hat{n}_{000} = \hat{\rho} \times \frac{n_{100}n_{010}}{n_{110}} = \frac{n_{111}n_{100}n_{010}n_{001}}{n_{011}n_{101}n_{110}}. \tag{1}$$

The specification of the individual capture histories in the form of an incomplete contingency table allows us to use log-linear models (Fienberg (1972)). The likelihood can be estimated through working with the conditional probabilities given the observed frequencies (i.e., n_{ijk} for $(i,j,k) \neq (0,0,0)$). Once the loglinear model parameters have been (conditionally) estimated, given the observed counts, the estimate of the unobserved, missing, cell count $\hat{n}_{000}$ can be generated based on the conditional maximum likelihood estimation (Darroch (1958); Fienberg (1972)). The log-linear based estimators are built from explicit considerations of the heterogeneity amongst individuals and the dependence between the lists (but in capture-recapture, the probabilities are assumed to be homogenous amongst individuals within the same capture profile). In addition, the goodness of fit of these models can be formally tested. Therefore, the log-linear modeling framework for capture-recapature is intuitively

appealing since it allows for dependence among lists and heterogeneity of capture (Gerritse et al. (2015)). Based on the selected model, the missing cell can be estimated thereby leading to the total population estimate. Additionally appealing is the fact that for all models, either closed form solutions exist or they can be estimated through iterative techniques (see Chapter 6 of (Bishop et al., 1975)).

Following the same notation introduced by Bishop et al. (1975), let μ_{ijk} be the expected number of individuals in the $(i,j,k)^{th}$ cell of the 2×2×2 contingency table, then the ('saturated') log-linear can be specified as

$$\log \mu_{ijk} = \lambda + \lambda_i^{(1)} + \lambda_j^{(2)} + \lambda_k^{(3)} + \lambda_{ij}^{(12)} + \lambda_{ik}^{(13)} + \lambda_{jk}^{(23)} + \lambda_{ijk}^{(123)}, \tag{2}$$

where $\lambda_i^{(1)}$, $\lambda_j^{(2)}$, $\lambda_k^{(3)}$ are the main effect terms, $\lambda_{ij}^{(12)}$, $\lambda_{ik}^{(13)}$, $\lambda_{jk}^{(23)}$ are the two-way interaction terms, and $\lambda_{ijk}^{(123)}$ is the three-way interaction term.

When we have an incomplete 2×2×2 contingency table, with μ_{000} representing the unobserved ('missing') cell, the saturated model is not indentifiable (in that we have eight parameters but seven observable cell counts). The implication of considering only hierarchical models (Fienberg (1972)), is that the highest order interaction, three-way term, $\lambda_{ijk}^{(123)}$, is set to zero, and our 'saturated model' becomes

$$\log \mu_{ijk} = \lambda + \lambda_i^{(1)} + \lambda_j^{(2)} + \lambda_k^{(3)} + \lambda_{ij}^{(12)} + \lambda_{ik}^{(13)} + \lambda_{jk}^{(23)}. \tag{3}$$

When $\lambda_{ijk} = 0$, this states that each of the two-factor effects (i.e. $\lambda_{ij}^{(12)}$, $\lambda_{ik}^{(13)}$, and $\lambda_{jk}^{(23)}$) is unaffected by the level of the third variable. Bartlett (1935) was the first to show that under the no-three-way interaction model, the cross product ratio specified by this model

$$\frac{\hat{\mu}_{001}\hat{\mu}_{010}\hat{\mu}_{100}\hat{\mu}_{111}}{\hat{\mu}_{000}\hat{\mu}_{011}\hat{\mu}_{101}\hat{\mu}_{110}} = 1 \quad \text{holds, which implies that the 2×2 odds are equal,} \tag{4}$$

where $\hat{\mu}_{ijk}$ are the maximum likehood estimates of the $(i,j,k)^{th}$ cell, under the specifed log-linear model.

This no-three-way interaction assumption is analagous to the assumption of independence in the 2×2 case: all pairs of lists can exhibit dependence, but the amount of dependence in each pair is assumed to be uninfluenced after conditioning on the third list (Darroch (1958), Fienberg (1972) and Darroch et al. (1993)). It now becomes possible to define various unsaturated hierarchical models by setting λ-terms to be equal to zero. The restriction for all models under consideration to be hierarchical implies that when a particular λ-term is set to zero then all of the higher-order relatives are also zero.

Crucially in capture-recapture population estimation, the best model is the one with the fewest possible parameters that allows for the dependencies amongst the lists (Fienberg (1972)). This model is then used to predict the missing cell, and subsequently estimate the population size. Closed form solutions exist for all models, apart from when all three lists are independent ((Fienberg (1972), Darroch et al. (1993)). For this case, the iterative proportional fitting algorithm (Deming and Stephan (1940)) can be used. Further, there are a number of techniques available to provide estimates of precision of these population estimates, such as the Supplemented EM algorithm (Meng and Rubin (1991)) or the profile likelihood (Cormack (1992)). For the estimators of the population total from log-linear models, Bishop et al. (1975, pages 237-242) derive variance estimates using the delta method.

Our preferred approach will be to use a bootstrap procedure similar to that suggested by Buckland and Garthwaite (1991). Note that the reason for fitting the most parsimonious model is in order that the variance of the estimate of the missing cell count, and hence the estimate of the unknown population size, can be as small as possible: the simpler the model, the smaller the variance (Fienberg (1972)).

3 An application to the 1988 US census dress rehearsal

3.1 Description

To investigate the performance of log-linear and latent class models for estimating the unknown population, an application to population censuses with data from the US Census Bureau was used. Previous censuses had shown that Black males are much more likely to be missed in the census processes (Darroch et al. (1993), Zaslavsky and Wolfgang (1990) and Zaslavsky and Wolfgang (1993)). Therefore, in the lead-up to the 1990 census, the Census Bureau carried out a census dress rehearsal in a district of St Louis, Missouri, an area chosen because most residents were expected to be Black renters. A census was carried out, shortly followed by a survey. From a combination of administrative registers, an administrative list was created and based on key demographic identifiers the three lists were matched. However, due to the difficulties in determining correct matches, a large number of records were removed, so that the final data set had around 1,000 observed people. The data have been restricted by age and sex to fall within four post-strata: Black Males aged 20-29 in Owned homes (Young Owners), Black Males aged 30-44 in Owned homes (Old Owners), Black Males aged 20-29 in Rented homes (Young Renters) and Black Males aged 30-44 in Rented homes (Old Renters), and are given in Table 3. After classifying the respondents in the dress rehearsal into whether or not they appeared on the census (i.e. First List, denoted C), post-enumeration survey (i.e. Second List, S) or the Administrative List (i.e. Third List, L), post-stratified by age and tenure, estimates of the total population can be derived (including those that are missing in all three lists).

Table 3: Three Sample data from 1988 US Census Dress Rehearsal

Cell	Young Owners	Young Renters	Old Owners	Old Renters
n_{000}	-	-	-	-
n_{001}	59	43	35	43
n_{010}	8	34	10	24
n_{011}	19	11	10	13
n_{100}	31	41	62	32
n_{101}	19	12	13	7
n_{110}	13	69	36	69
n_{111}	79	58	91	72
n	228	268	257	260

Source: Zaslavsky and Wolfgang (1993).

3.2 Estimation of the missing cell counts through log-linear modelling

The estimates of the missing cell under the different log-linear models are shown in Table 4, using results from Bishop et al. (1975, Chapter 6). The log-likelihood chi-squared statistic (i.e. the Deviance) is found by comparing the expected counts to the observed counts in all the cells but the missing cell. The results presented in the table are found by fitting 8 different models to each of the four strata (i.e. Young Owners, Young Renters, Old Owners and Old Renters). From Table 4, the three sources have some definite inter-relationships, and the size of the Deviance statistics show that the model assuming complete independence (i.e. the Census, Survey and Administrative List are independent of each other) poorly fits the data, across all four post-strata. There is evidence to suggest that there is some dependence between the census list and the survey list. Additionally, being counted in the survey appears related to whether or not an individual is found on the administrative list. It can be concluded from these that the best fitting model is the one that accounts for the pairwise interaction terms between the Census and Survey, and the Survey and Administrative List.

Table 4: Estimate of the missing cell counts, standard errors, and Deviance under different models. The bootstrap standard errors are provided in parenthesis.

Model		Young Owners	Young Renters	Old Owners	Old Renters	df (per stratum)
Independence	$\hat{n}_{000}$	13.8 (8.6)	28.4 (18.2)	14.3 (10.9)	18.2 (15.3)	3
{C, S, L}	*Deviance*	72.59	54.83	90.19	76.20	
{L,CS}	$\hat{n}_{000}$	24.0 (19.2)	26.0 (22.4)	24.4 (23.6)	17.3 (21.3)	2
	Deviance	59.01	54.23	62.54	76.06	
{S,CL}	$\hat{n}_{000}$	7.9 (6.4)	23.7 (19.8)	8.0 (7.7)	12.8 (13.4)	2
	Deviance	68.55	52.80	84.54	70.73	
{C,SL}	$\hat{n}_{000}$	26.2 (17.7)	76.4 (26.4)	33.2 (29.6)	58.4 (28.3)	2
	Deviance	34.46	12.19	59.27	15.71	
{CS, CL}	$\hat{n}_{000}$	19.1 (28.3)	20.2 (29.0)	17.2 (33.4)	11.1 (24.5)	1
	Deviance	58.71	51.58	61.25	69.90	
{CS, SL}	$\hat{n}_{000}$	96.2 (42.5)	146.8 (55.6)	166.8 (70.6)	196.2 (50.1)	1
	Deviance	3.15	6.53	3.55	3.04	
{CL, SL}	$\hat{n}_{000}$	24.8 (21.8)	132.8 (56.6)	35.0 (42.4)	79.3 (59.0)	1
	Deviance	34.44	8.78	59.25	14.73	
'Saturated'	$\hat{n}_{000}$	245.1 (0)	379.7 (0)	418.8 (0)	378.7 (0)	0
{CS, CL, SL}	*Deviance*	0	0	0	0	

It is, however, noticeable that the estimate of the missing cell under the 'selected' parsimonious model and the 'saturated' model are seemingly different - for some post-strata the estimate under the $\{CS, CL, SL\}$ model was almost three times the size of that under model $\{CS, SL\}$. The estimates of the missing are 96 Young Owners, 147 Young Renters, 167 Old Owners and 196 Old Renters under the reduced model, while the respective estimates under the 'saturated' model are 245, 380, 419 and 379. Additionally, the confidence intervals for all the post-strata under the chosen model do not contain the 'saturated' model estimates, implying that the estimates of the population size under the 'best' fitting model and the 'saturated model' are very different. This brings into doubt the assumption that

there is no three-way interaction between the Census, Survey and Third List.

A crucial prior assumption is that there is no unaccounted heterogeneity, and for cases where there is additional heterogeneity not fully corrected for by the post-stratification mechanism, the no-three-way assumption might fail. As mentioned earlier, the no-three-way assumption is important because we have an incomplete contingency table and the (likelihood) estimation of the parameters from each log-linear model relies on information from only the observed cells: in other words, the information about the dependence structure in the contingency table is fully provided by the observed cell counts.

3.3 An alternative parameterisation of the log-linear model with a grouping covariate

It is generally well accepted that post-stratification is the most efficient method of ensuring that there is homogeneity of capture, which implies that any remaining dependence is due to dependence between lists (International Working Group for Disease Monitoring and Forecasting (1995)). However, the advantage of the log-linear modelling framework is that it provides a convenient specification for including the post-stratification variables directly into the model. Here the post-stratified variables can be thought of as a grouping covariate, G, such that the 'saturated' model (with the G) becomes

$$\log \mu_{ijkg} = \lambda+\lambda_i^{(C)}+\lambda_j^{(S)}+\lambda_k^{(L)}+\lambda_G^{(g)}+\lambda_{ij}^{(CS)}+\lambda_{ik}^{(CL)}+\lambda_{jk}^{(SL)}+\lambda_{ig}^{(CG)}+\lambda_{jg}^{(SG)}+\lambda_{kg}^{(LG)}+\lambda_{ijg}^{(CSG)}+\lambda_{ikg}^{(CLG)}+\lambda_{jkg}^{(SLG)}. \quad (5)$$

The log-linear model, given in equation (5), directly includes the combined post-strata and contains parameters that identify both the list effects and the effects of heterogeneity of capture.

The benefit of incorporating the post-strata variables directly into the log-linear modelling framework is that there are more combinations of variables that can be considered. In particular, we can consider various two-way and three-way interaction terms. In essence, simultnaeously modelling over strata has the advantage over modelling each stratum separately since it enables selection of more parsimonious models through restricting certain parameters to be equal over specific strata (Agresti (1994)).

We re-analyzed the 1988 US census dress rehearsal data using the post-strata as covariates in the log-linear model. We contrast our results to the previous analysis. To select the best model, we use a stepwise selection procedure procedure to sequentially remove (backward elimination) or add in (forward selection) terms for which the resulting change in the AIC is smallest (or biggest). To ensure better model interpretability, we restricted the models under consideration to be hierarchical, and included all lower-order terms contained in the higher-order model term. The results are presented in Table 5.

We start with Model I which has all the pairwise interaction terms betweenthe grouping covariate G and the lists (C, S and L) . This model implies that there is conditional association and this association is the same across all variables, when controlling for the other variables. This model is a good starting point for us to assess whether there are any additional association terms that are needed to reconstruct the joint distribution which defines the simultaneous behaviour of the relationship between C, S and

Table 5: Estimate of the missing cell counts, standard errors and Deviance under different models - fitted simultaneously over the post-strata. The bootstrap standard errors are provided in parenthesis.

Models		Young Owners	Young Renters	Old Owners	Old Renters	df
I. {CS, CL, SL, CG, SG, LG}	$\hat{n}_{000g}$	199.6 (74.6)	528.4 (185.0)	242.9 (87.6)	368.6 (118.8)	9
	Deviance	20.54				
II. Model I - CS	$\hat{n}_{000g}$	11.7 (14.4)	26.1 (30.6)	12.0 (11.4)	16.6 (15.7)	10
	Deviance	269.56				
III. Model I - SL	$\hat{n}_{000g}$	35.7 (21.9)	98.9 (54.6)	45.9 (48.7)	63.7 (35.4)	10
	Deviance	121.62				
IV. Model I - CL	$\hat{n}_{000g}$	92.2 (34.7)	275.9 (107.6)	113.3 (45.5)	186.6 (73.1)	10
	Deviance	33.59				
V. Model I + CSG + CLG	$\hat{n}_{000g}$	170.1 (60.0)	908.3 (323.8)	239.6 (79.3)	543.3 (180.7)	3
	Deviance	8.92				
VI. Model I + CSG + SLG	$\hat{n}_{000g}$	225.7 (13.2)	344.4 (20.1)	391.3 (24.5)	460.4 (26.5)	3
	Deviance	0.41				
VII. Model I + CLG + SLG	$\hat{n}_{000g}$	380.0 (63.8)	402.8 (69.8)	343.3 (56.5)	221.9 (32.1)	3
	Deviance	2.76				
VIII. Model I - CL + CSG + SLG	$\hat{n}_{000g}$	96.3 (38.0)	146.9 (56.2)	166.9 (66.7)	196.5 (76.8)	4
	Deviance	16.27				
'**Saturated**': {CSG, SLG, CLG}	$\hat{n}_{000g}$	246.3 (0)	381.7 (0)	421.3 (0)	378.5 (0)	0
	Deviance	0				

L and G. This assessment between competing models is done by examining the deviance statistics, and allows us to remove some pairwise interactions. However, we can see that the CS (i.e. Model II), SL (i.e. Model III) or the CL (i.e. Model IV) interaction terms are all significant and therefore should not be removed in this application.

Building on Model I, we try combinations of the three-way interactions (CSG, SLG, CLG) and find that Models V, VI and VII are not significantly different from the saturated model, while still giving contradictory estimates for some of the missing cells. As a final check, we fit Model VIII, which removes the CL term from our best fitting model, Model VI; removing any dependence between C and L after controlling for the grouping covariate G. Notice that, Model VIII with a deviance of 16.27 on 4 degrees of freedom is the same as the model chosen to be best fitting model ($\{CS, SL\}$) when fitted to individual post-strata.

There is a remarkable difference in fit between the models with all the two-way interactions, specifically those between the three different lists (i.e. CS, SL, and CL). For these models (namely, Models V, VI and VII), while the model fit (given by the deviance) appears to be better, there are large differences in the estimated number of missing in the post-strata. According to Darroch et al. (1993) and Zaslavsky and Wolfgang (1993) this may be evidence that the no-three-way interaction term assumption is problematic. Although there is (obvious) direct dependence between the lists, there is more complicated dependence which may be due to differences in characteristics and behaviour between individuals found on the different lists. Additionally, Model VIII implies that the demographic

characteristics of individuals will determine their list capture behaviour, which essentially means that the conditional associations will not vary across the four post-strata. This might be too strong an assumption.

In reality, there is a radical difference in the way the census, survey and administrative lists were constructed. According to Zaslavsky and Wolfgang (1990) the administrative list was assembled through an exhaustive search of all admininstrative registers (e.g. drivers licence registry, employment records, Internal Revenue Service records, and Veterans Administration records) for a particular geographical area covered by the census and survey areas. This makes it plausible that there will be lower or negative association between being found on the administrative list, and being found on the census or the survey. Put differently, the probability of being found on the list given you were found on the census and survey varies from individual to individual.

One simple way of ensuring that there is homogeniety of capture on the third list, might be to replace the administrative list with another field sample, for example, a pre-enumeration survey (as suggested by Darroch et al. (1993)), and this would most likely produce data that would allow us to (realistically) assume that there is no three-way interaction. In the absence of this, we propose to fit the model under a latent class framework, and assume that the latent variable can be used to account for the unobserved heterogeniety, and in effect relaxing the (untestable) no three-way-interaction assumption. The basic premise is that the variation observed among the Census, Survey and Third List is due to each of the three variables' relationship to a latent variable, and this latent variable 'explains' the relationships between the (observed) variables. Consequently, controlling for this latent variable results in a better understanding of the 'true' characterisation of the observed relationships (Lazarsfeld and Henry (1968), Goodman (1974) and Haberman (1979)). Next, we investigate how a latent class model can be fitted to the data to offer an alternative solution.

4 Framework for latent class modelling in a census allowing for local dependence

4.1 The log-linear latent class model

In the previous section, it was noted that the dependence between any pair of samples can be accounted for through log-linear modelling, but the assumption is that the individuals within the contingency table are homogeneous. Usually, post-stratification can be used to subdivide the population (by demographic, socio-economic, housing, etc. characteristics) such that within each post-stratum the individuals are homogeneous. However, the choice of these characteristics to use for post-stratification can be difficult in practice, particularly for human populations. As will be illustrated in Section 4.2, a failure in the post-stratification mechanism can lead to biased population estimates.

If these post-strata are not known a priori, a latent class model can be used to identify these groups. The aim of latent class analysis here is to define a latent (unobservable) variable with a set of classes within which the observed (manifest) variables are locally independent, implying that within a

latent subgroup the manifest variables are independent of each other. Latent class analysis and post-stratification can achieve the same purpose of ensuring homogeneity of capture within groups. The heterogeneity is caused by some characteristics which could be assumed to be known (and therefore post-stratification can be used) or unknown (and therefore characterized as a latent variable). Our approach first uses post-stratification based on known covariates that influence capture, and then fits a latent class model to account for any remaining heterogeneity. Under the log-linear modelling framework, this capture-recapature model is simple to write down.

The standard latent class model assumes local independence between the latent and manifest variables, where the manifest variables are independent of each other within latent classes. In other words, the latent variable is taken to explain all the association between the manifest variables. In the initial work by Lazarsfeld and Henry (1968), Goodman (1974) and Haberman (1979), the local independence assumption was essential for the derivation of parameter estimates under latent class modelling: the criterion of local independence provided a method for determining whether relationships amongst a set of observed measures are due to some unmeasured explanatory variable (Lazarsfeld and Henry (1968)). However, the constraints imposed by local independence may be unrealistic, and often untrue, in practice (Hagenaars (1993)). Often manifest variables are in fact related or dependent. For example, multiple indicators of poverty, or tests of related symptoms for an underlying genetic condition. Such items are termed "conditionally dependent" or "locally dependent" because there is some association within latent classes. A failure to account for this leads to issues of mis-specification of the model (for example, we might choose a model with four latent classes instead of two).

Under a local dependence latent class model, the residual association not explained by the relationship between the latent and manifest variables can be directly included. In essence, the local dependence model that accounts for any residual association not explained by the latent model can be formulated. Hagenaars (1993) suggests either including an additional latent variable or alternatively add association terms between manifest variables. Within the capture-recpature literature, these latent variable models have been used in Biggeri et al. (1999) and Stanghellini and van der Heijden (2004), and more relevant to the current paper, Gerritse et al. (2015) and Di Cecco et al. (2018) have advocated a similar approach in the use of multiple administrative lists for population estimation.

To specify the latent model, let μ_{ijkx} be the expected counts in the $(i, j, k, x)^{th}$ cell, for the observed manifest variables C, S, L, and the latent variable X. Then the latent class model under local independence, on the one hand, is

$$\log \mu_{ijkx} = \lambda + \lambda_i^{(C)} + \lambda_j^{(S)} + \lambda_k^{(L)} + \lambda_x^{(X)} + \lambda_{ix}^{(CX)} + \lambda_{jx}^{(SX)} + \lambda_{kx}^{(LX)}. \tag{6}$$

This model is not identifiable: there are more unknown parameters to be estimated than the known cell frequencies. Identifying restrictions on the parameters are therefore necessary. The usual way, in log-linear modelling, is to express each effect in terms of deviations from the average effect and impose the restriction that the λ-parameters summed over any of its subscripts equal to zero (this is referred

to as 'sum-to-zero constraints) (Goodman (1974)). As such, the identifying constraints are given as

$$\sum_i \lambda_i^{(C)} = \sum_j \lambda_j^{(S)} = \sum_k \lambda_k^{(L)} = \sum_x \lambda_x^{(X)} = 0,$$
$$\sum_i \lambda_{ix}^{(CX)} = \sum_j \lambda_{jx}^{(SX)} = \sum_k \lambda_{kx}^{(LX)} = 0,$$

and

$$\sum_x \lambda_{ix}^{(CX)} = \sum_x \lambda_{jx}^{(SX)} = \sum_x \lambda_{kx}^{(LX)} = 0.$$

On the other hand, when there is local dependence due to residual association between the Census and Survey, for example, then we can use

$$\log \mu_{ijkx} = \lambda + \lambda_i^{(C)} + \lambda_j^{(S)} + \lambda_k^{(L)} + \lambda_x^{(X)} + \lambda_{ix}^{(CX)} + \lambda_{jx}^{(SX)} + \lambda_{kx}^{(LX)} + \lambda_{ij}^{(CS)} \tag{7}$$

with identifying constraints

$$\sum_i \lambda_i^{(C)} = \sum_j \lambda_j^{(S)} = \sum_k \lambda_k^{(L)} = \sum_x \lambda_x^{(X)} = 0,$$
$$\sum_i \lambda_{ix}^{(CX)} = \sum_j \lambda_{jx}^{(SX)} = \sum_k \lambda_{kx}^{(LX)} = \sum_i \lambda_{ij}^{(CS)} = \sum_j \lambda_{ij}^{(CS)} = 0,$$

and

$$\sum_x \lambda_{ix}^{(CX)} = \sum_x \lambda_{jx}^{(SX)} = \sum_x \lambda_{kx}^{(LX)} = 0.$$

Note that when n_{000} is unobserved, both latent models (local independence, i.e. equation (6) and local dependence, i.e. equation (7)) are not identified since there are too many parameters to be estimated for the data available, and as such additional constraints are needed. Biemer et al. (2001) suggests using a two-step estimation process which first estimates the missing cell and then fits a latent model to the 'full' contingency table, but this will only work for the local independence case. Other ways of coping with non-identifiability are to impose equality constraints on some of the parameters (Formann, 2003) or simply increase the number of capture occasions Brown, Biemer and Judson (2006). In the next section, we present a modelling strategy to address the issue of identifiability through adding covariate information.

4.2 Latent class modelling to account for heterogeneity and missingness

For capture-recapture log-linear modelling, recall that the over-riding assumption is that the most complicated model that can be fitted to the data is the homogeneous association model, meaning that the conditional odds ratios between any two variables are identical for each each category of the third variable. This is equivalent to assuming that the three-way-interaction term is zero. Further, it is expected that there is a less complicated model that fits the data equally well. Under these conditions, the 'saturated' and best-fitting models are anticipated to yield similar estimates of the missing counts.

The standard local independence latent model is expected to be a poor fit to the data for two reasons. First, with only seven observed terms, additional identifying constraints are required to

fit the model. Second, there is the need to account for interaction effects between the Census and Survey and the Survey and Third List. In other words, the latent variable does not fully account for all dependence between the Census, Survey and List. This residual dependence implies that a local dependence model is required. Obviously, this model is non-identifiable, since it is over-parameterised. The suggested solution is to bring a grouping covariate, such that the effect of each manifest variable is mediated through the latent variable, pictorially represented in Figure 1 (under local independence) and Figure 2 (under local dependence), to ensure model identifiability. For the figures, the single-headed arrows are used to denote the causal direction between two variables, while the double-headed arrows are used to specify that there is no causal direction between the two variables.

Figure 1: Path diagram of the local independence model.

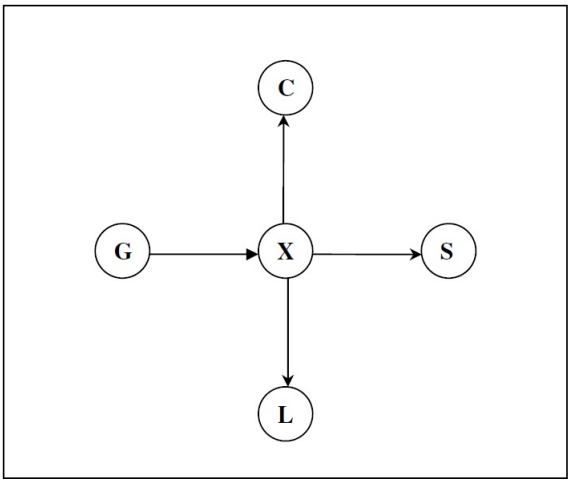
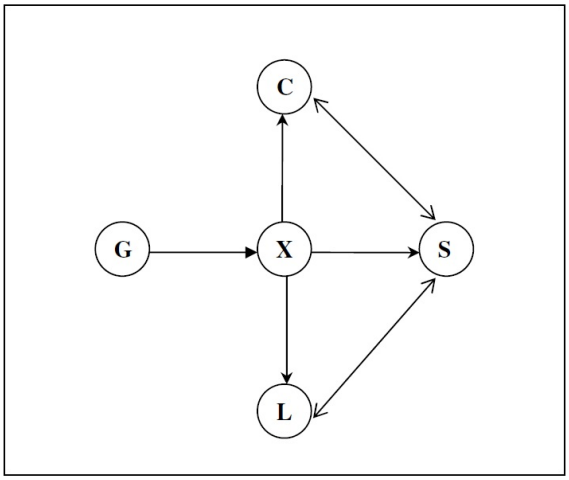


Figure 2: Path diagram of the local dependence model with two direct effects.



In the figures we see that the grouping variable G 'acts' through the latent class X to drive the counts observed on each list. In Figure 1 there is no relationship between the observed counts, after controlling for the latent class, while Figure 2 shows an additional dependency between CS and SL.

In our application, the proposed approach is to find a covariate G, that is only related to the latent variable, X, but not the manifest variables, C, S and L. This has the benefit of accounting

for the unobserved heterogeneity as well as any list dependence. In the simplest form, under local independence, this latent class model is given by

$$\log \mu_{ijkgx} = \lambda + \lambda_i^{(C)} + \lambda_j^{(S)} + \lambda_k^{(L)} + \lambda_x^{(X)} + \lambda_g^{(G)} + \lambda_{ix}^{(CX)} + \lambda_{jx}^{(SX)} + \lambda_{kx}^{(LX)} + \lambda_{gx}^{(GX)} \tag{8}$$

with additional constraints that

$$\sum_g \lambda_g^{(G)} = \sum_t \lambda_x^{(X)} = 0 \quad \text{and} \quad \sum_x \lambda_{gx}^{(GX)} = \sum_g \lambda_{gx}^{(GX)} = 0.$$

The interpretation of this model is that the residual dependence is fully accounted for through the post-strata and latent variable.

It is still possible to include further dependence terms. For instance, the model with path diagram presented in Figure 2 can be written as

$$\log \mu_{ijkgx} = \lambda + \lambda_i^{(C)} + \lambda_j^{(S)} + \lambda_k^{(L)} + \lambda_x^{(X)} + \lambda_g^{(G)} + \lambda_{ix}^{(CX)} + \lambda_{jx}^{(SX)} + \lambda_{kx}^{(LX)} + \lambda_{gx}^{(GX)} + \lambda_{ij}^{(CS)} + \lambda_{jk}^{(SL)}. \tag{9}$$

The latent class model as specified in this form does not have a closed form solution due to the number of identifying constraints. However, Haberman (1979) showed that maximum likelihood estimates for the log-linear model can be found using the iterative proportional fitting algorithm, which is equivalent to the M(aximisation)-step of the EM algorithm, since there is missing information as a result of the latent variable (Dempster et al. (1977)). This is similar to the approach suggested by Coull and Agresti (1999), Biggeri et al. (1999) and Stanghellini and van der Heijden (2004), amongst others. Initial values of the parameter estimates are essential to the convergence and speed of the EM algorithm (Dempster et al. (1977)) (and this is particularly true in the context of the capture-recapture latent class model). As such, the EM algorithm starts by finding some initial values $\mu_{ijkgx}^{(0)}$ which satisfy the log-linear model given by Equation (9). Now, given the data n_{ijkg}, with n_{000g} unobserved, the E-step consists of two sub-steps. Firstly, an estimate of the missing cells is obtained for each group g,

$$\hat{n}_{000g} = \sum_x \hat{\mu}_{000gx}, \tag{E1}$$

i.e. resulting in a 'full' observed contingency table. Secondly, the latent cells are estimated by

$$\hat{n}_{ijkgx} = \frac{n_{ijkg}}{\hat{\mu}_{ijkg}} \hat{\mu}_{ijkgx}. \tag{E2}$$

Then the M-step fits the log-linear model given by equation (9) to obtain $\hat{\mu}_{ijkgx}$. The estimated observed frequencies $\hat{\mu}_{ijkg}$ are identical to the observed frequencies n_{ijkg} when summed over the latent variable, at convergence.

This process of computing the expectation of the complete data likelihood conditional on the observed data and maximising, when repeated will converge to a solution that maximizes the (local) likelihood.

4.3 Population estimates under the latent class model - 1988 US census dress rehearsal

An investigation was carried out to determine if fitting latent class models to the US census data could improve the results. In particular, we wanted to examine whether using the age and tenancy

status of individuals as the grouping covariate, G improved the model fit. Recall that this grouping covariate has 4 levels and allows us to fit identifiable locally dependent latent class models. The interpretation of the grouping covariate is that once a person's age and gender status is accounted for the relationship amongst the manifest variables (here C, S and L) with the unobserved latent variable X, is the same in each of the four sub-tables representing the Young Owners, Young Renters, Old Owners and Old Renters. The assumption is that the effect of age and tenancy on C, S, L is completely mediated through the latent variable X. This assumption is less restrictive than the no-three-way interaction assumption required under the standard capture-recapture model with three lists. The latent variable X explains the remaining (unexplained) heterogeneity, that is not accounted for through the post-stratification scheme.

Various latent class models were fitted to the 1988 US census dress rehearsal data to investigate if latent models can account for both missingness and heterogeneity, and these results are compared to the previous results. Table 6 gives the estimates of the two latent classes under the local dependence model $\{CS, SL, CX, SX, LX, GX\}$, and bootstrapped standard errors. This model has 28 observations and 12 estimable parameters, and it is still possible to fit more complex, but identifiable, local dependence models, for instance the models with three-way interaction terms $\{SL, CX, SX, LX, GX, CSG\}$ and $\{CS, CX, SX, LX, GX, SLG\}$. However, these models were found to have qualitative similar results as the simpler model, as well as have issues with convergence during the bootstrapping. Though in our application, we have only two latent classes possible due to identifiability reasons, the choice of the number of classes plays a critical role, traditionally, in latent class modelling.

The first thing of note from the results in Table 6 is that the estimate of the missing cell in the second latent class is always zero, with all those estimated to be missing (n_{000}) placed in the first latent class. This was found to be the case whichever way the model is specified, be it under local independence or the various forms of local dependence. It can also be noticed that every person who appears in the n_{001}, n_{100} and n_{101} cell counts is placed in the first latent class. However, every person who is counted in both the Census and Survey, i.e. n_{110} and n_{111} cell counts, is placed in second latent class. The remaining cell counts, n_{010}, n_{011}, representing those people who were counted in the Survey, or the Survey and the Third list, are distributed by the latent model to both classes.

This leads us to conclude that the two latent classes represent an individual's catchability by the Survey. Put differently, a plausible interpretation is that the latent variable suggests that the unobserved heterogeneity is due to enumeration difficulty. Essentially, the observed contingency table data from the 1988 US census dress rehearsal in St Louis shows a mixture of two latent subgroups - one group of people can be described as being easy to count by the Survey, and the other subgroup are hard to count by the Survey. Furthermore, the results provide evidence that the current post-stratification mechanism fails to properly classify the population into suitably homogeneous groups such that there is no heterogeneity of capture amongst individuals within the same post-strata. The post-strata chosen on the basis of age, race and tenure is therefore inadequate, and the latent class modelling shows that there is additional heterogeneity in the data. Moreover, the estimates of the missing (roughly 155 Young Owners, 155 Young Renters, 150 Old Owners and 125 Old Renters) are closer those from the best fitting model ($\{CS, SL\}$) rather than the 'saturated model' ($\{CS, SL, CL\}$)

Table 6: Latent Class estimates of the US Census Data. The bootstrap standard errors are provided in parenthesis.

	n_{000}	n_{100}	n_{010}	n_{110}	n_{001}	n_{101}	n_{011}	n_{111}
Local Dependence (with CS and SL interactions)								
Latent Class 1								
Young Owners	**153.34**	31.00	6.56	0.00	59.00	19.00	10.84	0.00
	(**28.53**)	(5.61)	(3.72)	(0.00)	(7.47)	(4.13)	(17.92)	(0.00)
Young Renters	**155.34**	41.00	26.14	0.00	43.00	12.00	5.42	0.00
	(**28.75**)	(6.37)	(16.99)	(0.00)	(6.21)	(3.53)	(11.00)	(0.00)
Old Owners	**149.35**	62.00	7.70	0.00	35.00	13.00	4.94	0.00
	(**31.98**)	(7.36)	(4.96)	(0.00)	(5.45)	(3.58)	(9.58)	(0.00)
Old Renters	**127.26**	32.00	17.06	0.00	43.00	7.00	5.43	0.00
	(**25.21**)	(5.79)	(13.76)	(0.00)	(6.34)	(2.77)	(13.35)	(0.00)
Latent Class 2								
Young Owners	**0.00**	0.00	1.44	13.00	0.00	0.00	8.16	79.00
	(**0.00**)	(0.00)	(2.94)	(3.53)	(0.00)	(0.00)	(16.27)	(8.62)
Young Renters	**0.00**	0.00	7.86	69.00	0.00	0.00	5.58	58.00
	(**0.00**)	(0.00)	(15.61)	(8.09)	(0.00)	(0.00)	(10.99)	(7.51)
Old Owners	**0.00**	0.00	2.30	36.00	0.00	0.00	5.06	91.00
	(**0.00**)	(0.00)	(4.49)	(5.67)	(0.00)	(0.00)	(9.83)	(8.73)
Old Renters	**0.00**	0.00	6.94	69.00	0.00	0.00	7.57	72.00
	(**0.00**)	(0.00)	(13.72)	(8.31)	(0.00)	(0.00)	(14.63)	(8.52)

when compared to the results in Table 4. In this case, since the no-three-way interaction assumption cannot be justified - mainly due to a failure in the post-stratification scheme leading to the observed data being marginalised over a latent variable. As a consequence, there are issues surrounding the correct estimation of the population size.

In sum, the observed contingency table data from the 1988 US census dress rehearsal in St Louis, as they appear in Table 3, suffer from a failure in the post-stratification scheme. As such there is some residual heterogeneity not fully accounted for by the age, race and tenure post-strata. The consequence is that although the data have been post-stratified using demographic, socio-economic and household factors, some individuals with differing levels of catchability have been placed in the same post-stratum, leading to biased population estimates. Fitting the usual capture-recapture models fails to account for this, but the proposed latent class capture-recapture model offers an alternative and flexible approach, as shown by results in Table 6.

5 Conclusion

There is definitely value in using a latent variable as a way of coping with unobserved heterogeneity in the capture of individuals for population measurement when there are multiple lists. This has been fairly standard in ecological experiments (Seber (1986)), but less so within human populations. The current techniques for dealing with heterogeneity of capture rely on information being available on how sets of similar individuals are related with respect to their capture behaviour. This is because it is often

difficult to differentiate between *causal dependence*, where an individual's probability of appearing on one list depends on their probability of appearing on another list, and *apparent dependence*, where individuals do not have the same probability of appearing on a particular list. In theory an extensive, and exhaustive, post-stratification scheme should account for heterogeneity of individual capture. However, in practice, since post-stratification relies on creating discrete classifications of continuous covariates (such as age), there could be 'remaining' heterogeniety that has not been accounted for (Chen et al. (2010), Wolter (1986)).

In this paper, we have proposed a latent variable approach to estimate the population size when there is list dependence and unobserved heterogeneity of capture. Our main contribution has been to extend the current log-linear modelling framework to use the post-stratification information and by so doing expound a more flexible model. We also shed new light on the role of stratification variables in coping with both observed and unobserved heterogeneity. Our model provides a better way of examining the no-three-way interaction assumption which underpins capture-recapture population estimation with three lists. When applied to real-life census data from the US, our approach provides better population size estimates, and evidence to show that the failure of the post-stratification scheme induces dependence (i.e. heterogeneity) which invalidates the no-three-way interaction assumption.

The standard latent class model, under local independence, assumes that the population is composed of mutually exclusive latent classes such that within these classes the observed variables are unrelated. If there is reason to believe that, notwithstanding the relationships between the latent variable and the observed variables, there are relationships between the observed variables, then a local dependence model has to be considered. Within a capture-recapture framework with information available from three sampling occasions, this latent model is not identified. As such, the preferred solution is to rely on covariate information collected about the individuals to ensure identifiability and to improve the estimates of the population (Pollock (2002)).

The crucial part of any latent class analysis lies in the interpretation of the latent variable. In fact, in censuses there have been two interpretations, which lead to very different and conflicting population estimates. In the first instance, the latent classes represent enumeration difficulty, and so the estimate of the population total is the sum of the latent groups. In the second instance, they represent enumeration error, and as such the the total population is only those who are deemed to be real enumerations; any erroneous enumerations need to be removed (Brown, Biemer and Judson (2006)). The decision as how to interpret the latent classes after analysis can sometimes be challenging. However, in the context of this paper it is clear that neither class could relate to erroneous enumeration as it would imply either no true enumerations are missing from all lists (i.e. n_{000g1}) for latent class 1, or being erroneous when counted on all lists (i.e. n_{111g2}) for latent class 2. More generally, in the example shown in paper it is clear that latent class analysis can provide valuable insight into how to create an efficient post-stratification so as to produce unbiased estimates of the population.

We have demonstrated how the latent class approach could be applied to a census setting using data from the US census. In the original observed data, fitting log-linear models to the incomplete contingency tables and then using the capture recapture techniques to estimate the missing cells were shown to lead to inconclusive estimates of the population size. The reason for this indecision is due

to there being a difference between the 'saturated' model with no three-way interaction term and the best-fitting model. When there is some reason to believe that the conditional odds ratios between any two variables might differ across the different categories of the third variable, the estimates of the missing under different log-linear models might not be very reliable. Here the latent class model is beneficial in showing where the post-stratification scheme has not been adequate.

While capture-recapture techniques are increasingly being used in censuses, and through the log-linear modelling framework, they have an appealing representation to allow for dependency among the lists and heterogeneity in the population, they still rely on fairly strong assumptions, which cannot be tested from the data of the study. When there are two lists, there is the assumption of independence, and when there are three lists, there is the assumption of no three-way interaction. There is an enormous list of authors that have proposed innovative ways of coping with the failure of the independence assumption under dual-system estimation. Less is known about handling the failure of the no-three-way assumption under triple system estimation. Our paper has shown some insights into the advantages of a latwnt class model over a simple log-linear model, and the merits of using covariate information to cope with heterogeneity and dependence. The obvious extension is to move to four-lists (i.e. quadruple system estimation): there is the assumption here that the highest order interaction term is set to zero, implying there is no four-way-interaction. In fact, with a larger number of captures, this assumption is more likely to be correct (Fienberg (1972); International Working Group for Disease Monitoring and Forecasting (1995)) but four lists are rarely available in populations, hence the importance of including covariates.

References

Agresti, A. (1994). Simple capture-recapture models permitting unequal catchability and variable sampling effort. *Biometrics 50*(2), 494–500.

Alho, J. (1990). Logistic regression in capture-recapture models. *Biometrics 46*(3), 623–635.

Alho, J., M. Mulry, K. Wurdeman, and J. Kim (1993). Estimating heterogeneity in the probabilities of enumeration for dual-system estimation. *Journal of the American Statistical Association 88*(423), 1–130.

Bartlett, M. (1935). Contingency table interactions. *Journal of the Royal Statistical Society, Supplement 2*, 248–252.

Bartolucci, F. and A. Forcina (2006). A class of latent marginal models for capture-recapture data with continuous covariates. *Journal of the American Statistical Association 101*(474), 786–794.

Biemer, P., H. Woltmann, D. Raglin, and J. Hill (2001). Enumeration Accuracy in a Population Census: An Evaluation Using Latent Class Analysis. *Journal of Official Statistics 17*(1), 129–148.

Biggeri, A., F. Merletti, and M. Marchi (1999). Latent class models for varying catchability and correlation among sources in capture-recapture estimation. *Statistica Applicata 11*, 563–573.

Bishop, Y., S. Fienberg, P. Holland, J. Richard, and F. Mosteller (1975). *Discrete multivariate analysis: theory and practice*. MIT Press, Cambridge, Massachusetts.

Brown, G., P. Biemer, and D. Judson (2006). Estimating Erroneous Enumeration in the US Decennial Census using Four Lists. *Proceedings of the American Statistical Association, Section on Survey Research Methods*.

Brown, J., O. Abbott, and I. Diamond (2006). Dependence in the 2001 one-number census project. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 169*(4), 883–902.

Brown, J., I. Diamond, R. Chambers, L. Buckner, and A. Teague (1999). A methodological strategy for a one-number census in the UK. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 162*(2), 247–267.

Bruno, G., E. LaPorte, F. Merletti, A. Biggeri, D. McCarthy, and G. Pagano (1994). National diabetes programs: Application of capture-recapture to count diabetes? *Diabetes Care 17*(6), 548–556.

Buckland, S. and P. Garthwaite (1991). Quantifying precision of mark-recapture estimates using the bootstrap and related methods. *Biometrics 47*(1), 255–268.

Chandrasekar, C. and W. Deming (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association 44*(245), 101–115.

Chao, A. (2001). An overview of closed capture-recapture models. *Journal of Agricultural, Biological and Environmental Statistics 6*(2), 158–175.

Chen, S., C. Tang, and V. Mule (2010). Local post-stratification in dual system accuracy and coverage evaluation for the U.S. census. *Journal of the American Statistical Association 405*(489), 105–119.

Cormack, R. (1992). Interval estimation for mark-recapture studies of closed populations. *Biometrics 48*(2), 567–576.

Cormack, R. M. (1972). The logic of capture-recapture estimates. *Biometrics 28*(2), 337–343.

Coull, B. and A. Agresti (1999). The Use of Mixed Logit Models to Reflect Heterogeneity in Capture-Recapture Studies. *Biometrics 55*(1), 294–301.

Darroch, J. (1958). The Multiple-Recapture Census I. Estimation of a Closed Population. *Biometrika 45*(3-4), 343–359.

Darroch, J., S. Fienberg, G. Glonek, and B. Junker (1993). A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability. *Journal of the American Statistical Association 88*(423), 1137–1148.

Deming, W. and F. Stephan (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics 11*(4), 427–444.

Dempster, A., N. Laird, and D. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 39*(1), 1–38.

Di Cecco, D., M. Di Zio, D. Filipponi, and I. Rocchetti (2018). Population Size Estimation Using Multiple Incomplete Lists with Overcoverage. *Journal of Official Statistics 34*(2), 557 – 572.

Fienberg, S. (1972). The multiple recapture census for closed populations and incomplete 2k contingency tables. *Biometrika 59*(3), 591–603.

Fienberg, S., M. Johnson, and B. Junker (1999). Classical multilevel and Bayesian approaches to population size estimation using multiple lists. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 162*(3), 383–405.

Formann, A. (2003). Latent Class Model Diagnosis from a Frequentist Point of View. *Biometrics 59*(1), 189–196.

Gerritse, S. C., P. G. van der Heijden, and B. F. Bakker (2015). Sensitivity of Population Size Estimation for Violating Parametric Assumptions in Log-linear Models. *Journal of Official Statistics 31*(3), 357–379.

Goodman, L. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika 61*(2), 215–231.

Haberman, S. (1979). *Analysis of Qualitative Data: Vol. 2. New Developments*. Academic Press, New York.

Hagenaars, J. (1993). *Loglinear Models With Latent Variables*. Sage (University Paper Series).

Hogan, H. (1992). The 1990 Post-Enumeration Survey: An Overview. *American Statistician 46*, 261–269.

International Working Group for Disease Monitoring and Forecasting (1995). Capture-recapture and multiple-record systems estimations I: history and theoretical development. *American Journal of Official Epidemiology 142*, 1047–1058.

Isaki, C. and L. Schultz (1986). Dual system estimation using demographic analysis data. *Journal of Official Statistics 2*(2), 169–179.

Lazarsfeld, P. and N. Henry (1968). *Latent Structure Analysis*. Houghton, Mifflin.

Meng, X. and D. Rubin (1991). Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm. *Journal of the American Statistical Association 86*(416), 899–909.

Pollock, K. (2002). The use of auxiliary variables in capture-recapture modelling: an overview. *Journal of Applied Statistics 29*(1), 85–102.

Seber, G. (1986). A review of estimating animal abundance. *Biometrics 42*(2), 267–292.

Stanghellini, E. and P. van der Heijden (2004). A multiple-record systems estimation method that take observed and unobserved heterogeneity into account. *Biometrics 60*(2), 510–516.

Wang, Y. and J. Thandrayen (2009). Mutliple-Record Systems Estimation using Latent Class Models. *Australian and New Zealand Journal of Statistics 51*(1), 101–111.

Wolter, K. (1986). Some Coverage Error Models for Census Data. *Journal of the American Statistical Association 81*(394), 338–346.

Wolter, K. (1990). Capture-Recapture Estimation in the Presence of a Known Sex Ratio. *Biometrics 46*(1), 157–162.

Zaslavsky, A. and G. Wolfgang (1990). Triple System Modeling of Census, Post-Enumeration Survey and Administrative List Data. *Proceedings of the American Statistical Association, Section on Survey Research Methods*.

Zaslavsky, A. and G. Wolfgang (1993). Triple-System Modeling of Census, Post-Enumeration Survey, and Administrative-List Data. *Journal of Business and Economic Statistics 11*, 279–288.

Zhang, L.-C. (2019). A note on dual system population size estimator. *Journal of Official Statistics 35*(1), 279–282.

A EM algorithm - latent class modelling (with bootstrapped confidence intervals)

```
##
### Code to run the bootstrap
##

## This code has been updated to run the bootstrap for the models without the covariate

## We create 1000 bootstrap resamples

### Due to the fact we have missing data, simply using a standard bootstrap will not ensure that the
### resampled data are 'truly' from the underlying data distribution

set.seed(1234)

#
### Young Renters
#

#data <- array(c(NA,31,8,13,59,19,19,79), dim=c(2,2,2))
#dimnames(data) <- list(c("No", "Yes"), c("No", "Yes"), c("No", "Yes"))
#data <- data.frame(expand.grid(census = dimnames(data)[[1]],
#                               survey = dimnames(data)[[2]], admin = dimnames(data)[[3]]),
#                   count = c(data))



#
### Young Owners
#

#data <- array(c(NA,41,34,69,43,12,11,58), dim=c(2,2,2))
#dimnames(data) <- list(c("No", "Yes"), c("No", "Yes"), c("No", "Yes"))
#data <- data.frame(expand.grid(census = dimnames(data)[[1]],
#                               survey = dimnames(data)[[2]], admin = dimnames(data)[[3]]),
#                   count = c(data))



#
### Old Renters
#



#data <- array(c(NA,62,10,36,35,13,10,91), dim=c(2,2,2))
#dimnames(data) <- list(c("No", "Yes"), c("No", "Yes"), c("No", "Yes"))
#data <- data.frame(expand.grid(census = dimnames(data)[[1]],
```

```
#                                survey = dimnames(data)[[2]], admin = dimnames(data)[[3]]),
#                    count = c(data))



#
### Old Owners
#

data <- array(c(NA,32,24,69,43,7,13,72), dim=c(2,2,2))
dimnames(data) <- list(c("No", "Yes"), c("No", "Yes"), c("No", "Yes"))
data <- data.frame(expand.grid(census = dimnames(data)[[1]],
survey = dimnames(data)[[2]], admin = dimnames(data)[[3]]),
count = c(data))


### Initialise the data
data$em.data <- data$count


####################
###### MODELS    ###
####################


## I: Model I (independence)
#eqn <- em.data~census+survey+admin

## II: Model II - L, CS
#eqn <- em.data~census+survey+admin+census:survey

## III: Model III: S, CL
#eqn <- em.data~census+survey+admin+census:admin

## IV: Model IV: C, SL
#eqn <- em.data~census+survey+admin+survey:admin

## V: Model VI: CS, CL
#eqn <- em.data~census+survey+admin+census:survey+census:admin

## VI: Model V: CS, SL
eqn <- em.data~census+survey+admin+census:survey+survey:admin

## VII: Model VI: CL, SL
eqn <- em.data~census+survey+admin+census:admin+survey:admin
```

```
#### Begin the modelling and bootstrap code

#NA index
ii=is.na(data$em.data)

## initialise data (with replacement of missing values)
data$em.data[ii] <- 0

#only use non-NA, so that error should be less extreme
model <- glm(eqn, data = data[!ii,], family = poisson)


#Residual based bootstrap, using log difference

fit=fitted(model)
error=numeric(length(ii))

error[!ii]=(log(data$count)[!ii]-log(fit))

EM.sim=function(error,fit,data,ii,tol,eqn){


#simulate new data with residuals
data$em.data[!ii]=data$count[!ii]=exp(log(fit)+error[!ii])

model <- glm(eqn, data = data, family = poisson)

est <- 1
est <- cbind(est,model$coef)
fit <- model$fitted
## this does the same thing
#fit <- exp(model.matrix(model)%*%model$coef)

## E step
data$em.data[ii] <- fit[ii]

i <- 2
while(any(c(abs(est[, i] - est[, i - 1])) > tol,na.rm=T))
{
model <- glm(eqn, data = data, family = poisson)
est <- cbind(est,model$coef)
## M step
fit <- model$fitted
data$em.data[ii] <- fit[ii]
i <- i + 1
}
est<-est
```

```
fit<-fit
fit
}


#EM.sim(err,fit,data,ii,1e-5,eqn)

##
## It is better to use a for loop (more efficient)
##

#for loop style
## first create matrix
tmp=matrix(nrow=1000,ncol=length(ii))

for (i in 1:1000) {
err=numeric(length(ii))
err[!ii]=error[!ii][sample(1:sum(!ii),sum(!ii),T)]
tmp[i,]=EM.sim(err,fit,data,ii,1e-5,eqn)
}


#############################################################################################
### Collection of the bootstrap results to produce the estimates and precision estimates   ###
#############################################################################################

##### Computation of the bootstrap means, standard errors and confidence intervals

### bootstrap mean
round(apply(tmp,2,mean),2)

#bootstrap sd
round(apply(tmp,2,sd),2)

## quantiles (to compute the 95% confidence intervals)
round(apply(tmp, 2, quantile, probs = c(0.025,0.975)), 2)
```