Statistics Surveys
Vol. 18 (2024) 1–81
ISSN: 1935-7516
https://doi.org/10.1214/23-SS146

Bayesian dual systems population estimation for small domains*

Patrick Graham[1], Lucianne Varn[2], Matthew Hendtlass[1], Rebecca Green[1], and Andrew Richens[1]

[1] *Methods and Design, Statistics New Zealand*
e-mail: patrick.graham.br@gmail.com
matthew.hendtlass@stats.govt.nz; rebecca.green@stats.govt.nz
andrew.richens@stats.govt.nz

[2] *Population Insights, Statistics New Zealand*
e-mail: me@luciannevarn.org

Abstract: Dual systems estimation is a capture–recapture approach to population estimation restricted to two captures of the population. When applied to human populations, the population captures may be existing incomplete listings of the population, as provided by census and administrative datasets. In contrast to much capture-recapture analysis, dual systems applications in official statistics are usually concerned with estimating the distribution of a human population, over combinations of covariates, such as age, sex, ethnic group and small geographic area, rather than focussing primarily on the total population count. We synthesise theory and methods for Bayesian dual systems estimation for this problem, which we refer to as small domain population estimation. We primarily work within a model framework that combines a general model for the covariate distribution, such as an unrestricted multinomial in the case of categorical covariates, with hierarchical logistic models for the probability of inclusion on the lists. We explore the issue of dependence between the two lists which leads to a well-known identifiability problem. We illustrate the use of informative priors for the degree of dependence expressed as an odds ratio, or for certain aggregate level population totals. Although progress can be made using informative priors, the underlying identifiability problem means that inferences are sensitive to prior choices. We relate our approach to the popular log-linear modelling approach to capture-recapture analysis and note that the latter is primarily a re-parameterisation of our model set-up, though with a specific implied prior for the total population in the case of Poisson log-linear models.

MSC2020 subject classifications: Primary 62F15; secondary 62-02.
Keywords and phrases: Population estimation, Bayesian inference, small domain estimation.

Received February 2023.

Contents

*The views expressed in this paper are those of the authors and should not be taken as representing an official viewpoint of Statistics New Zealand. We thank editors and reviewers for their careful review of this paper.

Notation

N	Total population size.
Y	List inclusion cell indicator, taking values in $\mathcal{Y} = \{(1,1),(1,0),(0,1),(0,0)\}$.
$\mathbf{X}$	Covariate vector, taking values $\mathbf{x} \in \mathcal{X}$.
L_j	List inclusion indicator for List j, $j \in \{1,2\}$.
$\mathbf{Y}^{\text{com}}$	Complete population vector of list inclusion cells, with realised value $\mathbf{y}^{\text{com}}$.

$\mathbf{X}^{\text{com}}$	Complete population covariate matrix, with realised value $\mathbf{x}^{com}$; For q covariates $\mathbf{X}^{\text{com}}$ is an $N \times q$ random matrix and $\mathbf{x}^{com}$ is an $N \times q$ matrix.
n_{obs}	Number of people recorded on at least one list.
$\mathbf{x}^{\text{obs}}$	Observed covariate values.
$\mathbf{y}^{\text{obs}}$	Observed list inclusion cells.
$\mathbf{D}^{\text{obs}}$	Observed data.
$\mathbf{X}_{(0,0)}$	An $(N-n_{\text{obs}}) \times q$ matrix of covariate values for the group missed by both lists.
$\mathbf{x}_{[k]}$	The k^{th} covariate combination in a aggregated population structure.
K	Number of covariate combinations in an aggregated population structure.
$M_{k,y}$	Count of people with covariate level $\mathbf{x}_{[k]}$ in list inclusion cell y.
$M_{k,+}$	Population count for the k^{th} covariate combination; $\sum_y M_{k,y}$.
$M_{k,+}^{\text{obs}}$	Observable count of people recorded on at least one list at the k^{th} covariate combination; $\sum_{y \neq (0,0)} M_{k,y}$.
$\mathbf{M}_{.,y}$	Vector of counts for list inclusion cell y; $(M_{1,y}, M_{2,y}, \ldots, M_{K,y})'$.
$\mathbf{M}^{\text{com}}$	$\{M_{k,y},\, k \in \{1, \ldots, K\},\, y \in \mathcal{Y}\}$. Counts for the complete population, by covariate combination and list inclusion cell.
$\boldsymbol{\beta}_j$	Parameter vector for probability model for inclusion in List j, $j \in \{1, 2\}$.
$\boldsymbol{\beta}$	Full parameter vector for both coverage models; $(\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$.
$\rho(\mathbf{x})$	List inclusion dependence odds ratio for covariate setting $\mathbf{x}$.
$\boldsymbol{\rho}$	Vector of dependence odds ratio parameters; may be a vector of odds ratios for different groups or covariate combinations or a vector of parameters of a model relating odds ratios to covariates.
$p_{(0,0)}(\boldsymbol{\beta}, \boldsymbol{\theta})$	Population averaged probability of being missed by both lists, for a model assuming conditionally independent list inclusion.
$p_{(0,0)}(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\rho})$	Population averaged probability of being missed by both lists, for a model which allows conditionally dependent list inclusion, with dependence parameterised by the odds ratio parameters, $\boldsymbol{\rho}$.
$\tilde{\phi}_j(\mathbf{x}, \boldsymbol{\beta}_j)$	List inclusion probability for List j; $\Pr(L_j = 1) \vert \mathbf{X} = \mathbf{x}, \boldsymbol{\beta}_j)$.
$\phi_y(\mathbf{x}, \boldsymbol{\beta})$	Probability of being in list inclusion cell y, given covariate combination $\mathbf{X} = \mathbf{x}$, assuming conditionally independent list inclusion; $\Pr(Y = y \vert \mathbf{X} = \mathbf{x}, \boldsymbol{\beta})$.
$\boldsymbol{\phi}(\mathbf{x}, \boldsymbol{\beta})$	Vector of cell inclusion probabilities; $(\phi_{(1,1)}(\mathbf{x}, \boldsymbol{\beta}), \phi_{(1,0)}(\mathbf{x}, \boldsymbol{\beta}), \phi_{(0,1)}(\mathbf{x}, \boldsymbol{\beta}), \phi_{(0,0)}(\mathbf{x}, \boldsymbol{\beta}))'$.
$\phi_y(\mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\rho})$	Probability of being in list inclusion cell y, given covariate combination $\mathbf{x}$, for a model that allows conditionally dependent list inclusion with the dependence parameterised by the odds ratio parameters, $\boldsymbol{\rho}$.
$\boldsymbol{\theta}$	Parameter vector for covariate distribution.
$\boldsymbol{\eta}$; $\eta_{k,y}$	Vector of covariate combination and list inclusion cell probabilities in the multinomial log-linear model; Probability for covariate

	combination $\mathbf{x}_{[k]}$ and list inclusion cell y, in the multinomial log-linear model.
$\boldsymbol{\lambda}$	Parameter vector for $\log(\boldsymbol{\eta})$ in the multinomial log-linear model.
$\boldsymbol{\mu}$; $\mu_{k,y}$	Vector of expected cell counts, for the Poisson log-linear model; Expected count for list inclusion cell y for covariate combination $\mathbf{x}_{[k]}$.
$\boldsymbol{\xi}$	Parameter vector for $\log(\boldsymbol{\mu})$ in the Poisson log-linear model.
$\text{TNorm}(m, \sigma, a, b)$	Truncated normal distribution, with mean parameter m, standard deviation parameter σ, lower and upper truncation points a and b, respectively.
g subscript	Indicates a parameter or observable pertains to the g^{th} sub-population, as in $\boldsymbol{\theta}_g$ and $M_{g,k,y}$ which refer to the covariate distribution parameter vector for the g^{th} population and the count of people in sub-population g, with covariate combination $\mathbf{x}_{[k]}$ and in list inclusion cell y, respectively.

1. Introduction

Capture–recapture methods are widely used population estimation methods that use two or more listings of a population, in conjunction with statistical modelling, to estimate the size of the group not captured by any of the lists, and hence, the population size. The methods are used in ecology (Seber, 1982; Link and Barker, 2009, chapter 9), epidemiology (International Working Group for Disease Monitoring and Forecasting, 1995) and, increasingly in the estimation of the number of victims of human rights violations (Lum, Price and Banks, 2013; Manrique-Vallier, Price and Gohdes, 2013; Cruyff, van Dijk and van der Heijden, 2017; van Dijk, van der Heijden and Kragten-Heerdink, 2016; Silverman, 2020). Variants of the method are also commonly used by national statistics offices for estimating the size and distribution of human populations across dimensions such as age, sex, area and ethnic group. In this setting, the methods are commonly referred to as dual and multiple systems estimation and have, to date, generally been restricted to estimation of closed populations. Historically, official statistical applications of capture-recapture have concentrated on inferring the population from two sources (dual systems estimation), such as a census and a follow-up survey.

With the increasing availability of population listings based on administrative data, there is growing interest among official statisticians in the use of administrative lists for population estimation, either in combination with, or, as an alternative to, a traditional census (Statistics New Zealand, 2019). Although the use of administrative lists for population estimation increases the potential for multiple systems approaches to be applied, in practice the number of sources of administrative data that cover the full population age-range remain limited. For example, an administrative list sourced from education data could be expected to have good coverage of the young people but relatively poor coverage for older age groups. A list based on tax records may have good coverage of the working

age and older population but may have poor coverage of children. Thus, dual systems approaches remain of interest in official population estimation. While our set-up is primarily formulated for a data structure based on two large but incomplete population listings, such as provided by a traditional census dataset and an administrative list, or by two administrative lists, it also accommodates the census plus survey scenario, as we discuss in Appendix D of Supplementary Material (Graham et al., 2023).

Capture-recapture applications in official statistics differ from those in other areas because rather than estimation of the total population size being the primary focus, as is often the case in ecological, epidemiological and human rights applications, in official statistics it is usually necessary to estimate the distribution of the population across combinations of covariates, such as age, sex, ethnic group and location. These "small domain" population estimates are an important input to central and local government planning and resource allocation decisions. Our focus is on small domain population estimation when only two partial listings of the population are available. The problem is a missing data problem: if in addition to the number of people missed by both lists we could learn the covariate values for all such individuals, we would have complete covariate information for the population, and population estimation would amount to counting the number of people with each covariate combination of interest. Our estimation task can, therefore, be characterised as estimating the covariate distribution of the group missed by both available population lists.

We take a Bayesian approach to inference which is well-suited to the missing data characterisation of the problem and to small domain estimation. Small domain estimation often benefits from hierarchical modelling, which fits very naturally with the Bayesian approach to inference (see, for example, Gelman et al. (2014, chapter 5)).

While drawing on existing Bayesian and frequentist literature on population estimation, the paper is not intended as a comprehensive literature review and is more in the nature of a synthesis of important ideas in population estimation, viewed from a Bayesian perspective, with particular emphasis on the two-list situation. Where it seems helpful, we note connections between our approach and the frequentist literature on capture-recapture methods. However, we do not attempt formal comparisons with non-Bayesian methods. The paper is intended for those interested in exploring Bayesian theory and methods for small domain population estimation. There is already an extensive literature on frequentist inference for capture-recapture studies, though this is not generally focussed on small domain population estimation. Standard references for frequentist approaches to population estimation include Seber (1982) and International Working Group for Disease Monitoring and Forecasting (1995), emphasising ecological and epidemiological applications, respectively.

Although official statistics is the most obvious application area for Bayesian small domain dual systems estimation, applications in epidemiology and human rights are also possible. For example, if in addition to estimating the overall caseload for a particular disease, case numbers by age, sex, and small geographic area is an important consideration, the methods discussed in this paper are

relevant. One reason to consider geographic variation in case numbers is to help plan the provision of treatment services.

There have been several recent applications of Bayesian capture-recapture analysis in the human rights area (Manrique-Vallier, 2016; Manrique-Vallier, Ball and Sulmont, 2019; Sadinle, 2018; Silverman, 2020; Tuoto, Di Cecco and Tancredi, 2022). Although these applications utilise multiple population lists and focus on estimation of the total population, they illustrate the potential of Bayesian approaches to population estimation which automatically provide measures of uncertainty and provide a framework for incorporating prior information. Other work on Bayesian population estimation using capture-recapture methods includes George and Robert (1992); Madigan and York (1997); Fienberg, Johnson and Junker (1999); Tancredi and Liseo (2012); Tancredi, Steorts and Liseo (2020) and Di Cecco, Di Zio and Liseo (2020a). The early papers of George and Robert (1992) and Madigan and York (1997), concentrate on methods for multiple lists, with few covariates. Madigan and York (1997) adopt a graphical modelling approach to explore dependencies between lists. They emphasise direct evaluation of the posterior distribution, using marginal likelihood for the population size, obtained by integrating the likelihood over the prior distribution of other model parameters. However, Madigan and York (1997) note that Monte Carlo methods are likely to be necessary as the estimation problems become larger and more complex. Madigan and York (1997) also promote Bayesian model averaging, to combine inferences from multiple models.

George and Robert (1992) propose the Gibbs sampler as a computational framework for Bayesian population estimation from multiple lists, though in the simple case without covariates. In this case, the Gibbs sampler alternates between sampling from the conditional posterior for the population size given the list capture probabilities and sampling from the conditional list coverage probabilities given the population size. The Gibbs sampler is an appealing computational framework for population estimation because it is readily adapted to deal with problems such as missing data, measurement error and linkage error, that are likely to be encountered in practice, particularly as more use is made of administrative data. Several subsequent authors have adopted the Gibbs sampler as the basic computational engine for Bayesian population estimation. Fienberg, Johnson and Junker (1999) introduce random effects in the form of a Rasch model to allow for heterogeneity in capture probabilities over individuals, in the absence of covariates. Manrique-Vallier (2016) also adopt the Gibbs sampling framework to implement a Dirichlet Process mixture model for multiple systems estimation that allows for heterogeneity in capture probabilities that is not accounted for by covariates. They formulate the problem of population estimation from multiple incomplete lists using a missing data perspective that is similar to the approach taken in this paper. Both Fienberg, Johnson and Junker (1999) and Manrique-Vallier (2016), assume list inclusion is independent conditionally on latent individual level variables. The flexibility of Gibbs sampling for population estimation is exploited by Di Cecco, Di Zio and Liseo (2020a), who use Gibbs sampling in a multiple list problem with additional missing data due to some lists not operating in some sub-populations.

While we discuss the Gibbs sampler for population estimation, we also consider an alternative approach to obtaining the posterior predictive distribution of the covariate values for the group missed from both lists, made possible by recent advances in Bayesian computation.

Many of the above-cited references deal with inference from multiple lists and have a corresponding focus on modelling the dependence between inclusion on the lists. With our focus on the two-list case, there is less scope for modelling association between inclusion on the lists, and the modelling focus is on the association between list inclusion and covariates.

In estimation of human populations, record linkage of two or more lists replaces the physical capture or sighting methods of ecology as the mechanism for determining the number of lists that record a given individual. The papers by Tancredi and Liseo (2012) and Tancredi, Steorts and Liseo (2020) present a Bayesian approach to population estimation in which record linkage is dealt with by treating the links as unknowns, and, effectively, imputing a new linked data structure on each iteration of a Gibbs sampler. Though conceptually appealing, with large datasets the computational implications of re-establishing the linked dataset on each iteration of a Gibbs sampler are substantial. This approach also requires access to the identifying variables used to perform the record linkage, such as components of name and date of birth. For confidentiality reasons, these variables are usually not available to analysts using linked data for population estimation. Sadinle (2018) provides a potential solution by proposing an alternative, two-stage, approach to Bayesian population estimation in which, firstly, a Bayesian record linkage procedure draws linked data structures from the posterior for the true linkage structure, and secondly, population estimation is conducted for each linkage structure generated in the first stage. Sadinle (2018) uses the direct posterior computation method for total population size developed by Madigan and York (1997), but does not consider small domain population estimation.

We assume a more traditional approach to record linkage, whereby lists are linked by a process, typically involving a mix of deterministic and probabilistic linkage methods (Fellegi and Sunter, 1969), that result in a single linked dataset. We develop Bayesian theory and methods for dual systems population estimation assuming linkage error is absent. However, in practice, there is likely to be some error in the linkage process. While progress has been made on methods for adjusting for linkage error in frequentist dual systems estimation (Ding and Fienberg, 1994; Di Consiglio and Tuoto, 2015, 2018; de Wolf, van der Laan and Zult, 2019), we leave discussion of integrating linkage error adjustment with Bayesian dual systems estimation for future work.

The structure of the remainder of the paper is as follows. In Section 2, we introduce the basic structure of the problem and establish notation. In Sections 3 to 5, we set up the structure of Bayesian inference for dual systems estimation by formulating the dual systems likelihood with covariates (Section 3), discussing prior specifications (Section 4), and describing the computation of the posterior (Section 5). We present examples to illustrate and compare approaches to Bayesian small domain dual systems estimation in Section 6. In Section 7,

we review options for Bayesian dual systems estimation when inclusion on the two lists cannot be assumed conditionally independent given covariates. The latter is a standard assumption in dual systems estimation. In Section 8, we present a Bayesian view of another approach to population estimation that has been popular in frequentist applications: log-linear modelling. We relate the log-linear modelling approach to the approach developed in Sections 3 to 5. Finally, in Section 9, we briefly summarise the paper and suggest some priorities for further development of Bayesian small domain population estimation. In Supplementary Material (Graham et al., 2023), we present technical details of likelihood and other derivations, and in Appendix D of Supplementary Material (Graham et al., 2023), we show how our approach to Bayesian dual systems estimation accommodates the situation where one of the lists is an area based cluster sample survey.

In the interests of reproducibility we illustrate the methods with simulated data examples. An R package implementing the methodology in the paper is available from `https://github.com/lvarn/BDSE.git`.

2. Basic set-up and notation

We consider a population of N individuals, each with a set of attributes (covariates) $\mathbf{X}_i$, for $i \in \{1, \dots N\}$. N is an unknown that we seek to estimate, however, our overall objective is to estimate the distribution of the population across the covariate combinations. We let L_1 and L_2 denote indicators for inclusion on the two lists, and define the list coverage probabilities (i.e. probability of inclusion on the list given inclusion in the target population) as

$$\tilde{\phi}_1(\mathbf{x}, \boldsymbol{\beta}_1) = \Pr(L_1 = 1 | \mathbf{X} = \mathbf{x}, \boldsymbol{\beta}_1) \tag{1}$$

$$\tilde{\phi}_2(\mathbf{x}, \boldsymbol{\beta}_2) = \Pr(L_2 = 1 | \mathbf{X} = \mathbf{x}, \boldsymbol{\beta}_2). \tag{2}$$

Note that, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ will typically refer to parameter vectors of models that relate list inclusion to covariates, whereas $\tilde{\phi}_j(\mathbf{x}, \boldsymbol{\beta}_j)$, $j \in \{1, 2\}$, refers to the list inclusion probability for a particular setting of covariates, implied by the model parameterised by $\boldsymbol{\beta}_j$. We use similar notational conventions throughout the paper. We will model list inclusion using logistic models, so

$$\tilde{\phi}_j(\mathbf{x}, \boldsymbol{\beta}_j) = \Pr(L_j = 1 | \mathbf{X} = \mathbf{x}, \boldsymbol{\beta}_j) = \text{invlogit}(\beta_{j,0} + \mathbf{x}^T \boldsymbol{\beta}_{j,1}), \quad j \in \{1, 2\},$$

where invlogit(.) denotes the inverse logit function, and $\boldsymbol{\beta}_j = (\beta'_{j,0}, \boldsymbol{\beta}'_{j,1})'$, for $j \in \{1, 2\}$, and, where needed, we assume any interaction or non-linear terms required for modelling dependence of list inclusion on covariates are included in $\mathbf{X}$. We let $\boldsymbol{\beta} = (\boldsymbol{\beta}'_1, \boldsymbol{\beta}'_2)'$ denote the full vector of model parameters for both coverage models.

An individual in the target population may be included on both, one or neither of the two lists. We let Y denote the cell in the cross-classification of list inclusion indicators to which an individual belongs. Thus, Y can take the

values in $\mathcal{Y} = \{(1,1),(1,0),(0,1),(0,0)\}$, and we denote the probability of each of the list inclusion cells by

$$\phi_y(\mathbf{x},\boldsymbol{\beta}) = \Pr(Y = y|\mathbf{X} = \mathbf{x},\boldsymbol{\beta}), \quad y \in \mathcal{Y}.$$

Note that, in general, the list inclusion cell probabilities may depend on other parameters, in addition to the parameters of the list coverage models, however, until Section 7, we make the assumption of conditional independence of list inclusion, given covariates, that is often invoked in dual systems estimation. This implies the list inclusion cell probabilities can be obtained directly from the list coverage models as follows:

$$\phi_{(1,1)}(\mathbf{x},\boldsymbol{\beta}) = \tilde{\phi}_1(\mathbf{x},\boldsymbol{\beta}_1)\ \tilde{\phi}_2(\mathbf{x},\boldsymbol{\beta}_2) \tag{3}$$

$$\phi_{(1,0)}(\mathbf{x},\boldsymbol{\beta}) = \tilde{\phi}_1(\mathbf{x},\boldsymbol{\beta}_1)\ (1-\tilde{\phi}_2(\mathbf{x},\boldsymbol{\beta}_2)) \tag{4}$$

$$\phi_{(0,1)}(\mathbf{x},\boldsymbol{\beta}) = (1-\tilde{\phi}_1(\mathbf{x},\boldsymbol{\beta}_1))\ \tilde{\phi}_2(\mathbf{x},\boldsymbol{\beta}_2) \tag{5}$$

$$\phi_{(0,0)}(\mathbf{x},\boldsymbol{\beta}) = (1-\tilde{\phi}_1(\mathbf{x},\boldsymbol{\beta}_1))\ (1-\tilde{\phi}_2(\mathbf{x},\boldsymbol{\beta}_2)). \tag{6}$$

This is illustrated in Table 1. We denote the full vector of list inclusion cell probabilities, corresponding to covariate combination $\mathbf{X} = \mathbf{x}$, by $\boldsymbol{\phi}(\mathbf{x},\boldsymbol{\beta}) = \big(\phi_{(1,1)}(\mathbf{x},\boldsymbol{\beta}),\phi_{(1,0)}(\mathbf{x},\boldsymbol{\beta}),\phi_{(0,1)}(\mathbf{x},\boldsymbol{\beta}),\phi_{(0,0)}(\mathbf{x},\boldsymbol{\beta})\big)'$. For an individual with covariates $\mathbf{X} = \mathbf{x}$, we model list inclusion cell Y as a categorical random variable with probability vector $\boldsymbol{\phi}(\mathbf{x},\boldsymbol{\beta})$, or, equivalently, as multinomial over the four possible cells, with size parameter equal to one and probabilities given by $\boldsymbol{\phi}(\mathbf{x},\boldsymbol{\beta})$. Hence,

$$[Y|\mathbf{X} = \mathbf{x},\boldsymbol{\beta}] \sim \text{Categorical}(\boldsymbol{\phi}(\mathbf{x},\boldsymbol{\beta})).$$

TABLE 1
Probability model for the population distribution at covariate setting $\mathbf{X} = \mathbf{x}$*, under the conditional independence assumption.*

		List 2	
		1	0
List 1	1	$\tilde{\phi}_1(\mathbf{x},\boldsymbol{\beta}_1)\ \tilde{\phi}_2(\mathbf{x},\boldsymbol{\beta}_2)$	$\tilde{\phi}_1(\mathbf{x},\boldsymbol{\beta}_1)\ (1-\tilde{\phi}_2(\mathbf{x},\boldsymbol{\beta}_2))$
	0	$(1-\tilde{\phi}_1(\mathbf{x},\boldsymbol{\beta}_1))\ \tilde{\phi}_2(\mathbf{x},\boldsymbol{\beta}_2)$	$(1-\tilde{\phi}_1(\mathbf{x},\boldsymbol{\beta}_1))\ (1-\tilde{\phi}_2(\mathbf{x},\boldsymbol{\beta}_2))$

Individuals with $Y = (0,0)$ do not appear in the observed data which is the union of the three cells $(1,1)$, $(1,0)$, and $(0,1)$. Thus, the list inclusion cell Y is both the outcome of interest and indicator for missing-ness. The dependence between cell location and missing-ness means the missing-ness mechanism cannot be ignorable (Rubin, 1976) and some care must be taken in setting up the likelihood function, as discussed below.

The assumption of conditional independence of list inclusion given covariates, $\mathbf{X}$, has two aspects. Firstly, it rules out direct causal dependence between inclusion on the two lists, such as would be the case if inclusion on List 1, directly affected the probability of inclusion on List 2. In ecological applications, an example of causal dependence is the phenomenon of animals becoming trap

shy following a negative experience of an initial capture. In the human context, an analogous situation would arise if a negative experience of participation in List 1 (e.g. a census) makes List 1 participants less willing to provide data to List 2 (e.g. an administrative list), due, for example to distrust of government engendered by the negative experience of participation in List 1. In situations with more than two lists modelling such behavioural responses to capture is possible (Pledger, 2000).

Secondly, the conditional independence assumption, requires that, conditional on included covariates, inclusion probabilities are homogeneous over individuals for at least one of the lists. To see this suppose that conditionally on covariates $\mathbf{X}$ and $\mathbf{Z}$, list inclusion is conditionally independent. That is

$$\begin{aligned}\Pr(Y=(l_1,l_2)|\mathbf{X}=\mathbf{x},\mathbf{Z}=\mathbf{z}) = \\ \Pr(L_1=l_1|\mathbf{X}=\mathbf{x},\mathbf{Z}=\mathbf{z})\Pr(L_2=l_2|\mathbf{X}=\mathbf{x},\mathbf{Z}=\mathbf{z}),\ \ \forall(l_1,l_2,\mathbf{x},\mathbf{z})\end{aligned} \tag{7}$$

where we have omitted dependence on model parameters, to save space. Now, suppose $\mathbf{Z}$ is unobserved so that estimation can condition only on the observed covariates $\mathbf{X}$. The joint list inclusion probabilities given only the observed covariates $\mathbf{X}$ are

$$\begin{aligned}&\Pr(Y=(l_1,l_2)|\mathbf{X}=\mathbf{x})\\ &=\sum_{\mathbf{z}}\Pr(L_1=l_1,L_2=l_2|\mathbf{X}=\mathbf{x},\mathbf{Z}=\mathbf{z})\Pr(\mathbf{Z}=\mathbf{z}|\mathbf{X}=\mathbf{x}),\ \forall(l_1,l_2,\mathbf{x})\\ &=\sum_{\mathbf{z}}\Pr(L_1=l_1|\mathbf{X}=\mathbf{x},\mathbf{Z}=\mathbf{z})\Pr(L_2=l_2|\mathbf{X}=\mathbf{x},\mathbf{Z}=\mathbf{z})\Pr(\mathbf{Z}=\mathbf{z}|\mathbf{X}=\mathbf{x}),\\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\forall(l_1,l_2,\mathbf{x}).\end{aligned} \tag{8}$$

Now suppose that, in fact, conditionally on $\mathbf{X}$, inclusion on one of the lists, List 1 say, does not depend on $\mathbf{Z}$ so that, within levels of $\mathbf{X}$, inclusion probabilities for List 1 are homogeneous with respect to $\mathbf{Z}$. From (7) this implies

$$\begin{aligned}&\Pr(Y=(l_1,l_2)|\mathbf{X}=\mathbf{x})\\ &=\sum_{\mathbf{z}}\Pr(L_1=l_1|\mathbf{X}=\mathbf{x})\Pr(L_2=l_2|\mathbf{X}=\mathbf{x},\mathbf{Z}=\mathbf{z})\Pr(\mathbf{Z}=\mathbf{z}|\mathbf{X}=\mathbf{x}),\ \forall(l_1,l_2,\mathbf{x})\\ &=\Pr(L_1=l_1|\mathbf{X}=\mathbf{x})\sum_{\mathbf{z}}\Pr(L_2=l_2|\mathbf{X}=\mathbf{x},\mathbf{Z}=\mathbf{z})\Pr(\mathbf{Z}=\mathbf{z}|\mathbf{X}=\mathbf{x}),\ \forall(l_1,l_2,\mathbf{x})\\ &=\Pr(L_1=l_1|\mathbf{X}=\mathbf{x})\Pr(L_2=l_2|\mathbf{X}=\mathbf{x}),\ \forall(l_1,l_2,\mathbf{x})\end{aligned} \tag{9}$$

That is, given our initial assumption of conditional independence of list inclusion given $\mathbf{X}$ and $\mathbf{Z}$, we have shown the further assumption of homogeneity of inclusion probabilities, with respect to $\mathbf{Z}$, within levels of $\mathbf{X}$, for just one of the lists, implies list inclusion is conditionally independent given only $\mathbf{X}$. The argument also holds if inclusion probabilities are homogeneous, with respect to $\mathbf{Z}$, within levels of $\mathbf{X}$, for both lists. Conversely, if inclusion probabilities vary

with $\mathbf{Z}$, within levels of $\mathbf{X}$ for both lists, then (8) implies that list inclusion is not conditionally independent given only $\mathbf{X}$. It follows that if inclusion probability depends on $\mathbf{Z}$ within levels of $\mathbf{X}$ for both lists, $\mathbf{Z}$ must be included in the the analysis in order for the conditional independence assumption to hold. A good discussion of the two aspects of the conditional independence assumption can be found in International Working Group for Disease Monitoring and Forecasting (1995).

In multiple list problems, several authors relax the assumption of homogeneous capture probabilities by introducing latent variables to account for unobserved heterogeneity (Pledger, 2000; Fienberg, Johnson and Junker, 1999; Manrique-Vallier, 2016; Silverman, 2020). With two lists, there is less scope for such modelling, due to the intrinsic identifiability issues related to the group missed by both lists (i.e. the $(0,0)$ group) being unobservable. Consequently, in this paper, we generally rely on conditioning on observed covariates to control heterogeneity of list inclusion probability, although some possibilities for relaxing the conditional independence assumption using informative priors are considered in Section 7.

In addition to the assumption of conditional independence of list inclusion, we make the other standard assumptions of dual systems estimation, notably that the population is closed over the period spanned by the enumeration dates for the two lists (usually achieved by making the nominal enumeration date identical for the two lists), the lists are linked without error and there is no over-coverage on either list. For simplicity, we also assume that covariates are recorded consistently on the two lists. Consequently inclusion on either list is sufficient for individual's covariate values to be observed. Note that, we assume the covariates are defined for all individuals in the population, whether or not they are included on one or more of the lists.

3. Population data generating model and observed data likelihood

3.1. Unit-record population data structure

Given the total population size, N, we can imagine the population dataset comprising N records each containing the covariate values for an individual in the population. We call this the unit-record population structure or dataset. Appended to each record is the list inclusion cell Y. Thus, if q columns are required to describe the covariate values, the unit-record population data set is an $N \times (q+1)$ matrix. We assume the unit-record population covariate structure is generated as N independent draws from some q-dimensional covariate distribution, $H(\boldsymbol{\theta})$, and, conditionally on the covariates, $\mathbf{X} = \mathbf{x}$, the list inclusion cell Y is drawn, independently over individuals, as a categorical random variate with probability vector $\boldsymbol{\phi}(\mathbf{x}, \boldsymbol{\beta})$, that satisfies (3)–(6).

Given N and the coverage and covariate distribution model parameters $(\boldsymbol{\beta}, \boldsymbol{\theta})$, the data generating model we adopt is

$$[\mathbf{X}_i | \boldsymbol{\beta}, \boldsymbol{\theta}, N] \overset{\text{indep}}{\sim} H(\boldsymbol{\theta}), \quad i \in \{1, \ldots, N\}, \tag{10}$$

$$[Y_i|\mathbf{X}_i = \mathbf{x}_i, \boldsymbol{\beta}, \boldsymbol{\theta}, N] \overset{\text{indep}}{\sim} \text{Categorical}(\boldsymbol{\phi}(\mathbf{x}_i, \boldsymbol{\beta})), \quad i \in \{1, \ldots, N\} \tag{11}$$

or, equivalently,

$$\begin{aligned} p(\mathbf{y}^{\text{com}}, \mathbf{x}^{\text{com}}|N, \boldsymbol{\beta}, \boldsymbol{\theta}) &= \prod_{i=1}^{N} p(y_i|\mathbf{x}_i, \boldsymbol{\beta})\ p(\mathbf{x}_i|\theta) \\ &= \left(\prod_{i:y_i \neq (0,0)} \phi_{y_i}(\mathbf{x}_i, \boldsymbol{\beta})\ p(\mathbf{x}_i|\boldsymbol{\theta}) \right) \left(\prod_{i:y_i=(0,0)} \phi_{(0,0)}(\mathbf{x}_i, \boldsymbol{\beta})\ p(\mathbf{x}_i|\boldsymbol{\theta}) \right) \end{aligned} \tag{12}$$

where $\mathbf{x}^{\text{com}}$ is a realisation of the complete $N \times q$ population covariate matrix, $\mathbf{X}^{\text{com}}$, and $\mathbf{y}^{\text{com}} = (y_1, \ldots, y_N)$ is a realisation of the *complete* population vector of list inclusion cells, $\mathbf{Y}^{\text{com}}$. The i^{th} element of $\mathbf{Y}^{\text{com}}$ pertains to the same individual as the i^{th} row of $\mathbf{X}^{\text{com}}$, with a similar correspondence holding between the realised values $\mathbf{y}^{\text{com}}$ and $\mathbf{x}^{\text{com}}$. The products in (12) refer to subsets of the target population satisfying the condition indicated in the subscript. We follow similar notational conventions subsequently. In (12), we write $p(\mathbf{x}|\boldsymbol{\theta})$ for the joint covariate probability density or probability mass function evaluated at the realised value $\mathbf{x}$ with the understanding that the covariates are i.i.d draws from $H(\boldsymbol{\theta})$. In general, $\mathbf{X}$ may include both discrete and continuous covariates. Note that, the covariate distribution depends on $\boldsymbol{\theta}$ but not $\boldsymbol{\beta}$, while the conditional distribution for cell locations, given the covariates, depends on $\boldsymbol{\beta}$, but not $\boldsymbol{\theta}$. The model defined by (10) and (11) is a model for the complete data that would be observed if it were, somehow, possible to observe the covariate values for the group that is, in fact, missed by both lists. Since these covariate values cannot be observed, the likelihood for the model parameters is not given by the complete-data likelihood (12), but is, instead, obtained by integrating the complete data likelihood (12) over the unobserved covariate values for the group with $y = (0, 0)$.

Let $\mathbf{x}^{\text{obs}}$ denote the observable rows of $\mathbf{x}^{\text{com}}$, that is, the rows of $\mathbf{x}^{\text{com}}$ corresponding to elements of $\mathbf{y}^{\text{com}}$ with $y \neq (0, 0)$. Similarly, let $\mathbf{y}^{\text{obs}}$ denote the elements of $\mathbf{y}^{\text{com}}$ that are not equal to $(0, 0)$. Thus, if $\mathbf{y}^{\text{com}} = ((1, 1), (0, 1), (0, 0), (1, 0))'$, $\mathbf{y}^{\text{obs}} = ((1, 1), (0, 1), (1, 0))'$ and $\mathbf{x}^{\text{obs}}$ is the matrix comprising rows $1, 2$ and 4 of $\mathbf{x}^{\text{com}}$. The observed data is denoted by $\mathbf{D}^{\text{obs}} = (\mathbf{y}^{\text{obs}}, \mathbf{x}^{\text{obs}})$, and the likelihood by $p(\mathbf{D}^{\text{obs}}|N, \boldsymbol{\beta}, \boldsymbol{\theta})$. We let n_{obs} denote the number of individuals recorded in the observed data, so $\mathbf{y}^{\text{obs}}$ is a vector of length n_{obs} and $\mathbf{x}^{\text{obs}}$ is an $n_{\text{obs}} \times q$ matrix. Conditional on N, there are $N_{(0,0)} = N - n_{\text{obs}}$ individuals in the target population that are not recorded in the observed data. However, given only the observed data, it is unknown which rows in $\mathbf{x}^{\text{com}}$ are occupied by the covariate values for the $N - n_{\text{obs}}$ unrecorded individuals, and the covariate values for this group are unknown. We use the notation $\mathbf{x}_{(0,0)}$ to refer to the unobserved covariate values for the group missed by both lists.

In Appendix A of Supplementary Material (Graham et al., 2023), we show that the observed-data likelihood (i.e. the likelihood for the model parameters,

based on the data actually observed) is

$$p(\mathbf{D}^{\text{obs}}|N,\boldsymbol{\theta},\boldsymbol{\beta}) = \binom{N}{N-n_{\text{obs}}}\, p_{(0,0)}(\boldsymbol{\beta},\boldsymbol{\theta})^{N-n_{\text{obs}}} \prod_{i:y_i\neq(0,0)} \phi_{y_i}(\mathbf{x}_i,\boldsymbol{\beta})\; p(\mathbf{x}_i|\boldsymbol{\theta}) \tag{13}$$

$$\propto \frac{N!}{(N-n_{\text{obs}})!}\, p_{(0,0)}(\boldsymbol{\beta},\boldsymbol{\theta})^{N-n_{\text{obs}}} \prod_{i:y_i\neq(0,0)} \phi_{y_i}(\mathbf{x}_i,\boldsymbol{\beta})\; p(\mathbf{x}_i|\boldsymbol{\theta}), \tag{14}$$

where $p_{(0,0)}(\boldsymbol{\beta},\boldsymbol{\theta})$ is the population-averaged probability of being missed by both lists, defined formally as:

$$\begin{aligned} p_{(0,0)}(\boldsymbol{\beta},\boldsymbol{\theta}) &= \int \phi_{(0,0)}(\mathbf{x},\boldsymbol{\beta})\; p(\mathbf{x}|\theta)\,\mathrm{d}\mathbf{x} \\ &= \int (1-\tilde{\phi}_1(\mathbf{x},\boldsymbol{\beta}_1))\;(1-\tilde{\phi}_2(\mathbf{x},\boldsymbol{\beta}_2))\; p(\mathbf{x}|\boldsymbol{\theta})\,\mathrm{d}\mathbf{x}, \end{aligned} \tag{15}$$

assuming conditionally independent list inclusion.

In the simple case without covariates, (14) reduces to the standard capture-recapture likelihood used by several other authors (Fienberg, 1972; Pledger, 2000; Link and Barker, 2009), though simplified to the two list setting. Accommodating covariates in the dual systems estimation likelihood is a non-trivial extension to the no covariate case because of the need, firstly, to consider the covariate distribution and, secondly, to integrate over the distribution of the covariates for the unobserved component of the population. We note that our individual-level model described by (10) and (11), does not imply multinomial sampling of the population into the four possible cell locations because the dependence of cell probabilities on covariates implies the list inclusion cell probabilities vary over individuals. Consequently, our complete-data likelihood (12) differs from that adopted by some other authors (e.g. see King et al. (2016, Section 2)).

From (13), the observed data likelihood can be written as

$$p(\mathbf{D}^{\text{obs}}|N,\boldsymbol{\theta},\boldsymbol{\beta}) = \left[\binom{N}{N-n_{\text{obs}}}(1-p_{(0,0)}(\boldsymbol{\beta},\boldsymbol{\theta})^{n_{\text{obs}}} p_{(0,0)}(\boldsymbol{\beta},\boldsymbol{\theta})^{N-n_{\text{obs}}}\right] \left[\prod_{i:y_i\neq(0,0)} \frac{\phi_{y_i}(\mathbf{x}_i,\boldsymbol{\beta})p(\mathbf{x}_i|\boldsymbol{\theta})}{1-p_{(0,0)}(\boldsymbol{\beta},\boldsymbol{\theta})}\right]. \tag{16}$$

The first term in (16) is a binomial probability for observing n_{obs} people given a total population size N and probability of observation $(1-p_{(0,0)(\boldsymbol{\beta},\boldsymbol{\theta})})$, and the second term is the conditional likelihood for $(\boldsymbol{\beta},\boldsymbol{\theta})$ which is constructed from the likelihood contributions of the observed records, explicitly conditioned on the fact of being observed. More formally the conditional likelihood can be defined as

$$L_C(\mathbf{D}^{\text{obs}}|\boldsymbol{\beta},\boldsymbol{\theta}) = \prod_{i:y_i\neq(0,0)} p(y_i,\mathbf{x}_i|Y_i\neq(0,0),\boldsymbol{\beta},\boldsymbol{\theta})$$

$$= \prod_{i:y_i \neq (0,0)} p(y_i|Y_i \neq (0,0), \mathbf{X}_i = \mathbf{x}_i, \boldsymbol{\beta}, \boldsymbol{\theta})\; p(\mathbf{x}_i|Y_i \neq (0,0), \boldsymbol{\beta}, \boldsymbol{\theta})$$

$$= \left(\prod_{i:y_i \neq (0,0)} \frac{\phi_{y_i}(\mathbf{x}_i, \boldsymbol{\beta})}{\left(1 - \phi_{(0,0)}(\mathbf{x}_i, \boldsymbol{\beta})\right)} \right) \left(\prod_{i:y_i \neq (0,0)} \frac{\left(1 - \phi_{(0,0)}(\mathbf{x}_i, \boldsymbol{\beta})\right)\; p(\mathbf{x}_i|\boldsymbol{\theta})}{(1 - p_{(0,0)}(\boldsymbol{\beta}, \boldsymbol{\theta}))} \right) \tag{17}$$

$$= \prod_{i:y_i \neq (0,0)} \frac{\phi_{y_i}(\mathbf{x}_i, \boldsymbol{\beta})\; p(\mathbf{x}_i|\boldsymbol{\theta})}{1 - p_{(0,0)}(\boldsymbol{\beta}, \boldsymbol{\theta})} \tag{18}$$

which is the second term in (16). The decomposition of capture-recapture likelihoods into a binomial probability for the number of individuals recorded on at least one list and the conditional likelihood based only on the data for recorded individuals has been noted elsewhere (Sandland and Cormack, 1984; Huggins and Hwang, 2011). Since N appears only in the binomial probability component of (16), it seems clear that information in the data concerning the total population size is concentrated in n_{obs}. However, n_{obs} is only informative regarding total populations size in conjunction with the other model parameters, through $p_{(0,0)}(\boldsymbol{\beta}, \boldsymbol{\theta})$. The structure of (16) suggests the information concerning the list coverage model and covariate distribution parameters is contained primarily in the conditional likelihood, since these parameters contribute to the binomial probability for n_{obs} only through $p_{(0,0)}(\boldsymbol{\beta}, \boldsymbol{\theta})$.

A version of the conditional likelihood has proven useful in frequentist capture-recapture analyses, because maximising the conditional likelihood yields close approximations to the maximum likelihood estimates for coverage model parameter estimates (Sanathanan, 1972; Huggins and Hwang, 2011; Cormack and Jupp, 1991). Estimates of the population size then follow as the number of individuals observed in the data multiplied by the reciprocal of the conditional maximum likelihood estimate of the probability of inclusion in the data. This basic strategy has been extended to accommodate coverage probabilities that vary with covariates by Huggins (1989) and Alho (1990), by maximizing the first product in (17),

$$\tilde{L}_c(\mathbf{D}^{\text{obs}}|\boldsymbol{\beta}) = \prod_{i:y_i \neq (0,0)} \frac{\phi_{y_i}(\mathbf{x}_i, \boldsymbol{\beta})}{\left(1 - \phi_{(0,0)}(\mathbf{x}_i, \boldsymbol{\beta})\right)}, \tag{19}$$

which arises from conditioning on covariate values as well as the event of being recorded on at least one list. However, the emphasis in these papers remains on the estimation of the total population size rather than small domain estimation.

It seems a reasonable conjecture that the partial conditional likelihood, $\tilde{L}_c(\mathbf{D}^{\text{obs}}|\boldsymbol{\beta})$, is a good approximation to the likelihood for $\boldsymbol{\beta}$. However, in contrast to the approach adopted in this paper, which uses the full likelihood (14) or the complete conditional likelihood (18), frequentist applications based on the partial conditional likelihood (19) do not permit inference on the covariate distribution parameters. Although inference for these parameters is generally of lesser interest than the coverage model parameters and the population estimates, in our approach they play an important role in determining the posterior

predictive distribution of the covariates for the group missed by both lists, as we discuss in Section 5.3. In addition, although not considered in this paper, modelling the covariate distribution is useful for extending dual systems estimation to deal with issues such as covariate missing-ness and measurement error.

The decomposition of the likelihood into a binomial probability for n_{obs} and a conditional likelihood that explicitly conditions on the event of being observed is suggestive of an alternative data generating model in which the number of people to be recorded on at least one list is first generated from a binomial distribution with size parameter N, and the n_{obs} people to be observed are then distributed over covariate and list inclusion cell (excluding the $(0,0)$ cell), under a model that is appropriately conditioned on the event of being observed. In Appendix A of Supplementary Material (Graham et al., 2023), we present an alternative derivation of the likelihood, (14), or, equivalently, (16), motivated by this conceptual data-generating model.

3.2. Aggregated population data structure

If covariates are restricted to be categorical, the population can be aggregated to counts at the level of covariate combinations. Thus, instead of a unit-record structure, if there are K unique covariate combinations, the population dataset can be viewed as a $K \times (q+4)$ matrix with the first q columns recording the covariate values that label the rows and the final four columns recording counts in the $(1,1),(1,0),(0,1)$ and $(0,0)$ cells, for each covariate combination. The counts in the final four columns sum to N. However, the counts in the $(0,0)$ column are not observed for any covariate combination. We let $\mathbf{x}_{[k]}$ denote the k^{th} covariate combination, so for $y \in \mathcal{Y}$, $\phi_y(\mathbf{x}_{[k]}, \boldsymbol{\beta}) = \Pr(Y = y | \mathbf{X} = \mathbf{x}_{[k]}, \boldsymbol{\beta})$ denotes the probability of list inclusion cell y, for the k^{th} covariate combination. We let $\boldsymbol{\phi}(\mathbf{x}_{[k]}, \boldsymbol{\beta}) = (\phi_{(1,1)}(\mathbf{x}_{[k]}, \boldsymbol{\beta}), \phi_{(1,0)}(\mathbf{x}_{[k]}, \boldsymbol{\beta}), \phi_{(0,1)}(\mathbf{x}_{[k]}, \boldsymbol{\beta}), \phi_{(0,0)}(\mathbf{x}_{[k]}, \boldsymbol{\beta}))'$ denote the vector of list inclusion probabilities for the k^{th} covariate combination. The probability of the k^{th} covariate combination is denoted θ_k, and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)'$ is the K-vector of probabilities for the covariate combinations, so $\Pr(\mathbf{X} = \mathbf{x}_{[k]} | \boldsymbol{\theta}) = \theta_k$. The population-averaged probability of being missed by both lists is $p_{(0,0)}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \sum_k \phi_{(0,0)}(\mathbf{x}_{[k]})\, \theta_k$. We let $M_{k,y}$ denote the number of people in the population with covariate level $\mathbf{x}_{[k]}$ and list inclusion cell y, $M_{k,+} = \sum_y M_{k,y}$, the total population count at the k^{th} covariate combination, $M^{\text{obs}}_{k,+} = \sum_{y \neq (0,0)} M_{k,y}$ the observable count of people recorded on at least one list at the k^{th} covariate combination, and $\mathbf{M}_{.,y} = (M_{1,y}, M_{2,y}, \ldots, M_{K,y})'$ the K-vector of counts for list inclusion cell y. The unobserved counts are $\mathbf{M}_{.,(0,0)}$, and we note that $n_{\text{obs}} = \sum_k M^{\text{obs}}_{k,+} = \sum_{k, y \neq (0,0)} M_{k,y}$, and $(N - n_{\text{obs}}) = \sum_k M_{k,(0,0)} = N - \sum_{k, y \neq (0,0)} M_{k,y} = N - \sum_k M^{\text{obs}}_{k,+}$. The observable data is $\mathbf{D}^{\text{obs}} = (\mathbf{M}_{.,(1,1)}, \mathbf{M}_{.,(1,0)}, \mathbf{M}_{.,(0,1)})$.

If a Multinomial$(N, \boldsymbol{\theta})$ model is adopted for the counts by covariate combination, and a second multinomial model is adopted for the allocation of covariate combination counts to the four possible combinations of list inclusion indicators

at each covariate combination, we have the model:

$$[(M_{1,+},\ldots,M_{K,+})|N,\boldsymbol{\theta}] \sim \text{Multinomial}(N,\boldsymbol{\theta}) \tag{20}$$

$$[(M_{k,(1,1)},M_{k,(1,0)},M_{k,(0,1)},M_{k,(0,0)})|M_{k,+},\beta] \overset{\text{indep}}{\sim} \text{Multinomial}(M_{k,+},\boldsymbol{\phi}(\mathbf{x}_{[k]},\boldsymbol{\beta})),\ k\in\{1,\ldots,K\} \tag{21}$$

In this model, it is clear from (21) that there is multinomial sampling of individuals into the list inclusion cells, within levels of the covariates. In Appendix A.2 of Supplementary Material (Graham et al., 2023), we show that the observed-data likelihood that follows from the model given by (20) and (21) is

$$p(\mathbf{D}^{\text{obs}}|N,\boldsymbol{\beta},\boldsymbol{\theta}) = \frac{N!}{(N-n_{\text{obs}})!}\left[\frac{\prod_{k,y\neq(0,0)}\left(\phi_y(\mathbf{x}_{[k]},\boldsymbol{\beta})\,\theta_k\right)^{M_{k,y}}}{\prod_{k,y\neq(0,0)}M_{k,y}!}\right] p_{(0,0)}(\boldsymbol{\beta},\boldsymbol{\theta})^{(N-n_{\text{obs}})} \tag{22}$$

$$\propto \frac{N!}{(N-n_{\text{obs}})!}\left[\prod_{k,y\neq(0,0)}\left(\phi_y(\mathbf{x}_{[k]},\boldsymbol{\beta})\,\theta_k\right)^{M_{k,y}}\right] p_{(0,0)}(\boldsymbol{\beta},\boldsymbol{\theta})^{(N-n_{\text{obs}})}, \tag{23}$$

which, by comparison with (14), is an aggregated data version of the likelihood derived under the individual-level model.

From (22) it is easily seen that the likelihood can also be written as:

$$p(\mathbf{D}^{\text{obs}}|N,\boldsymbol{\beta},\boldsymbol{\theta}) \propto \binom{N}{n_{\text{obs}}}(1-p_{(0,0)}(\boldsymbol{\beta},\boldsymbol{\theta}))^{n_{\text{obs}}}\ p_{(0,0)}(\boldsymbol{\beta},\boldsymbol{\theta})^{(N-n_{\text{obs}})}\times \left[\frac{n_{\text{obs}}!}{\prod_{k,y\neq(0,0)}M_{k,y}!}\prod_{k,y\neq(0,0)}\left(\frac{\phi_y(\mathbf{x}_{[k]},\boldsymbol{\beta})\,\theta_k}{(1-p_{(0,0)}(\boldsymbol{\beta},\boldsymbol{\theta})}\right)^{M_{k,y}}\right] \tag{24}$$

$$\propto \binom{N}{n_{\text{obs}}}(1-p_{(0,0)}(\boldsymbol{\beta},\boldsymbol{\theta}))^{n_{\text{obs}}}p_{(0,0)}(\boldsymbol{\beta},\boldsymbol{\theta})^{(N-n_{\text{obs}})}\prod_{k,y\neq(0,0)}\left(\frac{\phi_y(\mathbf{x}_{[k]},\boldsymbol{\beta})\,\theta_k}{(1-p_{(0,0)}(\boldsymbol{\beta},\boldsymbol{\theta})}\right)^{M_{k,y}} \tag{25}$$

which is the product of the Binomial probability for n_{obs}, and the conditional likelihood for $(\boldsymbol{\beta},\boldsymbol{\theta})$, in keeping with the decomposition of the likelihood for the unit-record population structure given in (16).

4. Prior specification

The parameters of our population model are the population size, N, the coverage model parameters, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1',\boldsymbol{\beta}_2')'$, and the parameters of the covariate distribution, $\boldsymbol{\theta}$. We assume *a priori* independence for parameter blocks, that is $p(N,\boldsymbol{\beta},\boldsymbol{\theta}) = p(N)\ p(\boldsymbol{\beta})\ p(\boldsymbol{\theta})$.

One option for the prior for the total population size is a discrete uniform, which requires only specification of prior limits on the possible population size. In official statistics applications with good information on births, deaths and migration, updating of historical population estimates through straightforward demographic accounting should often provide a good basis for setting plausible bounds on the total population size. Other possibilities for setting informative priors for the total population size include the Poisson and Negative Binomial distributions. However, because these distributions imply that variance increases with the expectation, they may yield a prior variance that does not co-coincide with prior uncertainty about the population size. Another possibility is the Conway-Maxwell-Poisson distribution (Shmueli et al., 2005), which can model under-dispersion as well as over-dispersion with respect to the Poisson distribution.

A traditional and popular choice of prior for N, intended to be uninformative, is the Jeffreys' prior: $p(N) \propto 1/N$. Di Cecco (2019) suggests Rissanen's "universal prior for the integers" (Rissanen, 1983) as an alternative uninformative prior. This prior has the form $p(N) \propto 2^{\log^*(N)}$, where $\log^*(N)$ is the sum of positive terms in the sequence $\{\log_2(N), \log_2(\log_2(N)), \ldots\}$.

We model the parameters of the coverage models for Lists 1 and 2 independently, and so, adopt the prior specification $p(\boldsymbol{\beta}) = p(\boldsymbol{\beta}_1)p(\boldsymbol{\beta}_2)$. In realistic applications, some components of the parameter vectors for the list coverage models such as small area effects, may be modelled hierarchically, meaning that they are modelled as draws from a distribution with parameters that are themselves treated as unknowns that are assigned priors and estimated in the posterior computation, along with all other parameters. The use of hierarchical coverage models is illustrated in Section 6. However, regardless of the structure of the coverage models, setting priors for parameters of logistic regression models is a standard task in Bayesian statistics (Gelman et al., 2008).

For the general covariate case, including both continuous and categorical covariates, we have, so far, left the covariate distribution unstructured. While there are clearly a variety of modelling options, the categorical distribution over a set of possible covariate combinations provides a very general model. Under this model, the N covariates vectors are assumed to be sampled independently from a Categorical($\boldsymbol{\theta}$) distribution, or, equivalently, a Multinomial($1, \boldsymbol{\theta}$) distribution, over the number of covariate combinations considered possible. If this model is adopted, the conjugate Dirichlet distribution is a standard and convenient choice of prior for $\boldsymbol{\theta}$. In the presence of continuous covariates or categorical covariates with a large number of possible categories (e.g. small area geography, single year of age), it may be necessary to restrict the possible combinations in some way. One such restriction is to use only the covariate combinations represented in the observed data. If this restriction is adopted and an improper Dirichlet prior with parameters set uniformly to zero is assumed, the model for the covariate distribution is similar to that underpinning the Bayesian bootstrap (Rubin, 1981). However, while potentially helpful in restricting the number of possible covariate combinations, the restriction to observed covariate combinations is not necessary and may be unduly restrictive. For example, it may be

prudent to allow all ages within a range, to occur in all geographic areas, even though, in some areas no individuals are recorded for some ages. For small areas and smaller age-groups (e.g. older ages) it is possible that both lists have failed to record individuals at particular ages, even though they are present in the population. The set of allowed covariate combinations can be extended beyond those represented in the observed data, by giving positive prior probability to unobserved covariate combinations. This can be achieved by using proper Dirichlet distributions defined over a set of observed and unobserved covariate combinations that are deemed possible.

In the case of categorical covariates and a aggregated data structure, a Dirichlet prior is also the natural choice of prior for $\boldsymbol{\theta}$ in the multinomial model (20). In the examples considered in this paper, we use proper but weak Dirichlet prior distributions for the parameters of the covariate distribution.

5. Bayesian inference for the target population

Obtaining counts for population subgroups (e.g. counts by age, sex, ethnic group and area), or other statistics for the target population, would be straightforward if the number of individuals in the target population missed by both lists and the covariate values for this group were known. For a unit-record population structure, let $\mathbf{X}_{(0,0)}$ denote the $(N - n_{\text{obs}}) \times q$ matrix of unobserved covariate values for the group missed by both lists. Inference for target population counts and other population statistics follows from the posterior predictive distribution for $\mathbf{X}_{(0,0)}$. For the aggregate population structure, inference for target population counts follows from the posterior predictive distribution for the unobserved counts for the (0,0) cells, by covariate combination, $\mathbf{M}_{.,(0,0)}$.

For the unit-record population structure, appending a draw from the posterior predictive distribution of $\mathbf{X}_{(0,0)}$ to the observed covariate matrix $\mathbf{x}^{\text{obs}}$, and, for completeness, appending $(N - n_{\text{obs}})$ $(0, 0)$ entries to the observed vector of cell inclusion indicators, $\mathbf{y}^{\text{obs}}$, produces a realisation of the complete data, $\mathbf{D}^{\text{com}} = (\mathbf{Y}^{\text{com}}, \mathbf{x}^{\text{obs}}, \mathbf{X}_{(0,0)})$. For each simulated completed dataset, tabulations and other analyses can be conducted; repeating this for some number of simulations, and storing results, builds a sample from the posterior predictive distribution for the quantities of interest. This stored sample of results can be used for inference. For example, 95% credible intervals can be straightforwardly approximated by locating the 2.5% and 97.5% quantiles of the stored distribution for each quantity of interest (e.g. population counts). These intervals represent uncertainty due to the missing covariate information for the group missed by both lists and from estimation of the model parameters, including the total population size.

Both the number of people missed by both lists and the values of the covariates in $\mathbf{X}_{(0,0)}$ are unknown. Consequently, rather than computing the posterior predictive distribution of the missing data directly as $p(\mathbf{X}_{(0,0)}|\mathbf{D}^{\text{obs}})$, it is convenient to concentrate on the joint posterior distribution of the missing data

and the total population size:

$$p(\mathbf{X}_{(0,0)}, N|\mathbf{D}^{\text{obs}}) = \int\int p(\mathbf{X}_{(0,0)}, N, \boldsymbol{\beta}, \boldsymbol{\theta}|\mathbf{D}^{\text{obs}})\ \mathrm{d}\boldsymbol{\beta}\ \mathrm{d}\boldsymbol{\theta}$$

$$= \int\int p(\mathbf{X}_{(0,0)}|N, \boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{D}^{\text{obs}})\ p(N, \boldsymbol{\beta}, \boldsymbol{\theta}|\mathbf{D}^{\text{obs}})\ \mathrm{d}\boldsymbol{\beta}\ \mathrm{d}\boldsymbol{\theta} \tag{26}$$

$$= \int\int p(\mathbf{X}_{(0,0)}|N, \boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{D}^{\text{obs}})\ p(N|\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{D}^{\text{obs}})\ p(\boldsymbol{\beta}, \boldsymbol{\theta}|\mathbf{D}^{\text{obs}})\ \mathrm{d}\boldsymbol{\beta}\ \mathrm{d}\boldsymbol{\theta}. \tag{27}$$

Assuming we can sample from the joint posterior for the coverage model and covariate distribution parameters, $p(\boldsymbol{\beta}, \boldsymbol{\theta}|\mathbf{D}^{\text{obs}})$, we could simulate the joint posterior $p(\mathbf{X}_{(0,0)}, N|\mathbf{D}^{\text{obs}})$ by simulating the components of the integrand (27), proceeding from right to left, as described in Algorithm 1. (In Algorithm 1, and subsequently, numerical superscripts enclosed in parentheses indicate iteration number and not exponents.) Focusing on the generated values of $\mathbf{X}_{(0,0)}$ and N implicitly simulates the integration in (27), which is simply notation for marginalising the joint posterior for the unknowns $(\mathbf{X}_{(0,0)}, N, \boldsymbol{\beta}, \boldsymbol{\theta})$ with respect to $(\boldsymbol{\beta}, \boldsymbol{\theta})$.

Obtaining the joint posterior $p(N, \boldsymbol{\beta}, \boldsymbol{\theta}|\mathbf{D}^{\text{obs}})$, in (26), is, in general, difficult. However, by approximating the prior for N by a continuous distribution, we use the probabilistic programming language Stan (Stan Development Team, 2021) to obtain an approximation to the joint posterior for $(N, \boldsymbol{\beta}, \boldsymbol{\theta})$ for some of the examples considered in Sections 6 and 7. When employing the posterior decomposition in (26) to obtain the posterior predictive distribution for $\mathbf{X}_{(0,0)}$, Algorithm 1 is simplified to two steps: the first drawing values of $(N, \boldsymbol{\beta}, \boldsymbol{\theta})$ jointly from $p(N, \boldsymbol{\beta}, \boldsymbol{\theta}|\mathbf{D}^{\text{obs}})$ and the second drawing $\mathbf{X}_{(0,0)}$ from the conditional posterior predictive distribution, as in step (iii) of Algorithm 1.

Algorithm 1 Simulating the posterior predictive distribution for the covariate values for the group missed by both lists.

for t in $\{1, \ldots, T\}$ **do**
  step (i): draw $(\boldsymbol{\beta}^{(t)}, \boldsymbol{\theta}^{(t)})$ from $p(\boldsymbol{\beta}, \boldsymbol{\theta}|\mathbf{D}^{\text{obs}})$ using (28) in Section 5.1;
  step (ii): draw $n^{(t)}$ from $p(N|\boldsymbol{\beta}^{(t)}, \boldsymbol{\theta}^{(t)}, \mathbf{D}^{\text{obs}})$ using (33) in Section 5.2;
  step (iii): draw $\mathbf{x}_{(0,0)}^{(t)}$ from $p(\mathbf{X}_{(0,0)}|N = n^{(t)}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\theta}^{(t)})$ using (36) in Section 5.3;
  store $n^{(t)}, \mathbf{x}_{(0,0)}^{(t)}$
end for

In the case of a population data structure aggregated to covariate combinations, the missing data required to complete the population are the counts for the $(0, 0)$ cells by covariate combination, that is $\mathbf{M}_{\cdot,(0,0)}$. As noted above, inference for the target population then follows from the posterior predictive distribution of $\mathbf{M}_{\cdot,(0,0)}$, which can be obtained by replacing the individual covariate values $\mathbf{X}_{(0,0)}$ with $\mathbf{M}_{\cdot,(0,0)}$ in (27) and Algorithm 1. Assuming the data-generating

model given by (20) and (21), the conditional posterior predictive distribution for $\mathbf{M}_{\cdot,(0,0)}$ is a multinomial, defined in Section 5.3.

We discuss the components in the joint posterior distribution of the unknowns, namely, $p(\boldsymbol{\beta},\boldsymbol{\theta}|\mathbf{D}^{\text{obs}})$, $p(N|\boldsymbol{\beta},\boldsymbol{\theta},\mathbf{D}^{\text{obs}})$, and $p(\mathbf{X}_{(0,0)}|N,\boldsymbol{\beta},\boldsymbol{\theta},\mathbf{D}^{\text{obs}})$ or $p(\mathbf{M}_{\cdot,(0,0)}|N,\boldsymbol{\beta},\boldsymbol{\theta},\mathbf{D}^{\text{obs}})$, in Sections 5.1, 5.2, and 5.3, respectively, and in Section 5.4, we present a Gibbs sampler as an alternative approach to simulating the joint posterior of the unknowns.

5.1. *Computation of the posterior for the coverage model and covariate distribution parameters*

The first step in using (27) to generate the posterior predictive distribution of the covariates for the group missed by both lists is to draw from the posterior for the parameters of the coverage models and covariate distribution, $(\boldsymbol{\beta},\boldsymbol{\theta})$, which is the marginalisation of the joint posterior $p(N,\boldsymbol{\beta},\boldsymbol{\theta}|\mathbf{D}^{\text{obs}})$ over N:

$$\begin{aligned} p(\boldsymbol{\beta},\boldsymbol{\theta}|\mathbf{D}^{\text{obs}}) &= \sum_n p(N=n,\boldsymbol{\beta},\boldsymbol{\theta}|\mathbf{D}^{\text{obs}}) \\ &\propto p(\boldsymbol{\beta})\, p(\boldsymbol{\theta}) \sum_n p(\mathbf{D}^{\text{obs}}|N=n,\boldsymbol{\beta},\boldsymbol{\theta}) \Pr(N=n), \end{aligned} \tag{28}$$

since the parameters $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ and N are assumed to be a priori independent. Then, from (28) and (16), the marginal likelihood for $(\boldsymbol{\beta},\boldsymbol{\theta})$, is

$$\begin{aligned} &p(\mathbf{D}^{\text{obs}}|\boldsymbol{\beta},\boldsymbol{\theta}) \propto \sum_n p(\mathbf{D}^{\text{obs}}|N=n,\boldsymbol{\beta},\boldsymbol{\theta}) \Pr(N=n) \\ &= \prod_{i:y_i\neq(0,0)} \frac{\phi_{y_i}(\mathbf{x}_i,\boldsymbol{\beta})\, p(\mathbf{x}_i|\boldsymbol{\theta})}{(1-p_{(0,0)}(\boldsymbol{\beta},\boldsymbol{\theta}))} \times \\ &\sum_n \binom{n}{n-n_{\text{obs}}} (1-p_{(0,0)}(\boldsymbol{\beta},\boldsymbol{\theta}))^{n_{\text{obs}}} \left(p_{(0,0)}(\boldsymbol{\beta},\boldsymbol{\theta})\right)^{(n-n_{\text{obs}})} \Pr(N=n). \end{aligned} \tag{29}$$

In the special case of the Jeffreys' prior for N, we can write $\Pr(N=n) \propto 1/n$, and the marginal likelihood for $(\boldsymbol{\beta},\boldsymbol{\theta})$ becomes

$$\begin{aligned} &p(\mathbf{D}^{\text{obs}}|\boldsymbol{\beta},\boldsymbol{\theta}) \propto \\ &\begin{cases} \dfrac{1}{n_{\text{obs}}} \displaystyle\prod_{i:y_i\neq(0,0)} \dfrac{\phi_{y_i}(\mathbf{x}_i,\boldsymbol{\beta})\, p(\mathbf{x}_i|\boldsymbol{\theta})}{(1-p_{(0,0)}(\boldsymbol{\beta},\boldsymbol{\theta}))} \times \\ \displaystyle\sum_n \frac{(n-1)!}{(n_{\text{obs}}-1)!(n-n_{\text{obs}})!} (1-p_{(0,0)}(\boldsymbol{\beta},\boldsymbol{\theta}))^{n_{\text{obs}}} \left(p_{(0,0)}(\boldsymbol{\beta},\boldsymbol{\theta})\right)^{(n-n_{\text{obs}})} \end{cases} \end{aligned} \tag{30}$$

$$\begin{aligned} &= \frac{1}{n_{\text{obs}}} \prod_{i:y_i\neq(0,0)} \frac{\phi_{y_i}(\mathbf{x}_i,\boldsymbol{\beta})\, p(\mathbf{x}_i|\boldsymbol{\theta})}{(1-p_{(0,0)}(\boldsymbol{\beta},\boldsymbol{\theta}))} \\ &\propto \prod_{i:y_i\neq(0,0)} \frac{\phi_{y_i}(\mathbf{x}_i,\boldsymbol{\beta})\, p(\mathbf{x}_i|\boldsymbol{\theta})}{(1-p_{(0,0)}(\boldsymbol{\beta},\boldsymbol{\theta}))}, \end{aligned} \tag{31}$$

since the summand in (30) is the probability mass function for a Negative-Binomial $(n_{\text{obs}}, 1-p_{(0,0)}(\boldsymbol{\beta},\boldsymbol{\theta}))$ distribution, implying the summation reduces to one. The right hand side of (31) is the conditional likelihood for $(\boldsymbol{\beta},\boldsymbol{\theta})$ defined in (18). Thus, the marginal likelihood for $(\boldsymbol{\beta},\boldsymbol{\theta})$ obtained under the Jeffreys' prior for N is equivalent to the conditional likelihood for $(\boldsymbol{\beta},\boldsymbol{\theta})$.

For priors on N, other than the Jeffreys' prior, the complete conditional likelihood in (18) could be used to approximate the likelihood for $(\boldsymbol{\beta},\boldsymbol{\theta})$, defined in (29), in an otherwise fully Bayesian analysis. That is, we could approximate $p(\boldsymbol{\beta},\boldsymbol{\theta}|\mathbf{D}^{\text{obs}})$ by

$$p_C(\boldsymbol{\beta},\boldsymbol{\theta}|\mathbf{D}^{\text{obs}}) \propto p(\boldsymbol{\beta})\; p(\theta)\; L_C(\mathbf{D}^{\text{obs}}|\boldsymbol{\beta},\boldsymbol{\theta}) \tag{32}$$

and use $p_C(\boldsymbol{\beta},\boldsymbol{\theta}|\mathbf{D}^{\text{obs}})$ in (27). Clearly, the approximation is exact when the Jeffreys' prior is adopted for the population size, and, when other priors are adopted, is in error to the extent that the prior for N differs from the Jeffreys' prior.

In general, sampling from $p(\boldsymbol{\beta},\boldsymbol{\theta}|\mathbf{D}^{\text{obs}})$ is not straightforward, however we have implemented the approach based on the conditional likelihood in the Bayesian modelling language, Stan (Stan Development Team, 2021), and an example is given in Section 6.

5.2. Computation of the conditional posterior for the population size

The second step in generating the posterior predictive distribution of the covariates for the group missed by both lists is to draw from the conditional posterior of the total population size which, from (14), is given by

$$p(N|\boldsymbol{\theta},\boldsymbol{\beta},\mathbf{D}^{\text{obs}}) \propto p(N)\frac{N!}{(N-n_{\text{obs}})!}\left(p_{(0,0)}(\boldsymbol{\beta},\boldsymbol{\theta})\right)^{N-n_{\text{obs}}}. \tag{33}$$

For the improper Jeffreys' prior, $p(N) \propto 1/N$, the conditional posterior (33) is proportional to a Negative-Binomial probability mass function with parameters $(n_{\text{obs}}, 1-p_{(0,0)}(\boldsymbol{\beta},\boldsymbol{\theta}))$ (see Appendix B of Supplementary Material (Graham et al., 2023) for details). We note that choosing $p(N) \propto 1/N$ leads to the conditional posterior mean $E(N|\mathbf{D}^{\text{obs}},\boldsymbol{\theta},\boldsymbol{\beta}) = n_{\text{obs}}/(1-p_{(0,0)}(\boldsymbol{\beta},\boldsymbol{\theta}))$, which is also the MLE for N conditional on $(\boldsymbol{\beta},\boldsymbol{\theta})$. This is consistent with the notion of the Jeffreys' prior as a representation of lack of prior information.

For bounded priors on N, the right hand side of (33) can be obtained by direct evaluation for each value of N within the prior bounds, and the posterior for N can therefore be simulated by sampling directly from the discrete distribution implied by (33). Alternatively, rejection sampling or Metropolis-Hastings methods (Gelman et al., 2014, pp. 261–292) could be applied to obtain draws from (33), however, because the latter approach can take some time to converge to the target distribution, it may not prove practical to include a Metropolis-Hastings step to correct approximations to (33) in step (ii) of Algorithm 1.

5.3. Computation of the conditional posterior for the covariate values of the group missed by both lists

Given draws from the posterior for the coverage model and covariate distribution parameters and the population size, the final step in simulating the posterior predictive distribution of the covariate values for the group missed by both lists is to draw from the conditional posterior predictive distribution

$$p(\mathbf{X}_{(0,0)}|N,\boldsymbol{\beta},\boldsymbol{\theta},\mathbf{D}^{\text{obs}}) = \frac{p(\mathbf{X}_{(0,0)},\mathbf{D}^{\text{obs}}|N,\boldsymbol{\beta},\boldsymbol{\theta})}{p(\mathbf{D}^{\text{obs}}|N,\boldsymbol{\beta},\boldsymbol{\theta})}, \tag{34}$$

in the case of a unit-record population structure, or

$$p(\mathbf{M}_{\cdot,(0,0)}|N,\boldsymbol{\beta},\boldsymbol{\theta},\mathbf{D}^{\text{obs}}) = \frac{p(\mathbf{M}_{\cdot,(0,0)},\mathbf{D}^{\text{obs}}|N,\boldsymbol{\beta},\boldsymbol{\theta})}{p(\mathbf{D}^{\text{obs}}|N,\boldsymbol{\beta},\boldsymbol{\theta})} \tag{35}$$

in the case of an aggregate population structure. The denominator of (34) and (35) is the likelihood given by (13) in the unit-record case, and by (23) in the aggregated case, assuming the data generating model given by (20) and (21).

Straightforward evaluation of (34), given in Appendix C of Supplementary Material (Graham et al., 2023), shows that the conditional posterior predictive distribution of the covariate values for the group missed by both lists is given by

$$\Pr(\mathbf{X}_{(0,0)} = \mathbf{x}_{(0,0)}|N,\boldsymbol{\beta},\boldsymbol{\theta},\mathbf{D}^{\text{obs}}) = \prod_{i=1}^{(N-n_{\text{obs}})} \frac{\phi_{(0,0)}(\mathbf{x}_i,\boldsymbol{\beta})\; p(\mathbf{x}_i|\boldsymbol{\theta})}{p_{(0,0)}(\boldsymbol{\beta},\boldsymbol{\theta})} \tag{36}$$

where $p(\mathbf{x}_i|\boldsymbol{\theta}) = H_(\mathbf{x}_i|\boldsymbol{\theta})$, the population covariate distribution evaluated at $\mathbf{x}_i$. That is, conditional on N, the posterior predictive distribution for $\mathbf{X}_{(0,0)}$ is equivalent to $(N - n_{\text{obs}})$ independent draws from

$$p(\mathbf{X}|Y = (0,0),\boldsymbol{\beta},\boldsymbol{\theta}) \propto \phi_{(0,0)}(\mathbf{X},\boldsymbol{\beta})\; p(\mathbf{X}|\boldsymbol{\theta}),$$

which is a weighted version of the population covariate distribution where the weights are the probability of being missed by both lists for the specific covariate value. Therefore, the generation of the covariate values for the $Y = (0,0)$ group compensates for the selection bias inherent in the observed data covariate distribution, which by definition, under-represents covariate values predictive of being missed by both lists.

For an aggregate population structure, generated under the model given by (20) and (21), we show, in Appendix C of Supplementary Material (Graham et al., 2023), that the conditional posterior predictive distribution of the unobserved counts $\mathbf{M}_{\cdot,(0,0)}$ is given by

$$\begin{aligned}&\Pr(\mathbf{M}_{\cdot,(0,0)} = \mathbf{m}_{\cdot,(0,0)}|N,\boldsymbol{\beta},\boldsymbol{\theta},\mathbf{D}^{\text{obs}})\\ &\quad= \frac{(N-n_{\text{obs}})!}{\prod_k m_{k,(0,0)}!}\prod_k \left(\frac{\phi_{(0,0)}(\mathbf{x}_{[k]},\boldsymbol{\beta})\;\theta_k}{p_{(0,0)}(\boldsymbol{\beta},\boldsymbol{\theta})}\right)^{m_{k,(0,0)}}.\end{aligned} \tag{37}$$

Since $(N - n_{\text{obs}}) = \sum_k M_{k,(0,0)}$, and $p_{(0,0)}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \sum_k \phi_{(0,0)}(\mathbf{x}_{[k]})\, \theta_k$, (37) is the probability mass function of a K-category multinomial with size parameter $(N - n_{\text{obs}})$ and category probabilities $\phi_{(0,0)}(\mathbf{x}_{[k]}, \boldsymbol{\beta})\, \theta_k / p_{(0,0)}(\boldsymbol{\beta}, \boldsymbol{\theta})$, for $k \in \{1, \ldots K\}$.

5.4. The Gibbs sampler for small domain dual systems estimation

Gibbs sampling or data augmentation (Tanner and Wong, 1987; Gelman et al., 2014, pp. 275–292) provides an alternative to steps (i) to (iii) of Algorithm 1 for computing the joint posterior for the population size and covariate values for the group missed by both lists. The Gibbs sampler alternates between imputing the unobserved covariate values for the group missed by both lists conditional on the model parameters, using (33) and (36) (or (37) in the case of an aggregated population structure), and sampling from the posterior distribution of the parameters, conditional on the completed data formed by augmenting the observed data with imputed values of the missing data, $\mathbf{X}_{(0,0)}$ or $\mathbf{M}_{.,(0,0)}$. This approach to posterior computation extends readily to situations with complications such as measurement error and sporadic missing-ness of covariate values for individuals captured on at least one list, by adding additional imputation steps.

For small domain dual systems estimation, the unknowns are $(\mathbf{X}_{(0,0)}, N, \boldsymbol{\beta}, \boldsymbol{\theta})$, or, if the data is aggregated to covariate combinations, $(\mathbf{M}_{.,(0,0)}, N, \boldsymbol{\beta}, \boldsymbol{\theta})$. We describe the Gibbs sampler for the former case, but adaptation to the aggregated data case is immediate, using obvious substitutions. Rather than simulating the joint posterior $p(\mathbf{X}_{(0,0)}, N, \boldsymbol{\beta}, \boldsymbol{\theta} | \mathbf{D}^{\text{obs}})$ directly, using the decomposition in Algorithm 1, the Gibbs sampler proceeds by alternately sampling from the, full conditional posterior distributions: (i) $p(\mathbf{X}_{(0,0)}, N | \boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{D}^{\text{obs}})$; (ii) $p(\boldsymbol{\beta} | \mathbf{X}_{(0,0)}, N, \boldsymbol{\theta}, \mathbf{D}^{\text{obs}})$; (iii) $p(\boldsymbol{\theta} | \mathbf{X}_{(0,0)}, N, \boldsymbol{\beta}, \mathbf{D}^{\text{obs}})$. The Gibbs sampling algorithm for obtaining a Monte Carlo approximation to $p(\mathbf{X}_{(0,0)}, N, \boldsymbol{\phi}, \boldsymbol{\theta} | \mathbf{D}^{\text{obs}})$ is given in Algorithm 2. The algorithm describes the updating steps for a single Gibbs sampler chain, but in practice, multiple parallel chains are usually run to facilitate checking for convergence of the sampler (Gelman et al., 2014, pp. 281–288).

Note that, in step (i) of the Gibbs sampler in Algorithm 2, we update $(\mathbf{X}_{(0,0)}, N)$ jointly using the decomposition

$$p(\mathbf{X}_{(0,0)}, N | \boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{D}^{\text{obs}}) = p(\mathbf{X}_{(0,0)} | N, \boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{D}^{\text{obs}})\; p(N | \boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{D}^{\text{obs}}),$$

and so, make use of (36) and (33), rather than alternating between the full conditional distributions $p(N | \boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{X}_{(0,0)}, \mathbf{D}^{\text{obs}})$ and $p(\mathbf{X}_{(0,0)} | \boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{D}^{\text{obs}}, N)$. This is because, conditional on the observed data and $\mathbf{X}_{(0,0)}$, the total number of records in the population is known, implying that N is known. Consequently, simulating N conditionally on the observed data supplemented by $\mathbf{X}_{(0,0)}$ would mean N was fixed at its initially generated value. Fienberg, Johnson and Junker (1999) discuss a similar issue.

Updating the list coverage model parameters (step (ii), line 4 in Algorithm 2) and the covariate distribution parameters (step (iii), line 5) of the Gibbs sam-

Algorithm 2 A single Gibbs sampler chain for dual systems estimation.

1: Initialise $\boldsymbol{\beta} \to \boldsymbol{\beta}^{(0)}$, $\boldsymbol{\theta} \to \boldsymbol{\theta}^{(0)}$.
2: **for** t in $\{1, \ldots, T\}$ **do**
3: step (i): Draw $(\mathbf{x}_{(0,0)}^{(t)}, n^{(t)})$ by drawing
(a) $n^{(t)}$ from $p(N|\boldsymbol{\theta}^{(t-1)}, \boldsymbol{\beta}^{(t-1)}, \mathbf{D}^{\text{obs}})$) using (33) and
(b) $\mathbf{x}_{(0,0)}^{(t)}$ from $p(\mathbf{X}_{(0,0)}|N = n^{(t)}, \boldsymbol{\theta}^{(t-1)}, \boldsymbol{\beta}^{(t-1)}, \mathbf{D}^{\text{obs}})$ using (36);
Create an $(n^{(t)} - n_{\text{obs}})$ vector $\mathbf{y}_{(0,0)}^{(t)}$ with each element equal to $(0, 0)$;
Set

$$\mathbf{y}^{\text{com},(t)} = (\mathbf{y}^{obs,'}, \mathbf{y}_{(0,0)}^{(t),'})'$$
$$\mathbf{x}^{\text{com},(t)} = \begin{pmatrix} \mathbf{x}^{\text{obs}} \\ \mathbf{x}_{(0,0)}^{(t)} \end{pmatrix}$$
$$\mathbf{D}^{\text{com},(t)} = \left(\mathbf{y}^{\text{com},(t)}, \mathbf{x}^{\text{com},(t)}\right).$$

4: step (ii): Draw $\boldsymbol{\beta}^{(t)}$ from

$$p(\boldsymbol{\beta}|\boldsymbol{\theta}^{(t-1)}, n^{(t)}, \mathbf{D}^{\text{com},(t)}) = p(\boldsymbol{\beta}_1|n^{(t)}, \mathbf{D}^{\text{com},(t)})\ p(\boldsymbol{\beta}_2|n^{(t)}, \mathbf{D}^{\text{com},(t)})$$

using (39).
5: step (iii): Draw $\boldsymbol{\theta}^{(t)}$ from $p(\boldsymbol{\theta}|\boldsymbol{\beta}^{(t)}, n^{(t)}, \mathbf{D}^{\text{com},(t)})$.
6: Store $\mathbf{x}_{(0,0)}^{(t)}, n^{(t)}, \boldsymbol{\beta}^{(t)}$, and $\boldsymbol{\theta}^{(t)}$.
7: **end for**
8: Discard first B of T iterations as burn-in; use the remaining $T - B$ iterations for inference.

pling algorithm amount to reasonably standard Bayesian computations since they are conditional on the completed population data. Updating the coverage parameters in step (ii) decomposes into separate updates for List 1 and List 2 coverage model parameters, since, with $\mathbf{y}^{\text{com},(t)}$ and $\mathbf{x}^{\text{com},(t)}$ defined as in step (iii) of Algorithm 2, and recalling that $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$, we can write

$$\begin{aligned} p(\boldsymbol{\beta}|\boldsymbol{\theta}^{(t-1)}, n^{(t)}, &\mathbf{y}^{\text{com},(t)}, \mathbf{x}^{\text{com},(t)}) \propto p(\boldsymbol{\beta})\ p(\mathbf{y}^{\text{com},(t)}|\mathbf{x}^{\text{com},(t)}, N, \boldsymbol{\beta}) \\ &= p(\boldsymbol{\beta}_1)\ p(\boldsymbol{\beta}_2) \prod_{i:y_i^{\text{com},(t)}=(1,1)} \tilde{\phi}_1(\mathbf{x}_i^{\text{com},(t)}, \boldsymbol{\beta}_1)\ \tilde{\phi}_2(\mathbf{x}_i^{\text{com},(t)}, \boldsymbol{\beta}_2) \\ &\quad \times \prod_{i:y_i^{\text{com},(t)}=(1,0)} \tilde{\phi}_1(\mathbf{x}_i^{\text{com},(t)}, \boldsymbol{\beta}_1)\ (1 - \tilde{\phi}_2(\mathbf{x}_i^{\text{com},(t)}, \boldsymbol{\beta}_2)) \\ &\quad \times \prod_{i:y_i^{\text{com},(t)}=(0,1)} (1 - \tilde{\phi}_1(\mathbf{x}_i^{\text{com},(t)}, \boldsymbol{\beta}_1))\ \tilde{\phi}_2(\mathbf{x}_i^{\text{com},(t)}, \boldsymbol{\beta}_2) \\ &\quad \times \prod_{i:y_i^{\text{com},(t)}=(0,0)} (1 - \tilde{\phi}_1(\mathbf{x}_i^{\text{com},(t)}, \boldsymbol{\beta}_1))\ (1 - \tilde{\phi}_2(\mathbf{x}_i^{\text{com},(t)}, \boldsymbol{\beta}_2)). \end{aligned} \tag{38}$$

Rearranging (38) we have

$$p(\boldsymbol{\beta}|\boldsymbol{\theta}^{(t-1)}, n^{(t)}, \mathbf{y}^{\text{com},(t)}, \mathbf{x}^{\text{com},(t)}) =$$

$$\left[p(\boldsymbol{\beta}_1) \prod_{i:y_i^{\text{com},(t)} \in \{(1,1),(1,0)\}} \tilde{\phi}_1(\mathbf{x}_i^{\text{com},(t)}, \boldsymbol{\beta}_1) \times \prod_{i:y_i^{\text{com},(t)} \in \{(0,1),(0,0)\}} (1 - \tilde{\phi}_1(\mathbf{x}_i^{\text{com},(t)}, \boldsymbol{\beta}_1))\right] \times \left[p(\boldsymbol{\beta}_2) \prod_{i:y_i^{\text{com},(t)} \in \{(1,1),(0,1)\}} \tilde{\phi}_2(\mathbf{x}_i^{\text{com},(t)}, \boldsymbol{\beta}_2) \times \prod_{i:y_i^{\text{com},(t)} \in \{(1,0),(0,0)\}} (1 - \tilde{\phi}_2(\mathbf{x}_i^{\text{com},(t)}, \boldsymbol{\beta}_2))\right], \tag{39}$$

which clearly comprises distinct components for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, implying the posterior for $\boldsymbol{\beta}$ decomposes as the product of the posterior for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, as indicated in line 4 of Algorithm 2. Further details of the updates for the logistic coverage model parameters will depend on model specifications such as whether some parameters are modelled hierarchically. Since there is no conjugate prior for parameters of logistic regression models, draws from the conditional posterior for the coverage model parameters cannot be obtained directly and instead are usually drawn using a Metropolis-Hastings algorithm (Gelman et al., 2014, pp. 278–280). Nevertheless, posterior simulation for logistic regression models is a standard Bayesian inference problem. This is illustrated in the examples in Section 6.

Details of the updating step for the covariate distribution parameters will, of course, depend on the model adopted for the covariate distribution. If the covariate distribution is modelled as N independent draws from a categorical distribution with parameter $\boldsymbol{\theta}$, and a Dirichlet($\boldsymbol{\alpha}$) prior distribution is adopted for $\boldsymbol{\theta}$, the conditional posterior for $\boldsymbol{\theta}$ is also Dirichlet but with parameters updated to $\boldsymbol{\alpha} + (M_{1,+}^{(t)}, \ldots, M_{K,+}^{(t)})'$, at the t^{th} iteration of the sampler, where $M_{k,+}^{(t)}$ is obtained by adding the simulated count for the $(0,0)$ cell for the k^{th} covariate combination to the number of people observed with the k^{th} covariate combination, i.e. $M_{k,+}^{(t)} = M_{k,+}^{\text{obs}} + M_{k,(0,0)}^{(t)}$. The same holds for the aggregated population structure where counts by covariate combination are modelled by a multinomial as defined in (20).

6. Example: population estimation by single year of age, sex and geographic area

In order to illustrate the ideas and methods presented thus far, we consider an example based on simulated but realistic data. The example considered has a structure similar to problems faced by national statistical offices, who are charged with describing the structure of the total population by demographic variables, such as age, sex and area.

In Section 6.1, we outline the process of simulating realistic population data and describe the simulated dataset used for the demonstrations in this section. In Section 6.2, we estimate the population using the approaches discussed in

Section 5. In Section 6.3, we consider the issue of model misspecification and illustrate the use of posterior predictive model checking to help diagnose problems with model specification.

6.1. Creating the simulated populations

The simulated target populations are constructed in four steps:

- **(i)** For a target population of size N, N records with the desired covariates are drawn with replacement from a covariate distribution. The covariate distribution is based on the joint frequency distribution of demographic covariates (age, sex and area) in the 2013 New Zealand census. We set $N =$ 1,000,000. Each individual in the target population is then represented by their covariates vector;
- **(ii)** The coverage model for each of the two lists is specified by assigning population coverage patterns by age, sex and area. Logistic regression models were chosen for the coverage models, and therefore, this step amounts to specifying the form and coefficients of the linear component of the list inclusion probabilities $\tilde{\phi}_j(\mathbf{x}, \boldsymbol{\beta}_j) = \text{invlogit}(\mathbf{x}'\boldsymbol{\beta}_j)$, for $j \in \{1, 2\}$. The vector $\mathbf{x}$ includes predictor functions (such as interactions and basis functions). The exact models will be discussed below;
- **(iii)** The coverage models are then applied to the covariates of the N selected records in the target population to obtain list inclusion probabilities for each record, namely, $\tilde{\phi}_j(\mathbf{x}_i, \boldsymbol{\beta}_j)$, for $j \in \{1, 2\}$, and $i \in \{1, \ldots, N\}$;
- **(iv)** Finally, list inclusion indicators, for each record in the target population, are drawn independently from Bernoulli distributions with probability parameters set to the list inclusion probabilities obtained in step (iii): $[L_{i,j}|\mathbf{X}_i = \mathbf{x}_i, \boldsymbol{\beta}_j] \sim \text{Bernoulli}(\tilde{\phi}_j(\mathbf{x}_i, \boldsymbol{\beta}_j))$, $j \in \{1, 2\}$, and $i \in \{1, \ldots, N\}$.

The list inclusion cell, Y, and corresponding probability, $\phi_y(\mathbf{x}, \boldsymbol{\beta}), y \in \mathcal{Y}$, is then derived, for each record in the target population, from the generated list inclusion indicators and the corresponding marginal probabilities, according to (3)–(6).

The simulated observed data are obtained from the simulated target population by dropping all records with both list inclusion indicators equal to zero, that is, with list inclusion cell $(0, 0)$. For model fitting and analysis, the target population and the observed data are aggregated by unique covariate combinations.

In the examples considered in this section, the data is simulated using a clustered design, where individuals in the population are clustered within geographic areas. We use $a_i \in \{1, \ldots, A\}$ to denote the area for individual i, and $\check{\mathbf{x}}_i$ to denote their other covariates. Then, the full covariates vector for individual i becomes $\mathbf{x}_i = (a_i, \check{\mathbf{x}}_i')'$. The following data generating model is used for each individual, $i \in \{1, \ldots, N\}$, in the target population:

$$[\mathbf{X}_i|\boldsymbol{\theta}_{\text{census}}] \sim \text{Categorical}(\boldsymbol{\theta}_{\text{census}})$$

$$
\begin{aligned}
[L_{i,j}|\mathbf{X}_i = \mathbf{x}_i, \boldsymbol{\beta}_j] &\overset{\text{indep}}{\sim} \text{Bernoulli}\big(\tilde{\phi}_j(\mathbf{x}_i, \boldsymbol{\beta}_j)\big), \quad j \in \{1, 2\} \\
\tilde{\phi}_j(\mathbf{x}_i', \boldsymbol{\beta}_j) &= \text{invlogit}(\beta_{j,0,a_i} + \check{\mathbf{x}}_i' \, \boldsymbol{\beta}_{j,1}) \\
\beta_{j,0,a_i} &\overset{\text{indep}}{\sim} \text{N}(\beta_{j,0}, \sigma_j^2),
\end{aligned}
$$

where $\boldsymbol{\beta}_j = \big(\beta_{j,0,1}, \ldots, \beta_{j,0,A}, \boldsymbol{\beta}_{j,1}\big)'$ is the vector of parameters of the coverage model for List j, which includes random area effects for the A areas and the covariate coefficients vector.

We simulate a target population of size one million, spread across 11,929 covariate combinations. The covariate combinations range in size from 1 to 2,734, and the median of the covariate combination counts is 36; in fact, approximately 95% of the covariate combinations have a count smaller than 200. Therefore, this example illustrates a small domain population estimation problem. The covariate combinations are based on single year of age (truncated at age 89), sex (male and female) and 67 geographic areas. The area variable is included in the coverage models as an area-specific intercept; age is included as bases of a cubic spline with internal knots set to $\{5, 15, 25, 35, 45, 55, 65, 75\}$, and sex is included as a binary variable. The coverage model also includes interactions between sex and the age spline terms. We set the coverage model parameter vectors, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, to achieve pre-set overall coverage rates (approximately 70% for each list, and 90% for both lists combined). Summaries of the simulated target population and the observed data are plotted in Figures 1 and 2.

In Figure 1, the observed and target counts are plotted by single year of age and sex, marginalized over area. In Figure 2, the relative differences of the observed and target counts are plotted for each area, marginalized over age and sex. (Relative differences are plotted instead of absolute counts because the areas have high variability with respect to size.)

Figures 1 and 2 demonstrate the lists have appreciable under-coverage with respect to the target population. In order to estimate the size and structure of the target population, dual systems estimation needs to estimate and adjust for this under-coverage.

6.2. Estimating the target population from observed data

In this section, we apply the approaches discussed in Section 5 to estimate the simulated target population described in Section 6.1. Given the observed data (i.e. individuals observed on at least one of the two lists), the estimation problem is to infer the total population size, N, and the distribution (over the covariates) of the subset of the population missed by both lists, i.e. $\mathbf{X}_{(0,0)}$ for the unit-record population structure, or equivalently, $\mathbf{M}_{\cdot,(0,0)}$, for the aggregated population structure. We adopt the Jeffreys' prior for the total population size N. Priors for other model parameters are discussed below.

For the examples in this section, we use samplers implemented for the aggregated population structure. We demonstrate three approaches:

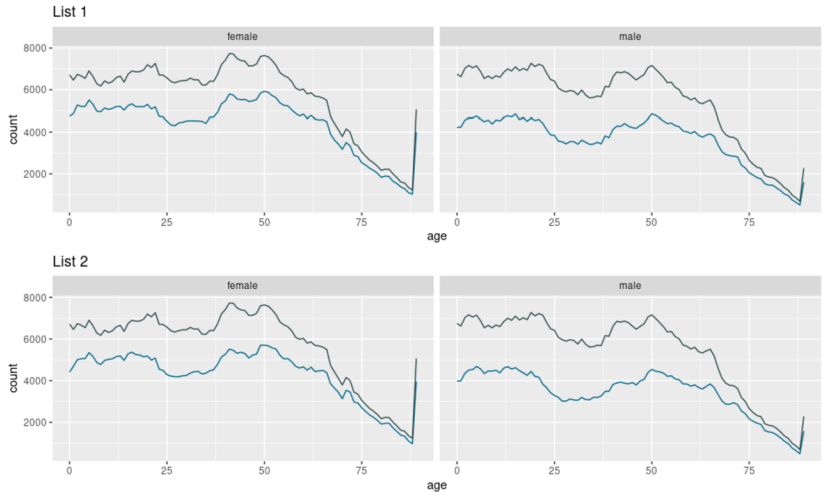

FIG 1. *Target counts (in gray) and observed counts (in blue) by single year of age (truncated at 89) and sex, marginalized over area, for each of the two lists.*

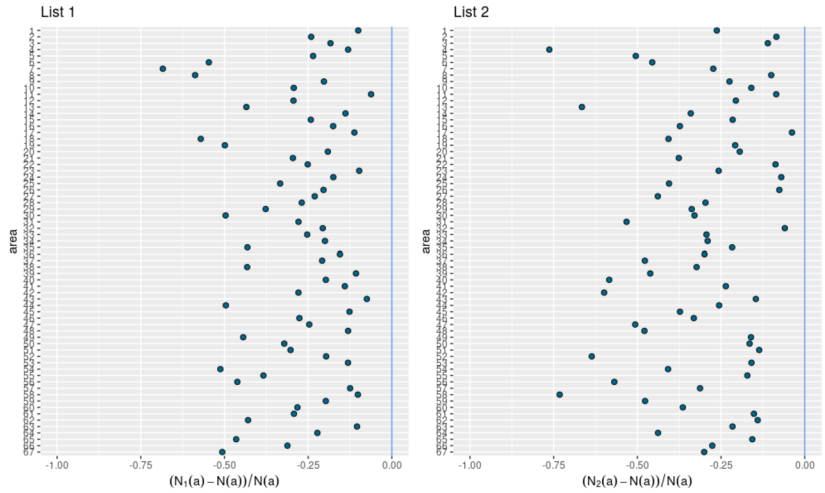

FIG 2. *Relative differences of the counts observed on each list, by geographic area, marginalized over age and sex. The relative difference for any given area is the difference between the observed and target counts relative to the target counts. The vertical blue line at zero signifies no difference between the observed and target counts.*

(i) The marginal likelihood approach, where we approximate (or sample from) the posterior distribution $p(\boldsymbol{\beta}, \boldsymbol{\theta}|\mathbf{D}_{\text{obs}})$, which is the full parameter posterior $p(N, \boldsymbol{\beta}, \boldsymbol{\theta}|\mathbf{D}_{\text{obs}})$ marginalised over N. Since we have adopted the Jeffreys' prior for N, the marginal likelihood for $(\boldsymbol{\beta}, \boldsymbol{\theta})$ is the marginal likelihood given in (31). We use the probabilistic programming software,

Stan (Stan Development Team, 2021) to implement this approach; Stan uses Hamiltonian Monte Carlo (HMC) which is a highly efficient approach to Markov Chain Monte Carlo (MCMC) posterior computation.

(ii) The full observed likelihood approach, where we approximate the posterior distribution of all the parameters, $p(N, \boldsymbol{\beta}, \boldsymbol{\theta}|\mathbf{D}_{\text{obs}})$. We also implement this approach in Stan and, since the sampler in Stan allows only continuous parameters, we use a continuous approximation to the prior for the population size N; that is, we let $p(N) \propto 1/N$, but do not restrict N to be integer-valued.

(iii) The data augmentation (or Gibbs sampler) approach, where we approximate the joint posterior $p(\mathbf{M}_{\cdot,(0,0)}, N, \boldsymbol{\beta}, \boldsymbol{\theta}|\mathbf{D}_{\text{obs}})$, as described in Section 5.4.

Note that, when using the marginal likelihood approach, given draws of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, the population size N and the counts vector $\mathbf{M}_{\cdot,(0,0)}$ are drawn from the conditional posterior distributions (33) and (37) described in Section 5.2 and 5.3, respectively. This step is straightforward since, under the Jeffreys' prior the conditional posterior distribution for N is Negative-Binomial, and the conditional posterior for $\mathbf{M}_{\cdot,(0,0)}$ is the multinomial given by (37), and both the Negative-Binomial and multinomial distributions are easy to sample from. Similarly, when using the full likelihood approach, given draws of $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ and N, $\mathbf{M}_{\cdot,(0,0)}$, is drawn from its multinomial conditional posterior (37).

Within the samplers, the coverage models for the two lists are two-level logistic regression models where the area-specific intercepts are modelled hierarchically. For each list $j \in \{1, 2\}$, area a and covariates vector $\check{\mathbf{x}}$, the coverage model is

$$\tilde{\phi}_j(\mathbf{x}, \boldsymbol{\beta}_j) = \text{invlogit}(\beta_{j,0,a} + \check{\mathbf{x}}' \, \boldsymbol{\beta}_{j,1}), \quad \forall(\check{\mathbf{x}}, a) \tag{40}$$

$$p(\beta_{j,0,a}|\beta_{j,0}, \sigma_j^2) = \text{N}(\beta_{j,0,a}|\beta_{j,0}, \sigma_j^2), \quad \forall a, \tag{41}$$

$$p(\boldsymbol{\beta}_{j,1}) = \text{MVN}(\beta_{j,1}|\mathbf{0}, \boldsymbol{\Sigma}_j), \tag{42}$$

$$p(\beta_{j,0}) = \text{N}(\beta_{j,0}|m_j, s_j^2) \tag{43}$$

$$p(\sigma_j) = \text{U}[l_j, u_j], \tag{44}$$

where the parameters of the prior/hyper-prior distributions are set to $\boldsymbol{\Sigma}_j = 100 \, \boldsymbol{I}$ (where $\boldsymbol{I}$ is the identity matrix), $m_j = 0$, $s_j^2 = 100$, and $[l_j, u_j] = [0.001, 3]$, for both lists. The parameters are assumed to be a priori independent. The covariate coefficients vector for List j, $\boldsymbol{\beta}_{j,1}$, includes coefficients for the age spline terms and sex, and their interactions. The hierarchical model for area effects, in which the area effects are modelled in (41) as conditionally independent draws from a common Normal distribution, with mean and variance also treated as unknowns, permits partial pooling of information across areas, which improves the precision of the estimates for area effects, particularly for smaller areas. An alternative would be to include area effects in (40) by replacing $\beta_{j,0}$ with a linear combination of area indicators, to form a single-level logistic model. If the prior for the logistic model parameters was specified as independent normal distributions, as in (42), the information used to estimate each of the area

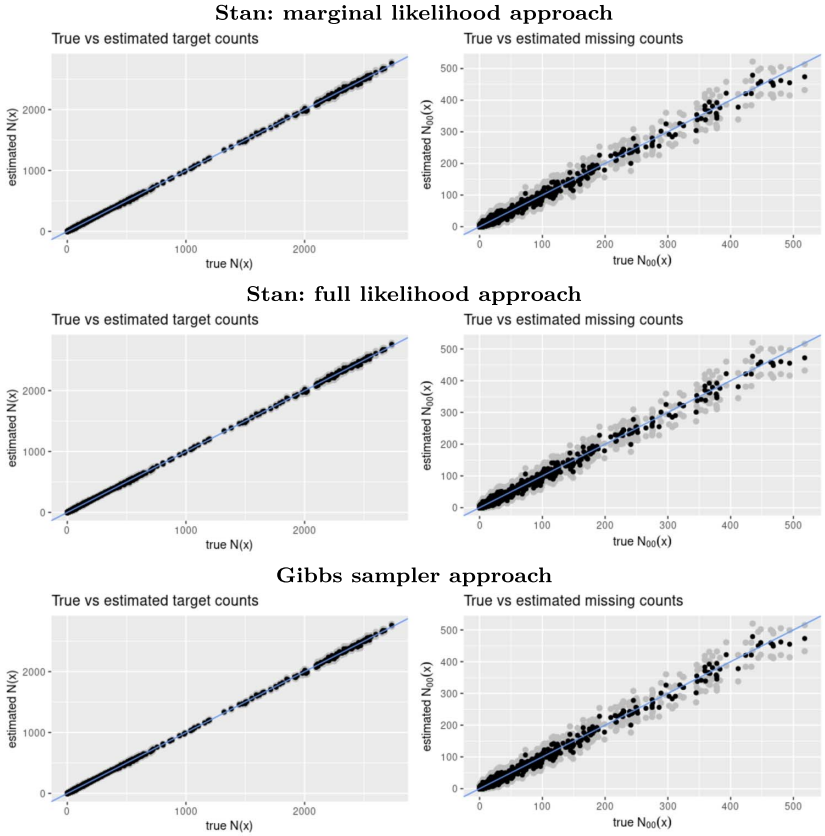

FIG 3. *Estimated covariate combination counts plotted against the true covariate combination counts, for marginal likelihood approach using Stan (row 1), full likelihood approach using Stan (row 2), and Gibbs sampler approach (row 3). Target (total) counts are plotted in the left column and missing counts, i.e. counts of individuals not observed by any list, are plotted on the right. In each plot, points correspond to covariate combinations. The black points are the median estimates, while the gray points are the 0.05 and 0.95 quantiles of the posterior draws.*

effects would be limited to the data for the corresponding area. That is, under the single-level logistic model formulation, with *a priori* independence assumed for the area effects, information regarding area effects would not be pooled over areas. The conditional independence model given by (41) is a convenient and widely used approach to allowing information to be pooled over areas (Gelman et al., 2014, Chapter 5).

In Figure 3, we have plotted the posterior median covariate combination counts against the true covariate combination counts, for the full target population (left column) and the subset missed by both lists (right column). The rows correspond to the marginal likelihood approach, the full likelihood approach, and the Gibbs sampler approach, respectively. Each point in the plots corre-

sponds to a unique covariate combination. To indicate the uncertainty in the estimates, the 0.05 and 0.95 quantiles of the posterior samples for each covariate combination are also plotted (in gray). All three approaches produce similar posterior distributions demonstrating the approaches are equivalent, and the estimates are generally centred on the $x = y$ line. Computationally, the Gibbs sampler, which uses the Metropolis-Hastings algorithm to update parameters, is run for many more iterations until convergence compared to the Stan (HMC) sampler. We ran three parallel chains, each for 200,000 iterations for the Gibbs sampler, including a burn-in of 80,000, and 4,000 iterations for the Stan sampler, with a burn-in of 1,000. The post-burn-in draws were thinned to get a posterior sample of approximately 2,000. The convergence diagnostic $\hat{R}$ was generally less than 1.05 for the Gibbs sampler and less than 1.02 for the Stan sampler. The runtimes for the examples presented here were comparable, however, in general, the Gibbs sampler has lower per-iteration runtime but must run for tens of thousands of iterations before convergence, while the Stan sampler has higher per-iteration runtime, but requires fewer iterations until convergence.

To demonstrate the population estimates and credible intervals by selected dimensions (covariates), in Figure 4, we have plotted the relative differences of the estimated counts (medians and 90% equal-tail-area credible intervals) by age and sex, marginalized over area, and in Figure 5, we have plotted the same for area, marginalized over age and sex. In both cases, we have used the posterior samples from the Gibbs sampler approach, but the posterior samples using Stan produce similar results. The 90% credible intervals of the relative differences overlap with the vertical/horizontal line at zero for most covariate combinations, implying that the estimated intervals capture the true value for most covariate combinations.

6.3. Model misspecification and posterior predictive model checking

The results reported above show that the underlying population structure can be recovered under any of the three computational approaches considered. However, these results were produced under the unrealistic scenario of a known data-generating model, i.e. we fitted a model that exactly matched the structure of the known data generating model. In practice, the data generating model is never exactly known and the fitted model may be misspecified. However, model checking can identify defects in the model specification and correcting these defects can lead to an improved model. The basic premise of posterior predictive model checking is that, if a fitted model is a reasonable representation of a model that could have generated the observed data, new data generated from the fitted model should look much like the observed data, allowing for random variation (Gelman et al., 2014, Chapter 6).

For Bayesian dual systems estimation, posterior predictive model checking is easily implemented: (i) For each draw from the joint posterior for the model parameters and covariate distribution for the group missed by both lists, that is, $p(\mathbf{M}_{\cdot,(0,0)}, N, \boldsymbol{\beta}, \boldsymbol{\theta}|\mathbf{D}_{\text{obs}})$, generate predicted list inclusion cell counts by covariate combination; (ii) Compare the observed counts against the posterior

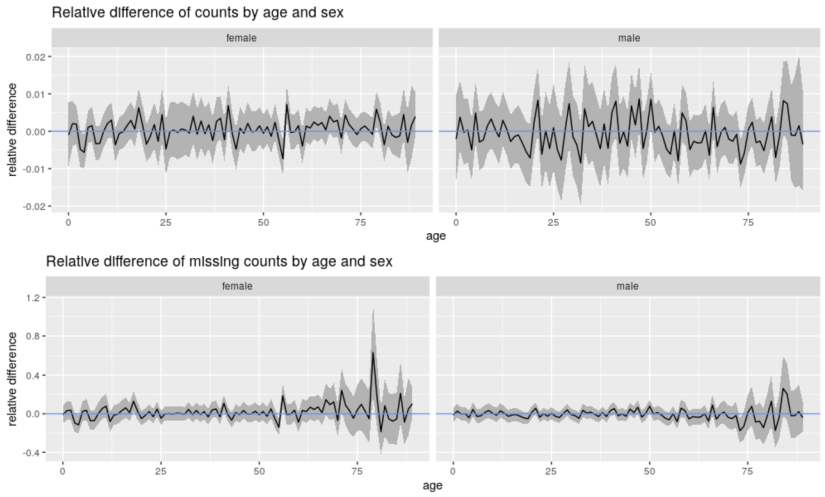

FIG 4. *Relative differences of estimated and true counts, relative to true counts, by single year of age and sex. Estimated covariate combination totals (which include observed counts) are plotted in the top row, and estimated covariate combination counts missing from both lists are plotted in the bottom row.*

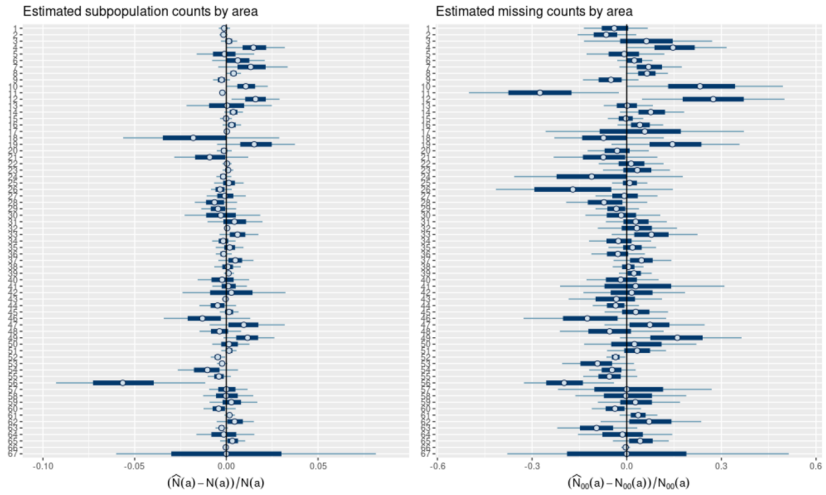

FIG 5. *Relative differences of estimated and true counts, relative to true counts, by geographic area. Estimated area totals (which include observed counts) are plotted in the left column, and estimated area counts missing from both lists are plotted in the right column.*

predictive distribution for the observable cell counts. If we find that observed counts often fall in the tail of the corresponding posterior predictive distribution, the model specification must be called into question. An algorithm for generating the posterior predictive distribution for the observable list inclusion

cell counts is given in Algorithm 3.

Algorithm 3 Generating the posterior predictive distribution of list inclusion cell counts.

1: **for** t in $\{1, \ldots, T\}$ **do**
2: step (i): Draw $(\mathbf{M}^{(t)}_{\cdot,(0,0)}, N^{(t)}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\theta}^{(t)})$ from $p(\mathbf{M}_{\cdot,(0,0)}, N, \boldsymbol{\beta}, \boldsymbol{\theta} | \mathbf{D}^{\text{obs}})$
3: **for** k in $1, \ldots, K$ **do**
4: step (ii): Set $M^{(t)}_{k,+} = M_{k,(1,1)} + M_{k,(1,0)} + M_{k,(0,1)} + M^{(t)}_{k,(0,0)}$
5: step (iii): Draw

$$(M^{\text{pred},(t)}_{k,(1,1)}, M^{\text{pred},(t)}_{k,(1,0)}, M^{\text{pred},(t)}_{k,(0,1)}, M^{\text{pred},(t)}_{k,(0,0)})' \sim \text{Multinomial}\left(M^{(t)}_{k,+}, \boldsymbol{\phi}(\mathbf{x}_{[k]}, \boldsymbol{\beta}^{(t)})\right)$$

6: step (iv): Set (predicted list counts)

$$\tilde{M}^{\text{pred},(t)}_{k,1} = M^{\text{pred},(t)}_{k,(1,1)} + M^{\text{pred},(t)}_{k,(1,0)}$$
$$\tilde{M}^{\text{pred},(t)}_{k,2} = M^{\text{pred},(t)}_{k,(1,1)} + M^{\text{pred},(t)}_{k,(0,1)}$$

7: **end for**
8: **end for**

In Algorithm 3, an aggregated data structure is assumed but the algorithm is easily adapted to a unit-record population structure by generating a predicted list inclusion cell for each record in the simulated population, based on the recorded covariates for observed records and the predicted covariate values for the unobserved group. Thus, in the unit-record case, for the t^{th} draw from the posterior, the predicted list inclusion cell indicators are generated by sampling from

$$Y^{\text{pred},(t)}_i \overset{\text{indep}}{\sim} \text{Categorical}\big(\boldsymbol{\phi}(\mathbf{x}^{(t)}_i, \boldsymbol{\beta}^{(t)})\big), \quad i \in \{1, \ldots, N^{(t)}\}. \tag{45}$$

In Algorithm 3, and in the adaptation to unit-record data just described, we condition on the covariate values for the n_{obs} individuals recorded on the observed list. This focuses attention on the adequacy of the coverage models, which, given our adoption of an unstructured multinomial for the covariate distribution, is likely to be the major source of model uncertainty. However, it is also possible to include prediction of the covariate distribution in the posterior predictive simulation by including a step to generate the full covariate distribution conditional on $(N, \boldsymbol{\beta}, \boldsymbol{\theta})$ before generating the list inclusion cell counts or indicators (step (iv) in Algorithm 3). This would facilitate checking aspects of the modelled covariate distribution.

To illustrate the posterior predictive checking for dual systems estimation and to explore how model misspecification manifests in posterior predictive checking, we fitted a misspecified version of the model discussed in Section 6.1, which omitted the sex by age interaction and modelled age effects using a simplified spline model with only three knots at ages 15, 25 and 65. Aside from these changes, the misspecified model had the same structure as the model given in (40)–(44). In particular, area-level effects were modelled hierarchically. The

misspecified model was fitted in Stan, using the marginal likelihood for $(\boldsymbol{\beta}, \boldsymbol{\theta})$ and assuming Jeffreys' prior for N. The same uninformative priors that were used for fitting the correctly specified model were adopted for the misspecified model.

We computed posterior predictive checks for the counts on each list, and for each of the three observable list inclusion cells, by covariate combination. Since the pattern of results was similar for each of the model checking targets, we report only results for the counts in the $(1, 1)$ cell below.

Several metrics for comparing the posterior predictive distributions with observed counts are presented in Table 2, for different levels of aggregation, beginning with the most granular level in the simulated data, which is area by sex by single year of age. The metrics reported in Table 2 are:

1. The difference between the posterior predictive median for the number in both lists and the observed count on both lists, averaged over covariate combinations (*mean diff*);
2. The relative median absolute deviation defined as the ratio of the median (over covariate combinations) of the absolute difference between the posterior predictive median and the observed count for the $(1, 1)$ cell, for the misspecified model relative to the correctly specified model (*rmad*);
3. The proportion of covariate combinations for which the posterior predictive 95% intervals include the observed count (*coverage*). We note that although we refer to this concept as "coverage" it is not the same concept as frequentist coverage of intervals, which refers to the proportion of intervals including a target value, under repeated sampling;
4. The average length of the 95% posterior predictive intervals under the misspecified model divided by the average length of the 95% posterior predictive intervals under the correctly specified model (*rel length*).

The latter metric (relative interval length) is not a measure of model-fit but is included to check whether it is likely that differences in the coverage of posterior predictive intervals could be attributed to differences in the length of the intervals. For example, excessive interval length can lead to high coverage even if posterior medians are suggestive of poor fit, and low interval coverage can result from intervals that are too narrow.

We use relative (to the correct model) measures of median absolute deviation and interval length because the value of these measures naturally increases as the level of aggregation increases, and the relevant comparison is between the misspecified and correct model. However, actual values of the mean difference between posterior predictive medians and observed values are reported in Table 2, because they show that average differences are small, for both the misspecified and correct model, at all levels of aggregation. In practice, we would not have a known correctly specified model to benchmark model checking results against, but working with simulated data affords us the opportunity to gain insight into how posterior predictive checking reveals issues with model misspecification.

TABLE 2
Comparison of posterior predictive checks for the number observed on both lists, for correct and misspecified models at different levels of aggregation.

granularity	measure	correct model	misspecified model
area, sex, age	mean diff[(a)]	0.224	−0.016
area, sex, age	rmad[(b)]	1.000	1.000
area, sex, age	coverage[(c)]	1.000	0.968
area, sex, age	rel. length[(d)]	1.000	1.000
area	mean diff	0.134	−0.127
area	rmad	1.000	0.500
area	coverage	1.000	1.000
area	rel. length	1.000	0.989
sex, age	mean diff	0.086	−0.047
sex, age	rmad	1.000	3.091
sex, age	coverage	0.972	0.500
sex, age	rel. length	1.000	0.965
sex, 5 year age	mean diff	0.472	−0.444
sex, 5 year age	rmad	1.000	10.619
sex, 5 year age	coverage	1.000	0.278
sex, 5 year age	rel. length	1.000	0.893

[(a)] Average (over covariate combinations) of the difference between the posterior predictive median and the observed count of people in the $(1, 1)$ cell;
[(b)] Ratio of the median absolute deviation of the posterior predictive median from the observed count for the misspecified model, compared to the correct model;
[(c)] Proportion of 95% posterior predictive intervals that include the observed count;
[(d)] Ratio of average length of 95% posterior predictive intervals obtained under the misspecified model, compared to the correct model.

At the lowest level of aggregation (11,929 combinations of area by sex by age combinations), there are no real differences between the model checking metrics for the misspecified and correct models. In particular, the coverage of the 95% posterior predictive intervals for the misspecified model is 96.8.% Although this is less than the coverage of 100% achieved by the correct model, 96.8% of 95% posterior predictive intervals including target values would not usually be regarded as indicative of problems with a model. Similarly, when aggregating counts to the area-level, there appears to be little difference between the misspecified model and correctly specified models. Since the misspecfied model differed from the correct model only in the specification of the effects of age and sex, this is not unexpected.

When counts are aggregated to sex by single year of age, resulting in 180 sex by age combinations, the coverage of the 95% posterior predictive intervals under the misspecified model drops dramatically, to 50.0%, whereas the coverage of the posterior predictive intervals for the correctly specified model remains high at 97.2%. The relative length of the posterior predictive intervals under the misspecified model is only slightly less than under the correctly specified model and is unlikely to explain the under-coverage of the posterior predictive intervals for

the misspecified model. The slightly reduced interval length for the misspecified model, reflects a small reduction in posterior variance for the coverage model parameters, due to fitting a model with fewer parameters. The poor coverage of the posterior predictive intervals for the misspecified model is associated with the magnitude of the discrepancies between the posterior predictive median and the observed counts: The median absolute deviation for the misspecified model, is just over 3 times the corresponding value for the correctly specified model. However the average difference between posterior medians and the observed counts remains low for both models, suggesting that although the magnitude of discrepancies is greater for the misspecified model, there are both positive and negative discrepancies that average out at close to zero, when averaging over covariate combinations.

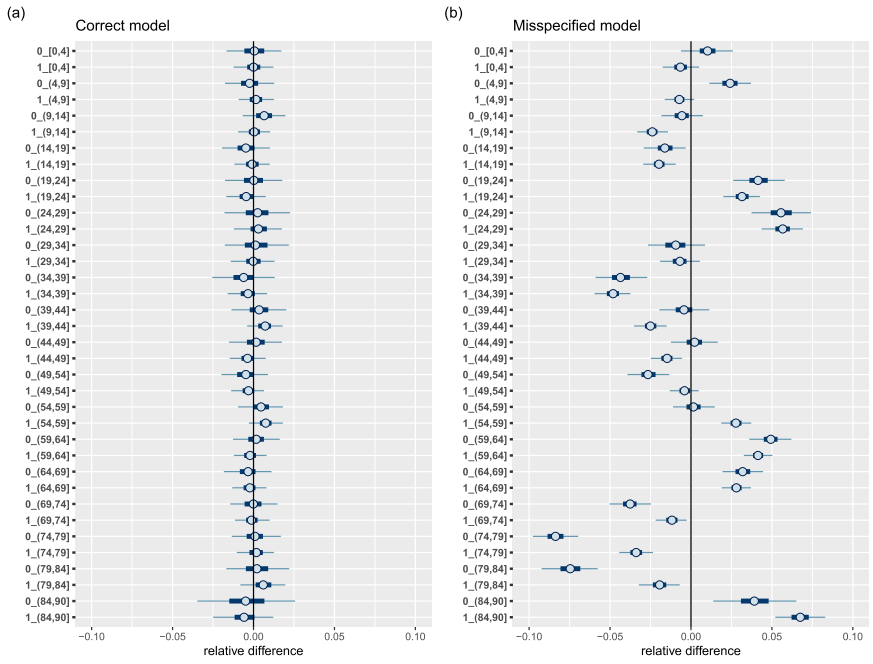

FIG 6. *Posterior predictive check for the number observed on both lists, by sex and 5-year age groups. The sex-age group combinations are indicated on the y-axis: The first character of the row label indicates sex (0 for male; 1 for female), with the age interval following. 95% and 50% (darker shaded inner interval) posterior predictive intervals are shown for the relative difference between model predictions and the observed count. For the misspecified model (panel (b)), only 10 of the 36 95% posterior predictive intervals include the observed value (indicated by the vertical line at 0.0). For the correct model (panel (a)), all the 95% posterior predictive intervals include the observed value.*

With further aggregation to sex and 5-year age groups, the coverage of the posterior predictive intervals declines further to 27.8% with the posterior predictive intervals including observed counts for only 10 of the 36 sex by 5-year age-group combinations. This is suggestive of serious problems with the mis-

specified model. In contrast, all 36 of the posterior predictive intervals for the correctly specified model included the observed count. The median absolute deviation for the misspecified model is about 10.6 times the median absolute deviation for the correctly specified model, suggesting the low coverage of the the posterior predictive intervals obtained under the misspecified model is attributable to problems with the location of these intervals. This is clearly illustrated in Figure 6 which shows posterior predictive 95% intervals for the relative difference between predicted values and the observed values, computed as (predicted − observed)/observed, for the correct and misspecified models. While the intervals for the correctly specified model are tightly clustered around the vertical line, indicating zero relative difference, the intervals for the misspecified model are more scattered, with the median relative difference exceeding 5% for several of the sex by 5-year age group combinations. As previously noted, only 10 of the posterior predictive intervals for the misspecified model include the null value of no difference.

It is interesting that model misspecification appears to have minimal effect on the model-fit when interest is confined to the most granular level of estimation, yet the mis-fit of the misspecified model becomes clearly apparent at higher levels of aggregation. At the most granular level (area by sex by age in our example), the impact of model misspecification is likely small relative to random variation. As attention turns to more aggregated levels of estimation, the impact of random variation, relative to the magnitude of the counts becomes less, and therefore, the impact of model misspecification becomes more apparent.

We note that the pattern of the posterior predictive model checking results for the misspecified and correctly specified models is replicated for population estimates obtained under the two models. In Table 3, we present similar metrics to Table 2, for comparing population estimates against true values. Thus, for example, the coverage measure is now the proportion of the covariate combinations for which the 95% credible interval includes the true population counts. For estimates by area, sex and age, it can be seen that there is very little difference between the performance of the misspecified and correctly specified models. This holds also for estimation by area (aggregating over levels of sex and age). However, the effect of model misspecification becomes clearly apparent when aggregating over area to obtain estimates by sex and age. For estimates by sex and single year of age, the median absolute deviation for the misspecified model is about three times the corresponding figure for the correctly specified model. For the misspecified model, the proportion of 95% credible intervals that include the true count is just less than 54%, whereas the coverage of the credible intervals for the correctly specified model is 96.7%. The performance of the misspecified model deteriorates further when age is aggregated to five year intervals: The median absolute deviation is 5.6 times the corresponding figure for the correctly specified model, and the coverage of the 95% credible intervals declines to 30.6%, compared to 97.2% for the correctly specified model. The contrast in the performance of the 95% credible intervals obtained under the two models is clearly apparent in Figure 7, which reports credible

TABLE 3
Metrics comparing summaries of the posterior distribution for small domain population counts against true values, at different levels of aggregation, for correct and misspecified models.

granularity	measure	correct model	misspecified model
area, sex, age	mean diff[(a)]	−0.228	−0.268
area, sex, age	rmad[(a)]	1.000	1.000
area, sex, age	coverage[(a)]	0.975	0.969
area, sex, age	rel. length[(a)]	1.000	0.998
area	mean diff	−1.269	−8.664
area	rmad	1.000	0.842
area	coverage	0.970	0.970
area	rel. length	1.000	0.991
sex, age	mean diff	−0.414	−3.194
sex, age	rmad	1.000	3.000
sex, age	coverage	0.967	0.539
sex, age	rel. length	1.000	0.968
sex, 5 year age	mean diff	−1.417	−14.889
sex, 5 year age	rmad	1.000	5.580
sex, 5 year age	coverage	0.972	0.306
sex, 5 year age	rel. length	1.000	0.896

[(a)] Average (over covariate combinations) of the difference between the posterior median and the true population count;
[(b)] Ratio of the median absolute deviation of the posterior median from the true value, for the misspecified model, compared to the correct model;
[(c)] Proportion of 95% credible intervals that include the true population count;
[(d)] Ratio of average length of 95% credible intervals obtained under the misspecified model, compared to the correct model.

intervals for the relative difference between posterior estimates and the true values.

From these investigations, we can conclude that posterior predictive model checking is useful for detecting problems with model fit that are likely to impact on population estimation. However, it is important to check model performance at all levels of aggregation for which estimates are likely to be produced. The impact of model misspecification may not be apparent at low levels of aggregation because the impact of random variation may outweigh biases due to model misspecification. Posterior predictive checking can reveal components of the model specification that need attention. For example our model-checking results revealed no issues at the area level but discrepancies between model predictions and observed values, by sex and age. In a real data situation, this would clearly point to problems in the specification of the sex by age effects in the coverage models.

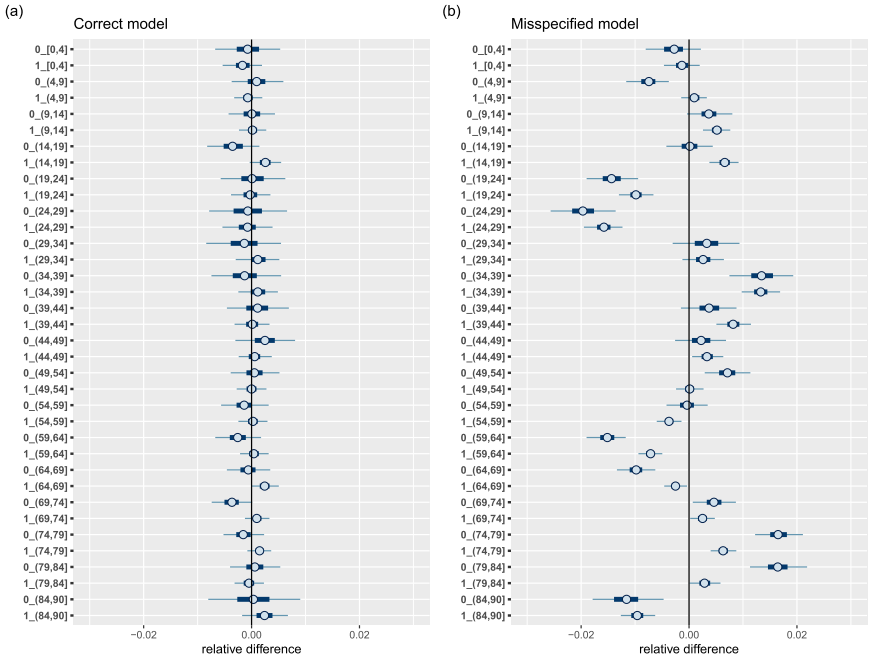

FIG 7. *Posterior medians and* 95% *credible intervals, by sex and 5-year age groups, for the correct (panel (a)) and misspecified (panel (b)) models. Estimates are presented as relative difference from the true value. The open circles indicate the difference between the posterior median and the true value, divided by the true value. Interval endpoints are defined similarly. 95% and 50% (darker shaded inner interval) credible intervals are shown. The sex-age group combinations are indicated on the y-axis: The first character of the row label indicates sex (0 for male; 1 for female), with the age interval following. For the misspecified model only 11 of the 36 95% posterior predictive intervals include the observed value (indicated by the vertical line at 0.0). For the correct model, the 95% credible values include the observed value for all 36 sex-age group combinations.*

7. Relaxing the conditional independence assumption

7.1. Identifiability considerations

So far we have followed standard practice by assuming conditional independence between inclusion on the two lists, given covariates. Any attempt to model dependence between inclusion on the two lists must deal with an inherent identifiability issue, related to the group omitted from both lists. If we allow inclusion on List 2, say, to depend on inclusion on List 1, after controlling for covariates, then we are saying that, within levels of the covariates, the probability of inclusion on List 2 differs for the groups included and omitted from List 1, i.e. $\Pr(L_2 = 1|\mathbf{X} = \mathbf{x}, L_1 = 1) \neq \Pr(L_2 = 1|\mathbf{X} = \mathbf{x}, L_1 = 0)$. However, only people included on List 2 are observable for the group omitted from List 1. Thus, although we observe data from which we can estimate $\Pr(L_2 = 1|\mathbf{X} = \mathbf{x}, L_1 = 1)$,

we cannot estimate $\Pr(L_2 = 1|\mathbf{X} = \mathbf{x}, L_1 = 0)$ directly from the observed data because, for the $L_1 = 0$ group, only people with $L_2 = 1$ are observed. Intuitively, therefore, it seems clear that estimating the association between inclusion on List 1 and inclusion on List 2, is not, directly, supported by the usual data structure for dual systems estimation. This is the essence of the identifiability issue that makes moving away from the assumption of conditional independence difficult.

Some mathematical insight into the identifiability issue can be gained by considering the marginal likelihood for the coverage model and covariate distribution parameters, $(\boldsymbol{\beta}, \boldsymbol{\theta})$, with Jeffreys' prior for total population size, as given in (18). Suppose we place no restrictions on the inclusion cell probabilities, except the requirement that they sum to one, at each covariate setting. Thus we are removing the conditional independence assumption. The population averaged probability of being missed by both lists is

$$\begin{aligned} p_{(0,0)}(\boldsymbol{\beta}, \boldsymbol{\theta}) &= \int \left(1 - (\phi_{(1,1)}(\mathbf{x}, \boldsymbol{\beta}) + \phi_{(1,0)}(\mathbf{x}, \boldsymbol{\beta}) + \phi_{(0,1)}(\mathbf{x}, \boldsymbol{\beta}))\right)\ p(\mathbf{x}|\boldsymbol{\theta})\, \mathrm{d}\mathbf{x} \\ &= 1 - \int (\phi_{(1,1)}(\mathbf{x}, \boldsymbol{\beta}) + \phi_{(1,0)}(\mathbf{x}, \boldsymbol{\beta}) + \phi_{(0,1)}(\mathbf{x}, \boldsymbol{\beta}))\ p(\mathbf{x}|\boldsymbol{\theta})\, \mathrm{d}\mathbf{x}. \end{aligned}$$

Now, suppose we multiply each of $\phi_{(1,1)}(\mathbf{x}, \boldsymbol{\beta}), \phi_{(1,0)}(\mathbf{x}, \boldsymbol{\beta})$ and $\phi_{(0,1)}(\mathbf{x}, \boldsymbol{\beta})$ by α satisfying $0 < \alpha < 1$, for each covariate combination $\mathbf{x}$, to obtain new cell probabilities $\phi_{\alpha,y}(\mathbf{x}, \boldsymbol{\beta})$, for $y \neq (0, 0)$, and $\phi_{\alpha,(0,0)}(\mathbf{x}, \boldsymbol{\beta}) = 1 - \alpha \sum_{y \neq (0,0)} \phi_{\alpha,y}(\mathbf{x}, \boldsymbol{\beta})$. With these modified cell probabilities, the population-averaged probability of being missed by both lists is

$$\begin{aligned} p_{\alpha,(0,0)}(\boldsymbol{\beta}, \boldsymbol{\theta}) &= \int \phi_{\alpha,(0,0)}(\mathbf{x}, \boldsymbol{\beta})\ p(\mathbf{x}|\boldsymbol{\theta}) d\mathbf{x} \\ &= 1 - \alpha \int (\phi_{(1,1)}(\mathbf{x}, \boldsymbol{\beta}) + \phi_{(1,0)}(\mathbf{x}, \boldsymbol{\beta}) + \phi_{(0,1)}(\mathbf{x}, \boldsymbol{\beta}))\ p(\mathbf{x}|\boldsymbol{\theta})\, \mathrm{d}\mathbf{x} \\ &= 1 - \alpha(1 - p_{(0,0)}(\boldsymbol{\beta}, \boldsymbol{\theta})) \end{aligned}$$

and, from (18), the marginal likelihood for $(\boldsymbol{\beta}, \boldsymbol{\theta})$ is therefore

$$\prod_{i: y_i \neq (0,0)} \frac{\alpha\ \phi_{y_i}(\mathbf{x}_i, \boldsymbol{\beta})\ p(\mathbf{x}_i|\boldsymbol{\theta})}{\alpha\ (1 - p_{(0,0)}(\boldsymbol{\beta}, \boldsymbol{\theta}))} = \prod_{i: y_i \neq (0,0)} \frac{\phi_{y_i}(\mathbf{x}_i, \boldsymbol{\beta})\ p(\mathbf{x}_i|\boldsymbol{\theta})}{1 - p_{(0,0)}(\boldsymbol{\beta}, \boldsymbol{\theta})},$$

which is just the marginal likelihood for the original inclusion cell probabilities and covariate distribution parameters. Thus, for any particular value of cell probabilities $\{\phi^*_{(1,1)}(\mathbf{x}, \boldsymbol{\beta}), \phi^*_{(1,0)}(\mathbf{x}, \boldsymbol{\beta}), \phi^*_{(0,1)}(\mathbf{x}, \boldsymbol{\beta}), \forall \mathbf{x} \in \mathcal{X}\}$, the marginal likelihood evaluated at that value of the cell probabilities is identical to the marginal likelihood evaluated at $\{\alpha\phi^*_{(1,1)}(\mathbf{x}, \boldsymbol{\beta}), \alpha\phi^*_{(1,0)}(\mathbf{x}, \boldsymbol{\beta}), \alpha\phi^*_{(0,1)}(\mathbf{x}, \boldsymbol{\beta}), \forall \mathbf{x} \in \mathcal{X}\}$, for all $0 < \alpha < 1$, and for any $\boldsymbol{\theta}$. Thus, even though the implications for the $(0, 0)$ cell probability of multiplying the observable cell probabilities by $0 < \alpha < 1$ may be substantial (e.g. consider $\alpha = 0.01$), the marginal likelihood cannot discriminate between $\{\phi^*_{(1,1)}(\mathbf{x}, \boldsymbol{\beta}), \phi^*_{(1,0)}(\mathbf{x}, \boldsymbol{\beta}), \phi^*_{(0,1)}(\mathbf{x}, \boldsymbol{\beta}), \forall \mathbf{x} \in \mathcal{X}\}$,

and $\{\alpha\phi^*_{(1,1)}(\mathbf{x},\boldsymbol{\beta}), \alpha\phi^*_{(1,0)}(\mathbf{x}), \alpha\phi^*_{(0,1)}(\mathbf{x}), \forall \mathbf{x} \in \mathcal{X}\}$. Moreover, this phenomenon holds for any realisation of the observable data. Without some restrictions on the cell probabilities, the observable data cannot identify the cell probabilities. It can be easily verified that under the conditional independence assumption, multiplying the two sets of list inclusion probabilities by a constant $0 < \alpha < 1$ does lead to different values of the marginal likelihood, assuming Jeffreys' prior for N.

While the above argument is specific to the marginal likelihood under the Jeffreys' prior for N, it does illustrate the difficulty of learning about the cell inclusion probabilities from just the observable data. In 7.3, we discuss the use of an informative prior for sub-population totals to help strengthen inferences from dependent lists.

7.2. Structuring list inclusion dependence using the odds ratio

Some structure on the cell probabilities is necessary for identifiability but the model of conditional independence is not the only identifying structure that can be considered. A natural alternative to the conditional independence model is a model that maintains the logistic coverage models for each list but introduces a specific parametric form for the association between inclusion on the lists. This makes the list inclusion cell probabilities dependent on both the parameters of the marginal list coverage models, $\boldsymbol{\beta}$, and parameters of the model for dependence between inclusion on the two lists. For binary variables, a standard measure of association is the odds ratio. Accordingly, we let $\rho(\mathbf{x}) = \big(\phi_{(1,1)}(\mathbf{x},\boldsymbol{\beta},\boldsymbol{\rho})\ \phi_{(0,0)}(\mathbf{x},\boldsymbol{\beta},\boldsymbol{\rho})\big)\ /\ \big(\phi_{(1,0)}(\mathbf{x},\boldsymbol{\beta},\boldsymbol{\rho})\ \phi_{(0,1)}(\mathbf{x},\boldsymbol{\beta},\boldsymbol{\rho})\big)$ denote the odds ratio for list inclusion, for covariate combination $\mathbf{x}$, where $\boldsymbol{\rho}$ is either a vector of association parameters, which may be odds ratios for each value of $\mathbf{x}$ or the parameters of a model relating the odds ratio to covariates, or a scalar if a common odds ratio is assumed for all covariate combinations. The conditional independence model corresponds to the case with $\rho(\mathbf{x}) = 1, \forall \mathbf{x} \in \mathcal{X}$. Setting the odds ratios to other fixed values also yields an identified likelihood.

For a fixed value of $\rho(\mathbf{x}) \neq 1$ and marginal coverage probabilities $\tilde{\phi}_1(\mathbf{x},\boldsymbol{\beta}_1)$ and $\tilde{\phi}_2(\mathbf{x},\boldsymbol{\beta}_2)$, for Lists 1 and 2 respectively, it follows from Lee (1997) that the probability of inclusion in the $(0,0)$ cell for covariate combination $\mathbf{x}$ is given by the solution of the quadratic equation

$$\begin{aligned}(\rho(\mathbf{x})-1)\ \phi^2_{(0,0)}(\mathbf{x},\boldsymbol{\beta},\boldsymbol{\rho})&\\ -\phi_{(0,0)}(\mathbf{x},\boldsymbol{\beta},\boldsymbol{\rho})\big(1+(\rho(\mathbf{x})-1)(2-\tilde{\phi}_1(\mathbf{x},\boldsymbol{\beta}_1)-\tilde{\phi}_2(\mathbf{x},\boldsymbol{\beta}_2))\big)&\\ +\rho(\mathbf{x})(1-\tilde{\phi}_1(\mathbf{x},\boldsymbol{\beta}_1))(1-\tilde{\phi}_2(\mathbf{x},\boldsymbol{\beta}_2))=0&\end{aligned} \tag{46}$$

that satisfies

$$\begin{aligned}\max\big(0,1-\tilde{\phi}_1(\mathbf{x},\boldsymbol{\beta}_1)-\tilde{\phi}_2(\mathbf{x},\boldsymbol{\beta}_2)\big) &\leq \phi_{00}(\mathbf{x},\boldsymbol{\beta},\boldsymbol{\rho})\\ &\leq \min\big(1-\tilde{\phi}_1(\mathbf{x},\boldsymbol{\beta}_1),1-\tilde{\phi}_2(\mathbf{x},\boldsymbol{\beta}_2)\big).\end{aligned} \tag{47}$$

With $\phi_{(0,0)}(\mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\rho})$ obtained from (46) and (47), the remaining cell probabilities follow easily as

$$\phi_{(1,0)}(\mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\rho}) = (1 - \tilde{\phi}_2(\mathbf{x}, \boldsymbol{\beta}_2)) - \phi_{(0,0)}(\mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\rho}) \tag{48}$$

$$\phi_{(0,1)}(\mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\rho}) = (1 - \tilde{\phi}_1(\mathbf{x}, \boldsymbol{\beta}_1)) - \phi_{(0,0)}(\mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\rho}) \tag{49}$$

$$\phi_{(1,1)}(\mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\rho}) = \tilde{\phi}_1(\mathbf{x}, \boldsymbol{\beta}_1) - \phi_{(1,0)}(\mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\rho}) = \tilde{\phi}_2(\mathbf{x}, \boldsymbol{\beta}_2) - \phi_{(0,1)}(\mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\rho}). \tag{50}$$

For given values of $\boldsymbol{\rho}$, all likelihood calculations follow as in the conditional independence case, but with cell probabilities obtained from (46) and (47) and (48)–(50) rather than as products of the corresponding marginal probabilities. Posterior computation using the Gibbs sampler is more complex in the case of dependent list inclusions because the conditional posterior for the marginal coverage model parameters does not separate into distinct conditional posterior distributions for the two marginal coverage models. If the full or marginal likelihood is used directly to obtain the posterior for the model parameters only minor modifications are required to the conditional independence implementation to compute the cell probabilities case using (46), (47) and (48)–(50)) instead of the products of the marginal coverage probabilities.

Using any of the approaches to fitting dual systems estimation models considered in this paper, the sensitivity of population estimates to departures from conditional independence can be assessed by running additional analyses with fixed values of the odds ratios. A special case that may suffice for an initial investigation of sensitivity to dependence is a model with a common odds ratio for each covariate combination. A simple illustration of this idea is given in Section 7.2.1 below.

As an alternative, or adjunct, to running a series of analyses with fixed values of the odds ratios (or other measures of association), a more fully Bayesian perspective leads to placing a prior on the odds ratio parameter(s). In view of the identifiability issues noted above, we should not expect the data to be informative about the odds ratios, but adopting a prior for the odds ratio parameters allows uncertainty concerning the dependence between inclusion on List 1 and List 2 to be propagated through to population estimation. In the example considered in Section 7.2.1 below, uncertainty concerning dependence makes a substantial contribution to uncertainty concerning population size.

If the odds ratio is allowed to vary with covariate combinations, a model could be formulated such as

$$\log(\rho(\mathbf{x})) = \delta_0 + \mathbf{x}'\boldsymbol{\delta}_1$$

and prior distributions specified on the parameters of this model.

Alternatively, a hierarchically structured prior for covariate specific odds ratios could be considered, so that covariate specific odds ratios are modelled as varying around some common mean. One such hierarchical structure is:

$$[\rho_k|\rho_0] \overset{\text{indep}}{\sim} \text{TNorm}(\rho_0, \tau, d_l, d_u),\ k \in \{1, \ldots, K\}$$
$$\rho_0 \sim \text{TNorm}(\mu, \sigma, c_l, c_u)$$

$$\tau \sim F_\tau \tag{51}$$

where, for simplicity, truncated normal priors have been proposed, with truncation points c_l and c_u for the second-level model, and d_l and d_u for the covariate-level model. Given the identifiability issue, τ must be specified, or assigned a tight prior, rather than relying on the data to inform the posterior for this parameter. The model is a way of structuring prior beliefs about the structure of dependence. For example, if a scenario where the assumption of a common dependence odds ratio across all covariate combinations was thought unlikely to hold but odds ratios were nevertheless considered unlikely to vary greatly by covariate combination, τ could either be set to a modest value or assigned a prior density concentrated on small values.

In some circumstances the prior for the odds ratio parameters may be informed by analysis external to the dual systems estimation. For example, in the context of population estimation based on a census and a coverage survey, Brown, Abbott and Diamond (2006) estimate odds ratios, by small geographic area, from preliminary analysis of household-level data and extrapolate the household-level odds ratios to the person-level odds ratios, so that it can be incorporated in person-level dual systems estimation. In the application Brown, Abbott and Diamond (2006) describe, the household-level odds ratios can be estimated because of the existence of an assumed high quality estimate of the total number of households by small geographic area. For a known population size, the odds ratio is identified from dual systems data, because the number of units (households in this case) missed by both lists is known. Consequently, given known population size, all four cells of the tables cross-classifying list inclusion indicators are observed and the odds ratios (for households) can be straightforwardly estimated. We return to this idea in Section 7.3.

Brown, Abbott and Diamond (2006) note that the confidence intervals obtained from person-level dual systems estimation with dependence odds ratios fixed at the values obtained by extrapolating the household-level odds ratios do not reflect uncertainty in the household-level odds ratios or the extrapolation to the person level. From a Bayesian viewpoint, Brown et al's approach could be used to inform priors for small area odds ratio parameters, thereby allowing uncertainty in the estimated odds ratios to be incorporated in the analysis.

By placing a prior on the odds ratio parameters, we are regarding the odds ratios as model parameters and the joint posterior of the model parameters is now

$$p(N, \boldsymbol{\rho}, \boldsymbol{\beta}, \boldsymbol{\theta} | \mathbf{D}^{\text{obs}}) \propto p(N, \boldsymbol{\rho}, \boldsymbol{\beta}, \boldsymbol{\theta})\, p(\mathbf{D}^{\text{obs}} | N, \boldsymbol{\rho}, \boldsymbol{\beta}, \boldsymbol{\theta})$$

The likelihood $p(\mathbf{D}^{\text{obs}}|N, \boldsymbol{\rho}, \boldsymbol{\beta}, \boldsymbol{\theta})$ is as given by (14) or (23) for the unit record and aggregated population data structures, respectively, except that the cell probabilities should now be viewed as functions of $\boldsymbol{\rho}$ as well as $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$, and are obtained from (46) and (47) and (48)–(50). Similarly, the population averaged probability of being missed by both lists, is now a function of $\boldsymbol{\rho}$, as well as $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$.

7.2.1. Numerical illustration of dual systems estimation with dependence parameterised using odds ratio

To illustrate dual systems estimation with dependent list inclusion, we consider a simple example based on a version of the synthetic dataset described in Section 6.1, aggregated over geographic area to produce a dataset with 180 sex by age combinations and no other covariates. We simulated list inclusion cell counts, using logistic coverage models with coefficients for sex, age spline and sex by age spline interaction effects set to the same values used to generate the synthetic data of Section 6.1. We fixed the odds ratio for the association between inclusion on the two lists to $\rho = 5$, for all sex by age combinations, and generated the cell inclusion counts using (46), (47) and (48)–(50) to define the cell probabilities for covariate-specific multinomial distributions with size, $M_{k,+}$ defined by the sex by age totals in the simulated population dataset of Section 6. We also generated a second dataset using the same sex by age covariate structure and marginal coverage models, but assuming conditional independence for list inclusion.

The assumed dependence odds ratio of 5 represents a moderately strong degree of dependence between inclusion on the lists. For example, if $\Pr(L_2 = 1|L_1 = 0, \mathbf{X} = \mathbf{x}) = 0.8$, then an odds ratio of 5 implies $\Pr(L_2 = 1|L_1 = 1, \mathbf{X} = \mathbf{x}) = 0.95$, whereas if $\Pr(L_2 = 1|L_1 = 0, \mathbf{X} = \mathbf{x}) = 0.6$, an odds ratio of 5 implies $\Pr(L_2 = 1|L_1 = 1, \mathbf{X} = \mathbf{x}) = 0.88$. We adopted this moderately high value of the odds ratio so that any difficulties in recovering the true value of the odds ratio from the estimated models would be readily apparent.

We fitted a series of models to the synthetic data generated with dependence, assuming different values for the fixed odds ratio. For each of these models, we adopted Jeffreys' prior for N, a Dirichlet($\boldsymbol{\alpha}$) prior for the covariate distribution parameters, with $\boldsymbol{\alpha} = \mathbf{0.01}$, and Normal$(0, 10)$ priors for all coverage model parameters. We fitted the model in Stan using the marginal likelihood for $(\boldsymbol{\beta}, \boldsymbol{\theta})$ to obtain a sample from the posterior for these parameters, followed by application of Algorithm 1 to obtain a posterior sample over the completed population. To obtain the posterior samples for $(\boldsymbol{\beta}, \boldsymbol{\theta})$ in Stan, we ran five parallel HMC chains of 2,900 iterations, discarded the first 2,000 iterations as burn in, and thinned the chains by three to produce a nominal Monte Carlo sample size of 1,500. These HMC settings ensured $\hat{R} \leq 1.01$ and an effective posterior sample size of at least several hundred for all parameters.

Figure 8 shows how estimates change as models with different assumed dependence odds ratios are fitted. Population estimates increase as the strength of the assumed association between inclusion on the two lists increases. Consequently, when the assumed value of the odds ratio is less than the true value ($\rho = 5$), the population counts are underestimated, whereas population counts are over-estimated when the assumed odds ratio is greater than the true value. The proportion of covariate combinations for which 95% credible intervals for coverage probabilities or population counts contain the true value varies dramatically depending on the distance of the assumed dependence odds ratio from the true value of 5. For example, when ρ is assumed to equal 1, none of the 95%

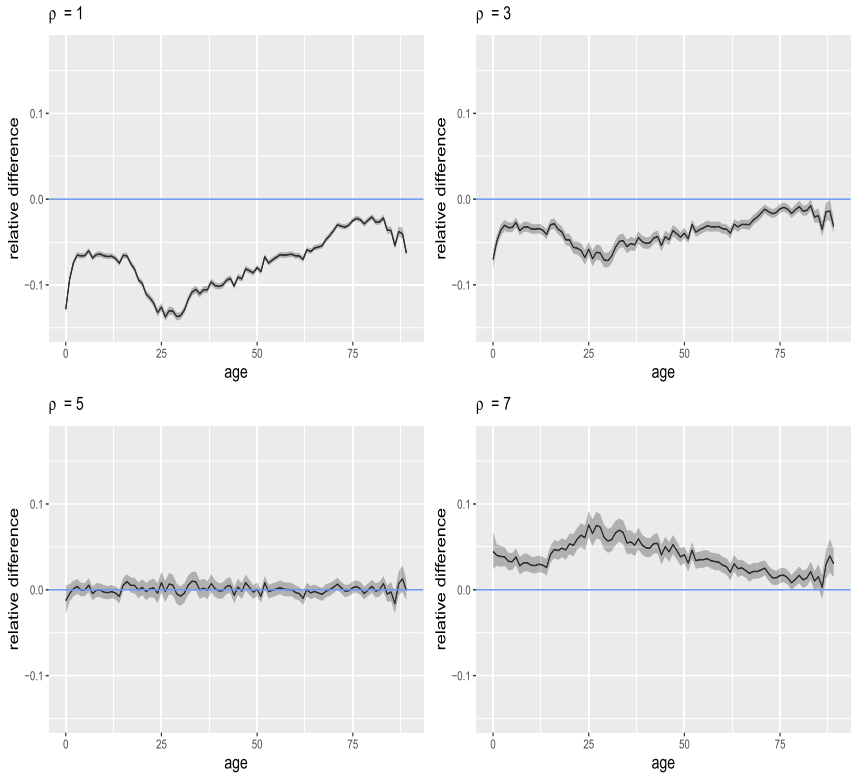

FIG 8. *Posterior median (black line) and 95% credible intervals (grey shading) for population counts for females, by age, expressed as a relative difference between estimated and true values, for four models fitted with different assumed values of the dependence odds ratio. The horizontal blue line indicates exact equivalence between estimated and true counts. The models were all fitted to a dataset generated under an assumption of a common odds ratio of $\rho = 5$ for all sex and age combinations. (Results for males are not shown.) Assuming a dependence odds ratio less than the true value leads to underestimation of the population for all covariate combinations. Assuming a dependence odds ratio greater than the true value leads to overestimation of the population.*

credible intervals for counts or list coverage probabilities include the true value. In contrast, when ρ is assumed to equal the true value of 5, 95% credible intervals for population counts, List 1 coverage probabilities and List 2 coverage probabilities, included 94%, 98% and 90% of the true values, respectively. Posterior uncertainty increases with the assumed value of ρ.

In spite of the clear differences between the fixed ρ models in terms of their ability to recover the underlying population structure, a range of posterior predictive checks similar to those discussed in Section 6.3 suggested each of the fixed ρ models fitted the observed data equally well. For example, for each of the models all 95% posterior predictive intervals for the counts of people recorded on both lists, by sex and age, included the true values. This is a reflection of the

underlying identifiability problem: observed data cannot discriminate between different values of ρ and, consequently, models that assume different values for ρ fit the data equally well.

Given the sensitivity of population estimates to the dependence odds ratio, illustrated in Figure 8, it may also be of interest to integrate over a plausible range of uncertainty concerning the odds ratio. This can be achieved by specifying a prior for ρ and including ρ as a parameter in the model. This can be conveniently implemented in Stan using either the marginal or full likelihood approach. For illustration, we considered a truncated normal prior centred at the true value of $\rho = 5$, with truncation points set to 0.001 and 9.999 and standard deviation parameter set to 1.7845 to ensure the prior probability that ρ was less than one was 0.01. Values of ρ less than one indicate negative dependence, which is possible, but positive dependence is the more common concern in applications. We ran five parallel HMC chains for 4,900 iterations in Stan, and discarded the first 3,000 iterations of each chain as burn-in. This produced adequate convergence with all $\hat{R}$ statistics less than 1.034. Effective Monte Carlo sample sizes were at least 140 (after thinning the chains by three). Despite the apparent convergence of the MCMC procedure, the procedure failed to recover the true parameter values. For example, ρ was underestimated and the coverage model intercept parameters were overestimated, as shown in the first row of Figure 9. Similar results were obtained when we reran the model for 100,000 iterations, discarding the first 50,000 as burn-in, suggesting the results are not due to the sampler getting trapped around local modes. From Figure 9, it can be seen that ρ and the two intercept parameters are clearly highly correlated in the posterior and it seems likely these high correlations make it difficult for the sampler to concentrate near the underlying parameter values. The underestimation of ρ and overestimation of the intercept parameters leads to a substantial underestimation of the population for all sex by age groups, as shown in Figure 11 (first column).

For comparison, we also fitted a model with a much tighter truncated normal prior for ρ, $\text{TNorm}(5, 0.4295, 0.001, 9.999)$, which assigns prior probability of 0.95 to the interval $(4.03, 5.97)$. Under this prior, the posterior obtained using the marginal likelihood approach, implemented in Stan assuming Jeffreys' prior for N, did recover the true model parameters, as shown in the second row of Figure 9. Consequently, credible intervals for sex and age-specific counts are approximately centred on the true values as shown in the first column of Figure 12.

An alternative approximation to the posterior can be obtained by observing that the joint posterior for the model parameters can be written as

$$p(N, \rho, \boldsymbol{\beta}, \boldsymbol{\theta} | \mathbf{D}^{\text{obs}}) = p(N, \boldsymbol{\beta}, \boldsymbol{\theta} | \mathbf{D}^{\text{obs}}, \rho)\ p(\rho | \mathbf{D}^{\text{obs}}),$$

and since we do not expect the data to be informative with respect to the odds ratio, a reasonable approximation to the joint posterior for the model parameters, is therefore

$$p(N, \rho, \boldsymbol{\beta}, \boldsymbol{\theta} | \mathbf{D}^{\text{obs}}) \approx p(N, \boldsymbol{\beta}, \boldsymbol{\theta} | \mathbf{D}^{\text{obs}}, \rho)\ p(\rho). \tag{52}$$

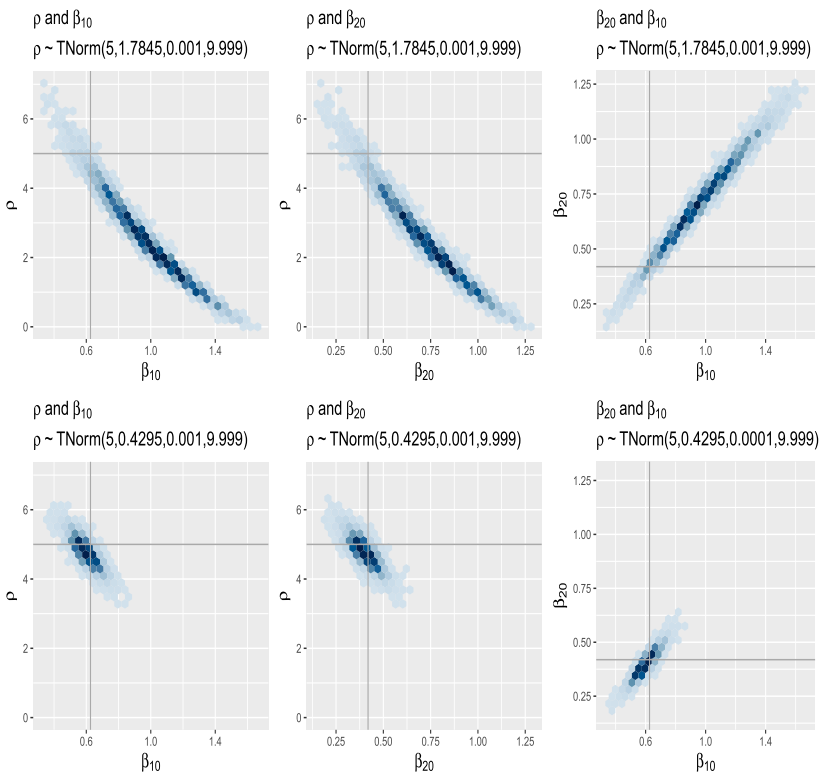

FIG 9. *Joint posterior distribution of the odds ratio (ρ) and the intercept parameters of the list coverage models, obtained from fitting a dependent dual systems model in Stan, under a TNorm(5,1.7845,0.001,9.999) prior for ρ (top row), and a much less diffuse TNorm(5,0.4295,0.001,9.999) prior for ρ (bottom row). All coverage model parameters were assigned $N(0, 3)$ priors. The data were generated from a model with the odds ratio set to 5 for all 180 covariate combinations. Areas of higher density are indicated by darker shading. The true parameter values are represented by the solid lines drawn perpendicular to the axes. The posterior under the more diffuse TNorm(5,1.7845,0.001,9.999) prior for ρ fails to recover the true parameters values. The posterior is more nearly centred on the true parameter values under the tighter TNorm(5,0.4295,0.001,9.999) prior for ρ. This prior also results in a substantial reduction in posterior variance for all three parameters. Note that scales differ for each plot to accommodate the different locations of the posterior distribution.*

That is, we approximate the posterior for ρ by the prior for ρ because the lack of information in the data concerning ρ suggests the posterior should be similar to the prior. The decomposition in (52) is equivalent to the posterior arising from a "cut model" in which the data is not allowed to inform the posterior for one or more parameters, that is, the feedback from the data is cut for some parameters, ρ, in this case (Plummer, 2015; Carmona and Nicholls, 2020). The cut model approximation to the posterior can be implemented by looping over draws from the prior for ρ and, for each sampled value of ρ, fitting a dependent dual systems model with the odds ratio set to the sampled

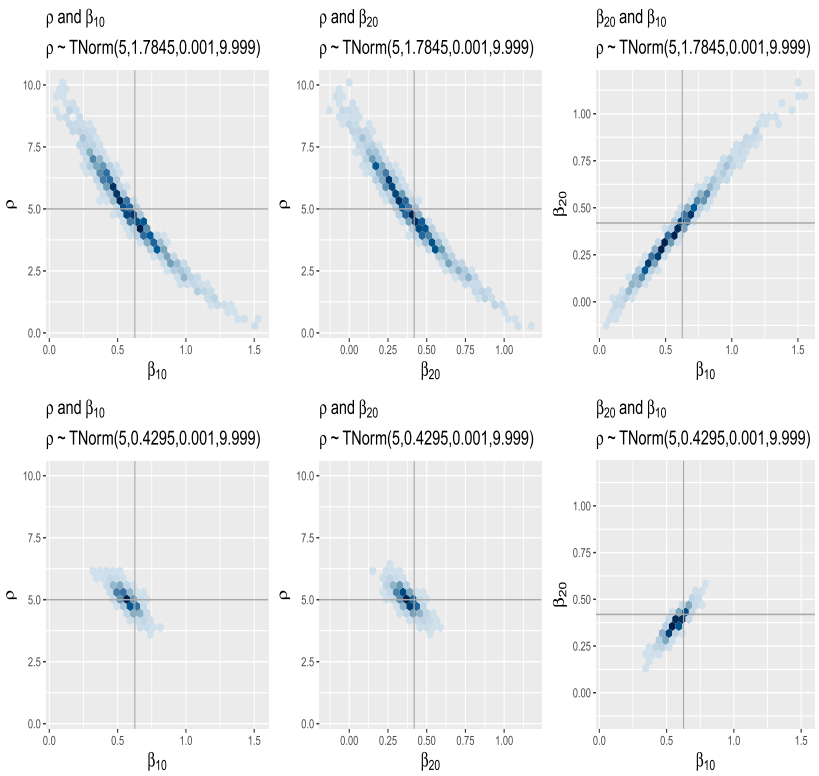

FIG 10. *Approximate joint posterior distribution of the odds ratio (ρ) and the intercept parameters of the list coverage models, obtained from fitting a dependent dual systems model using a cut model, whereby values of ρ are repeatedly drawn from the prior and coverage model parameters estimated conditional on the generated ρ value. The prior for ρ is not updated by the data. The top row shows results under a TNorm(5,1.7845,0.001,9.999) prior for ρ, while the bottom row shows results for the less diffuse TNorm(5,0.4295,0.001,9.999) prior for ρ. Jeffreys' prior was assumed for the total population size and all coverage model parameters were assigned $N(0, 10)$ priors. The data were generated from a model with the odds ratio set to 5 for all 180 covariate combinations. Areas of higher density are indicated by darker shading. The true parameter values are represented by the solid lines drawn perpendicular to the axes. Using the cut model, posterior distributions are approximately centred on the underlying parameter values for both priors. The less diffuse prior for ρ results in a substantial reduction in posterior variance for all three parameter.*

value. The collection of generated values of ρ and other parameters constitutes a sample from the approximate posterior (52). We implemented this procedure for 1,000 draws from the prior for ρ for both the TNorm(5,1.7845,0.001,9.999) and TNorm(5,0.4295,0.001,9.999) priors. For each draw from the prior for ρ, we ran a dependent dual systems model in Stan, for 4,000 iterations (including a burn-in of $3,500$, and stored the parameter values obtained on the last iteration. This procedure produces a sample from the approximate posterior in (52). Results are summarised in Figure 10. Posterior correlations between the model

intercept parameters remain very high, but the posterior samples are located close to the underlying population values. Graphs of population estimates by sex and age are presented in Figures 11 (diffuse prior) and 12 (tight prior), for the fully Bayesian approach (left column) and the cut model (right column). The plots suggest that the cut model posterior is centred approximately on the true values, for both priors. (Note that the scales of the y-axis are different in the two sets of plots.) The cut model approach appears to be a more reliable method than fully Bayesian modelling in Stan for fitting dependent dual systems models that include a non-degenerate prior on the dependence parameter.

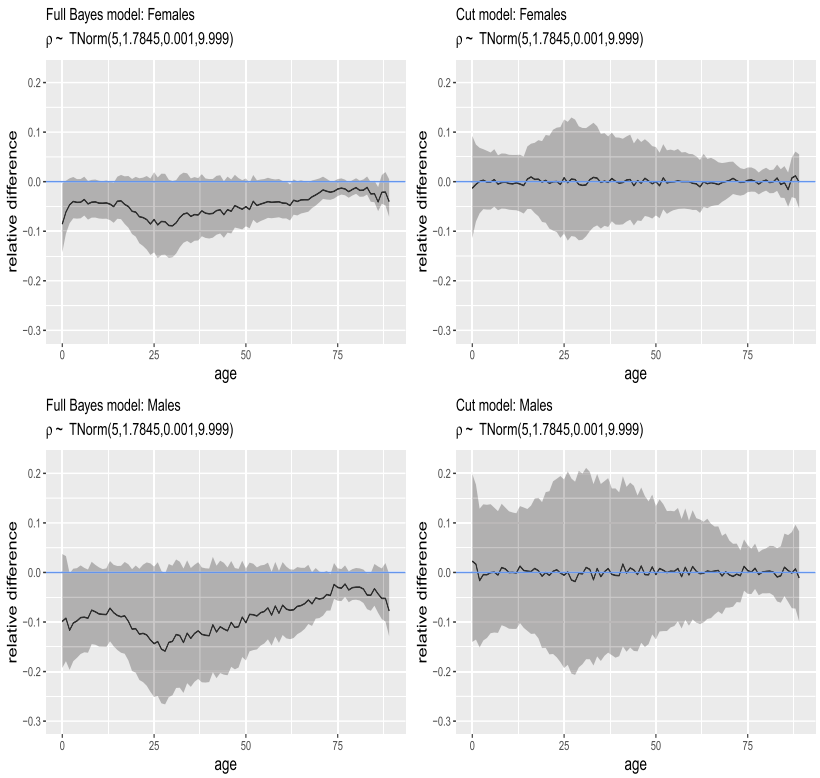

FIG 11. *Population estimates by sex and age, under fully Bayesian (first column) and cut model (second column) versions of dependent dual systems models under the* $\mathrm{TNorm}(5, 1.7845, 0.001, 9.999)$ *prior for the dependence odds ratio* ρ*, assuming Jeffreys' prior for the total population size. Estimates are expressed as relative difference from true values. The shaded area represents an equal-tail-area* 95% *credible interval, and the solid black line represents the posterior median, transformed to relative difference from the true population counts. The full Bayesian approach appears to underestimate at all ages, for both males and females.*

We note that the preceding analyses are intended only as illustrations of the sensitivity of population estimation to assumptions concerning dependence between inclusion on the two lists and to highlight some difficulties in obtaining

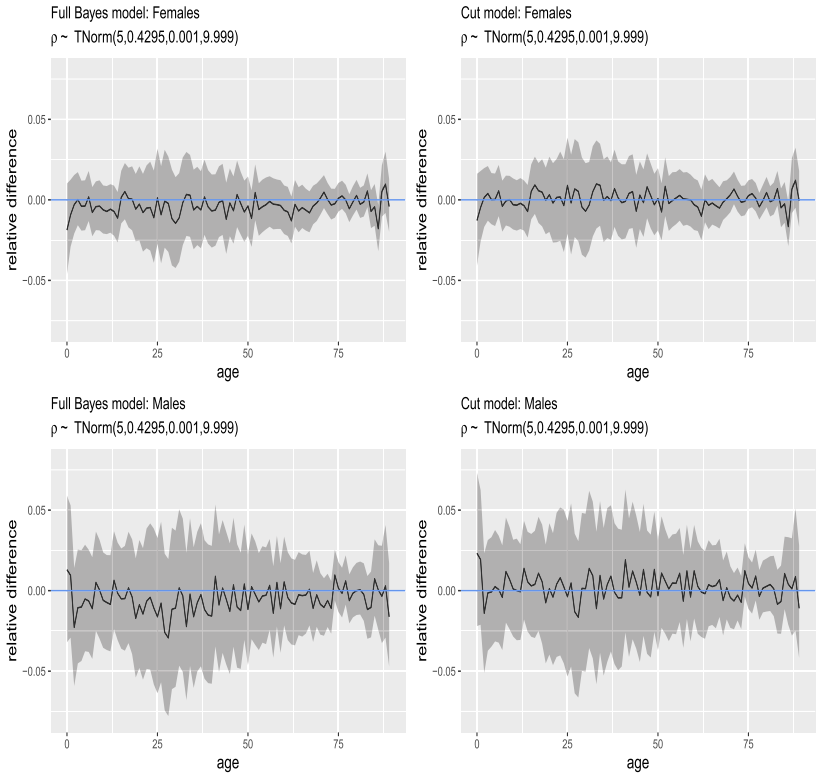

FIG 12. *Population estimates by sex and age, under fully Bayesian (first column) and cut model (second column) versions of dependent dual systems models under the* $\text{TNorm}(5, 0.4295, 0.001, 9.999)$ *prior for the dependence odds ratio* ρ*, and assuming Jeffreys' prior for the total population size. Estimates are expressed as relative difference from true values. The shaded area represents an equal-tail-area* 95% *credible interval, and the solid black line represents the posterior median, transformed to relative difference from the true population counts. Both full Bayesian and cut model implementations of the model appear to recover the underlying population counts.*

posterior distributions for models that are not fully identified. In reality, we would not have the luxury of being able to centre the prior for ρ on a known true value, and would be dependent on prior information to determine a reasonable prior.

7.3. Using external information from demographic analyses

In some situations, high quality aggregate-level population estimates may be available prior to small area dual systems estimation. For example, national statistical offices usually produce regular updates of population estimates. Given a reliable base population estimate, comprehensive births and deaths registration and high quality data on external migration, the total population can be easily

updated by adding births, subtracting deaths and adding net (inward) migration. With accurate recording of sex and age, this basic accounting approach to population estimation can be extended to give population estimates by age and sex. If these accounting-based estimates are regarded as sufficiently accurate that they can be considered known, the remaining task for small domain population estimation is to estimate the population distribution by small area and other demographic characteristics (such as ethnic group), within levels of the assumed known marginal totals. We demonstrate below that, given known marginal totals, it is possible to recover the true dependence odds ratio parameters, though under the assumption that they are constant over covariate combinations within the known margins. In practice, there may be some uncertainty in accounting based population estimates, arising for example, from uncertainty in the population used as the base and potential delays in recording births or deaths or errors in recording age or sex in migration data. Uncertainty about the aggregate population estimates can clearly be accommodated in the prior. Indeed, in the case that a prior population estimate is available only for the total population size, our model set-up already explicitly accommodates prior uncertainty on N.

To accommodate prior information on sub-population totals, some changes are needed to the likelihood structure and we discuss these in Section 7.3.1 below.

7.3.1. Incorporating prior information on sub-population totals

Suppose there are G aggregate groups with population totals $N_1, \ldots, N_G$. We assume the groups form a mutually exclusive and exhaustive partition of the population, so $N = \sum_{g=1}^{G} N_g$. We re-define $\mathbf{X}$ to refer to the covariates other than those used to define the G sub-populations, and let $\boldsymbol{\theta}_g$ denote the parameters of the distribution of the $\mathbf{X}$ covariates within the g^{th} sub-population. For simplicity, we consider discrete covariates and let

$$\theta_{g,k} = \Pr(\mathbf{X} = \mathbf{x}_{[k]} | B = g, \boldsymbol{\theta}_g),$$

where B is a discrete variable taking values in $\{1, \ldots, G\}$, representing the sub-population to which an individual belongs, and $\boldsymbol{\theta}_g = (\theta_{g,1}, \ldots, \theta_{g,K})$ is the probability vector for the K covariate combinations within the g^{th} sub-population. We also let $M_{g,k,y}$ denote the number of people in sub-population g with covariate combination $\mathbf{x}_{[k]}$, in list inclusion cell y, $M_{g,k,+} = \sum_y M_{g,k,y}$, the total number of people in sub-population g with covariate combination k, and $M_{g,+,y} = \sum_{k=1}^{K} M_{g,k,y}$, the total count in sub-population g in inclusion cell y. It follows that the number of individuals recorded on at least one list within sub-population g is $n_{g,\text{obs}} = N_g - M_{g,+,(0,0)}$. It is also convenient to define the cell inclusion probabilities in a sub-population specific manner as

$$\phi_{g,y}(\mathbf{x}_{[k]}, \boldsymbol{\beta}, \rho_g) = \Pr(Y = y | B = g, \mathbf{X} = \mathbf{x}_{[k]}, \boldsymbol{\beta}, \rho_g), \text{ for } y \in \mathcal{Y},$$

where ρ_g is the dependence odds ratio for the g^{th} sub-population, and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$ are the parameters for the models for inclusion on Lists 1 and 2. We denote the vector of list inclusion cell probabilities corresponding to the k^{th} covariate combination in sub-population g by

$$\boldsymbol{\phi}_g(\mathbf{x}_{[k]}, \boldsymbol{\beta}, \rho_g) = (\phi_{g,(1,1)}(\mathbf{x}_{[k]}, \boldsymbol{\beta}, \rho_g), \phi_{g,(1,0)}(\mathbf{x}_{[k]}, \boldsymbol{\beta}, \rho_g), \phi_{g,(0,1)}(\mathbf{x}_{[k]}, \boldsymbol{\beta}, \rho_g), \phi_{g,(0,0)}(\mathbf{x}_{[k]}, \boldsymbol{\beta}, \rho_g))'.$$

The sub-population cell inclusion probabilities are obtained from marginal list inclusion models and sub-population odds ratios, exactly as described by equations (46)–(50). The marginal inclusion models are not changed by the existence of the G sub-populations for which strong prior information on population totals is available, but it is useful to explicitly represent these sub-populations in the model notation, so we write the model for inclusion in List j as

$$\tilde{\phi}_{g,j}(\mathbf{x}_{[k]}, \boldsymbol{\beta}) = \Pr(L_j = 1 | B = g, \mathbf{X} = \mathbf{x}_{[k]}, \boldsymbol{\beta}_j),\ j \in \{1, 2\}, \tag{53}$$

whereas, in our previous notation, the sub-population g would have been incorporated in $\mathbf{x}_{[k]}$. We can also define a sub-population-averaged probability of being missed by both lists as

$$p_{g,(0,0)}(\boldsymbol{\beta}, \boldsymbol{\theta}_g, \rho_g) = \sum_k \phi_{g,(0,0)}(\mathbf{x}_{[k]}, \boldsymbol{\beta}, \rho_g)\ \theta_{g,k}. \tag{54}$$

To accommodate prior information on the sub-population totals in dual systems estimation, we regard them as model parameters and the likelihood must be structured to accommodate them. To recognise the G sub-populations in the model structure, we assume a product multinomial form for the joint distribution of the $\mathbf{x}$ covariates. The distribution over the inclusion cells is modelled conditionally on sub-population and covariate combination and is also assumed to be multinomial. The data model is therefore

$$[\{M_{g,k,+},\ k = 1, \ldots, K\} | \boldsymbol{\theta}_g, N_g] \overset{\text{indep}}{\sim} \text{Multinomial}(N_g, \boldsymbol{\theta}_g),\quad g \in \{1, \ldots, G\} \tag{55}$$

$$\begin{aligned}\left[\left(M_{g,k,(1,1)}, M_{g,k,(1,0)}, M_{g,k,(0,1)}, M_{g,k,(0,0)}\right) | \boldsymbol{\beta}, \boldsymbol{\rho}, M_{g,k,+}\right] &\overset{\text{indep}}{\sim} \\ \text{Multinomial}\left(M_{g,k,+}, \boldsymbol{\phi}_g(\mathbf{x}_{[k]}, \boldsymbol{\beta}, \rho_g)\right),& \\ g \in \{1, \ldots, G\}, k \in \{1, \ldots, K\},&\end{aligned} \tag{56}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \ldots, \boldsymbol{\theta}_G')$, and $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_G)'$.

Under the model defined in (56), the likelihood function is

$$\begin{aligned}p(\mathbf{D}^{\text{obs}} | N_1, \ldots, N_G, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\rho}) = \prod_{g=1}^{G} \frac{N_g!}{(N_g - n_{g,\text{obs}})!} \\ \prod_{k, y \neq (0,0)} \left(\phi_{g,y}(\mathbf{x}_{[k]}, \boldsymbol{\beta}, \rho_g)\ \theta_{g,k}\right)^{M_{g,k,y}}\ p_{g,(0,0)}(\boldsymbol{\beta}, \boldsymbol{\theta}_g, \rho_g)^{(N_g - n_{g,\text{obs}})}.\end{aligned} \tag{57}$$

The derivation of (57) is given in Appendix A.4 of Supplementary Material (Graham et al., 2023), along with the derivation for the case without the restriction to categorical covariates and the product multinomial model for the joint covariate distribution.

For $G > 2$ a multinomial prior could be adopted for the sub-population counts, $N_1, \ldots, N_G$, constrained by the total population count N, so that

$$p(N_1, \ldots, N_G | N, \boldsymbol{\varpi}) = \frac{N!}{N_1! \ \ldots \ N_G!} \prod_g \varpi_g^{N_g},$$

where $\boldsymbol{\varpi} = (\varpi_1, \ldots, \varpi_G)$, and a hyperprior is placed on the total population count.

The prior for the covariate distribution parameters could be used to encourage pooling of information over the G sub-populations through a hierarchically structured prior of the form,

$$p(\boldsymbol{\theta}|\boldsymbol{\zeta}) = \prod_g p(\boldsymbol{\theta}_g|\boldsymbol{\zeta}),$$

accompanied by a prior on the hyper-parameters, $\boldsymbol{\zeta}$. However, in the simple illustration reported below we adopted a simple independence prior

$$p(\boldsymbol{\theta}) = \prod_g p(\boldsymbol{\theta}_g),$$

and Dirichlet prior distributions for each of the G sub-population covariate distribution parameter vectors.

To verify that introducing prior information on specific sub-populations allows dependence odds ratios specific to those sub-populations to be estimated, we consider a simple example based on a similar data structure to that considered in Section 7.2, but assume prior information is available for both male and female population counts. These data were generated under an assumption of a dependence odds ratio equal to five, for all age-groups for both males and females. In addition, we consider a second simulated dataset, generated, with the same sex by age population totals as the data used in Section 7.2, but generated under an assumption of conditional independence for all ages within the female group and a common dependence odds ratio of five for all ages within the male group. The latter data enables us to check, that given strong prior information on the female and male population sizes, the dual systems estimation model can adapt to estimate very different common odds ratios for the female and male groups.

We specified priors for the female and male population totals, via the sex ratio and the population total. There has been some interest in the literature in the use of known sex ratios in dual systems estimation, since these may be stable demographic parameters that could be reliably estimated prior to a dual systems estimation (Wolter, 1990; Bell, 1993; Elliot and Little, 2000). Elliot and Little (2000) present a Bayesian model which integrates known sex-ratios with

dual systems estimation using a bespoke model in which observed counts are modelled as Normal random variables conditional on the unobserved true counts. In the Elliot and Little (2000) model, no specific prior information on the total population size is included in the analysis but estimation is carried out under the assumption that conditional independence holds for the female group, but not for the male group. Models with the same assumption were also considered by Wolter (1990) and Bell (1993), from a frequentist perspective. Consequently, in addition to models with an informative prior for the population total and the sex-ratio, we consider models that assume conditionally independent inclusion for females, in conjunction with an informative prior for the sex-ratio and an uninformative prior for the female population total. The simulated population of 1,000,000 comprised 513,975 females and 486,025 males. Thus, the true sex ratio (male to female) for the simulated population was $r = N_{\text{males}}/N_{\text{females}} = 0.95$. As in previous examples, the (simulated) observed data was obtained from the true population by omitting records in the $(0, 0)$ cells for each combination of sex and age.

We fitted several models, corresponding to different prior assumptions for the sex-ratio and total populations size. Models were fitted in Stan using the full likelihood given by (57). Similarly to previous analyses, we adopted independent Dirichlet(**0.01**) priors for the sex-specific, age distribution parameters, and independent Normal(0,10) priors for all list coverage model parameters. We adopted independent truncated normal priors for the odds ratio parameters, specifically $\text{TNorm}(1, 2.19, 0.001, 9.999)$, for both male and female odds ratios, for all models except for the models where independence was assumed for females. Although the prior mode for the $\text{TNorm}(1, 2.19, 0.001, 9.999)$ prior is equal to 1, the prior mean is 2.16 and the prior median is 1.92, because of the asymmetry of the truncation points. Under the $\text{TNorm}(1, 2.19, 0.001, 9.999)$ prior, the prior probability that the odds ratio exceeds 5 is 0.05. Thus, the prior is located well away from the underlying true value of 5, and therefore, provides a useful test of the informative-ness of data with respect to the odds ratio: Informative data will move the posterior away from the prior towards the true value of 5. The $\text{TNorm}(1, 2.19, 0.001, 9.999)$ prior for the male and female dependence odds ratios is intended to represent a situation where an analyst believes independence is a plausible assumption but wishes to allow for the possibility that independence does not hold, and is open to the possibility of both negative and positive dependence. Other prior choices could be made, the log-normal being a natural alternative to the truncated Normal.

The results reported below are based on posterior samples obtained by thinning the last 900 draws from each of five parallel HMC chains by a factor of three to yield a nominal posterior sample size of 1500. $\hat{R}$ statistics were less than 1.02 for all parameters and effective posterior sample sizes were at least several hundred, though the median (over parameters for a given model) was close to the nominal size of 1500 for each of the models considered. We used burn-in periods of 1,000 or 2,000 for models with fixed N or a prior on N, respectively.

Some key summaries are reported in Table 4 for the data generated under the assumption of conditionally independent list inclusion for females and a

dependence odds ratio of 5 for males ($\rho_{\text{female}} = 1$, $\rho_{\text{male}} = 5$), and in Table 5 and Figure 13, for the data generated assuming a dependence odds ratio of 5 for both females and males ($\rho_{\text{female}} = \rho_{\text{male}} = 5$).

TABLE 4

Posterior summaries for male and female dependence odds ratios, population counts and sex-ratio (r) under alternative prior settings. The models are all fitted to simulated data generated under the assumption that dependence odds ratio is equal to 5 for males (ρ_{male}) and equal to 1 for females (ρ_{female}). N denotes the total population size, and N_{males} and N_{females} denote male and female counts, respectively. The covariate structure of the data comprises 180 sex-age combinations.

prior	Estimand[(a)]	2.5%	50%	97.5%
$\rho_{\text{male}} \sim \text{TNorm}(1, 2.190, 0.001, 9.999)$	ρ_{male}	4.90	4.96	5.02
$\rho_{\text{female}} \sim \text{TNorm}(1, 2.190, 0.001, 9.999)$	ρ_{female}	0.99	1.00	1.02
$r = 0.95$	r	0.95	0.95	0.95
$N = 1{,}000{,}000$	N_{male}	486,025	486,025	486,025
	N_{female}	513,975	513,975	513,975
$\rho_{\text{male}} \sim \text{TNorm}(1, 2.190, 0.001, 9.999)$	ρ_{male}	4.50	4.94	5.38
$\rho_{\text{female}} \sim \text{TNorm}(1, 2.190, 0.001, 9.999)$	ρ_{female}	0.66	0.99	1.29
$r = 0.95$	r	0.95	0.95	0.95
$N \sim \text{Normal}(1000000, 7812.5)$	N_{male}	477,889	485,871	493,122
	N_{female}	505,168	513,606	521,270
$\rho_{\text{male}} \sim \text{TNorm}(1, 2.190, 0.001, 9.999)$	ρ_{male}	4.47	4.94	5.38
$\rho_{\text{female}} \sim \text{TNorm}(1, 2.190, 0.001, 9.999)$	ρ_{female}	0.64	1.00	1.38
$r \sim \text{TNorm}(0.95, 0.01, 0.001, 1.999)$	r	0.93	0.94	0.97
$N \sim \text{Normal}(1000000, 7812.5)$	N_{male}	477,208	485,672	493,980
	N_{female}	504,639	513,930	523,557
$\rho_{\text{male}} \sim \text{TNorm}(1, 2.190, 0.001, 9.999)$	ρ_{male}	4.37	4.92	5.46
$\rho_{\text{female}} = 1$	ρ_{female}	1	1	1
$r \sim \text{TNorm}(0.95, 0.01, 0.001, 1.999)$	r	0.92	0.94	0.96
	N_{male}	475,179	485,372	495,280
$p(N_{\text{female}}) \propto 1/N_{\text{female}}$	N_{female}	513,425	513,869	514,264

[(a)] True values are $\rho_{\text{male}} = 5$, $\rho_{\text{female}} = 1$, $r = 0.95$, $N_{\text{male}} = 486{,}025$, $N_{\text{female}} = 513{,}975$, $N = 1{,}000{,}000$.

For a model with the total population size and sex ratio fixed at their true values ($N = 1{,}000{,}000$, $r = 0.95$), the common odds ratios for both males and females are estimated well with posterior medians close to the true values and narrow 95% credible intervals (Tables 4 and 5, panel 1). The posterior interval for the female odds ratio for the model fitted to the data generated under an assumption of conditionally independent list inclusion for females is particularly narrow, presumably reflecting the fact that the prior mode coincides with the true value in this case. When a modest amount of uncertainty in the total population size is accommodated in the analysis by a normal prior

TABLE 5

Posterior summaries for male and female dependence odds ratios, population counts and the sex-ratio, (r), under alternative prior settings. The models are all fitted to simulated data generated under the assumption of a common dependence odds ratio equal to 5 for all ages, for both the male (ρ_{male}) and female (ρ_{female}) groups. N denotes the total population size, and N_{males} and N_{females} denote male and female counts, respectively. The covariate structure of the data comprises 180 age-sex combinations.

prior	Estimand[(a)]	2.5%	50%	97.5%
$\rho_{\text{male}} \sim \text{TNorm}(1, 2.190, 0.001, 9.999)$	ρ_{male}	4.87	4.94	5.00
$\rho_{\text{female}} \sim \text{TNorm}(1, 2.190, 0.001, 9.999)$	ρ_{female}	4.89	4.97	5.04
$r = 0.95$	r	0.95	0.95	0.95
$N = 1{,}000{,}000$	N_{male}	486,025	486,025	486,025
	N_{female}	513,975	513,975	513,975
$\rho_{\text{male}} \sim \text{TNorm}(1, 2.19, 0.001, 9.999)$	ρ_{male}	4.52	4.90	5.28
$\rho_{\text{female}} \sim \text{TNorm}(1, 2.19, 0.001, 9.999)$	ρ_{female}	4.18	4.90	5.59
$r = 0.95$	r	0.95	0.95	0.95
$N \sim \text{Normal}(1000000, 7812.5)$	N_{male}	478,244	485,431	492,193
	N_{female}	505,544	513,140	520,289
$\rho_{\text{male}} \sim \text{TNorm}(1, 2.19, 0.001, 9.999)$	ρ_{male}	4.41	4.90	5.38
$\rho_{\text{female}} \sim \text{TNorm}(1, 2.19, 0.001, 9.999)$	ρ_{female}	3.94	4.86	5.77
$r \sim \text{TNorm}(0.95, 0.01, 0.001, 1.999)$	r	0.93	0.95	0.97
$N \sim \text{Normal}(1000000, 7812.5)$	N_{male}	476,168	485,399	494,512
	N_{female}	502,989	512,768	522,165
$\rho_{\text{male}} \sim \text{TNorm}(1, 2.19, 0.001, 9.999)$	ρ_{male}	2.36	2.85	3.32
$\rho_{\text{female}} = 1$	ρ_{female}	1	1	1
$r \sim \text{TNorm}(0.95, 0.01, 0.001, 1.999)$	r	0.93	0.95	0.96
	N_{male}	437,356	446,665	455,434
$p(N_{\text{female}}) \propto 1/N_{\text{female}}$	N_{female}	471,824	472,069	472,309

[(a)] True values are $\rho_{\text{male}} = 5$, $\rho_{\text{female}} = 5$, $r = 0.95$, $N_{\text{male}} = 486{,}025$, $N_{\text{female}} = 513{,}975$, $N = 1{,}000{,}000$.

centred on the true value with standard deviation 7812.5, which implies the prior probability for the total population is within ±1.0% of the true value of 1,000,000 is 0.8, the true value for the male and female odds ratios are recovered by the models, though with notably wider credible intervals. The 95% credible interval for the total population size is virtually identical to the prior 95% interval (Tables 4 and 5, panel 2). Adding a $\text{TNorm}(0.95, 0.01, 0.001, 1.999)$ prior to the sex ratio leads to a modest further increase in credible interval widths (Tables 4 and 5, panel 3). Increasing the prior standard deviation on the sex ratio to 0.02, led to a further modest increase in posterior uncertainty (results not shown). Centering the priors for the total population size at 990,000 led to a predictable reduction in estimates of both dependence odds ratios and population sizes, though credible intervals still included the true values (results

not shown).

We also fitted a model which assumed independence for females but not for males, along with a TNorm$(0.95, 0.01, 0.001, 1.999)$ prior for the sex ratio. For this model, we adopted the Jeffreys prior for the female population size. Unsurprisingly, when fitted to the data generated under an assumption of independence for females (Table 4, panel 4) this model recovered the true male and female population totals, male dependence odds ratio and sex ratio. Age-specific estimates were also well estimated: for example, 95.2% of the female credible intervals and 88.9% of the male credible intervals included the true age-specific counts. However, when fitted to the dataset generated under the assumption of a common odds ratio of five for all sex by age groups, the dependence odds ratio for males and both the male and female population totals were badly underestimated (Table 5, panel 4). The underestimation for males, may be partly attributable to the small prior variance adopted for the sex-ratio parameter. The assumption of conditional independence for females leads to underestimation for the female group, and a tight prior on the sex ratio could therefore be expected to lead to underestimation of the male population. However, in fact, the location of the posterior distributions did not change appreciably when the prior standard deviation on the sex ratio was increased, first to 0.02 and then to 0.05 (results not shown). However, in the latter case, the posterior variance for both the sex ratio and the male population count increased markedly. Under the TNorm$(0.95, 0.05, 0.001, 1.999)$ prior for the sex ratio parameter, r, the equal-tail-area 95% credible intervals for r, ρ_{male} and N_{male} were (0.87 to 1.03), (0.90 to 4.90) and $(409,747$ to $484,996)$, respectively. These are all considerably wider than the corresponding results obtained under the TNorm$(0.95, 0.01, 0.001, 1.999)$ prior reported in Table 5 (panel 4). In contrast, posterior estimates for the female population total were virtually unchanged by increasing the prior variance for the sex ratio.

Consequences of erroneously assuming conditional independence for the female group are shown in Figure 13 (panel (d)) for age-specific population estimates of males, where it is clear that assuming conditional independence for females leads to underestimation of the population for males at all age groups. A similar pattern holds for females, though with notably narrower credible intervals for age-specific population estimates, resulting from the female population being estimated under an assumption of independence. For both the male and female groups, the incorrect assumption of conditional independence for the female group resulted in none of the 95% credible intervals for population counts by age including the true value.

We emphasise these results pertain to analysis of a single simulated dataset. A proper simulation study involving analysis of repeated draws of data from the simulation model would be required to quantify bias in the estimates. Nevertheless, the results corresponding to the assumption of independence for females are instructive and illustrative of the likely impact of erroneous independence assumptions, when in fact there is appreciable dependence between inclusion on the two lists. The analysis illustrates the potential for recovering information on dependence odds ratios, given prior information on certain sub-population

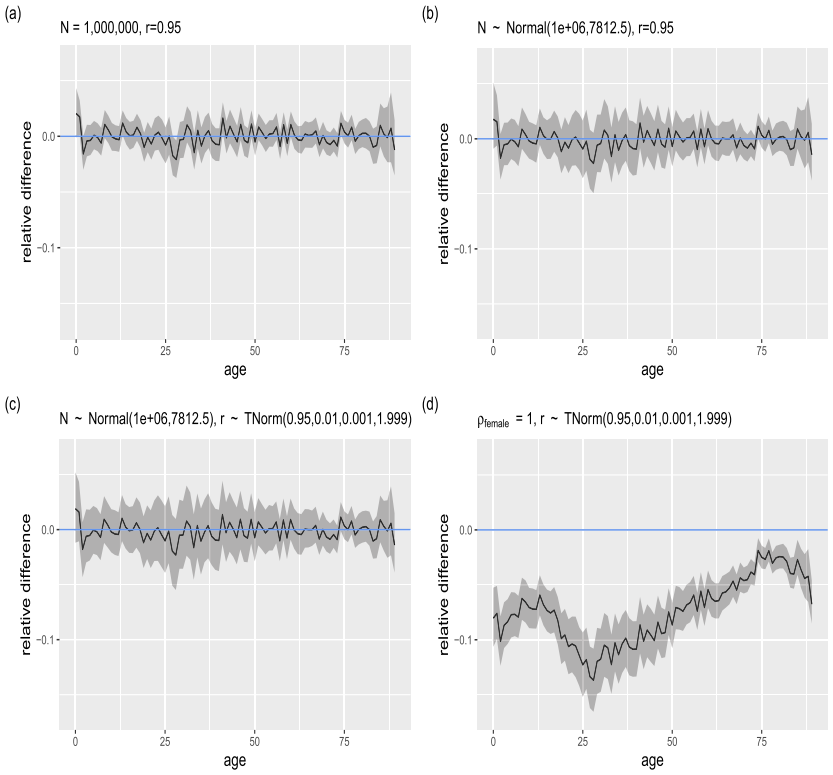

FIG 13. *Posterior medians (black line) and* 95% *credible intervals (grey shading), for counts by age, for males, expressed as relative differences from the true value, under different prior assumptions for population size,* N*, sex-ratio,* r*, and the dependence odds ratio,* ρ*. The blue horizontal line indicates agreement between estimated and true population counts. All models were fitted to data generated under an assumption of common dependence odds ratio of 5, for all sex by age combinations. In panel (a),* N *and* r *are both set to their true values. Independent* $\mathrm{TNorm}(1, 2.19, 0.001, 9.999)$ *priors are specified for the male and female dependence odds ratios. Estimated age-specific counts agree closely to the true values. In panel (b), adding a normal prior for* N *centred at the true value, with modest standard deviation, leads to noticeable increases in posterior uncertainty, but credible intervals remain approximately centred on true values. In panel (c), a truncated normal prior for* r*, centred on the true value, with small standard deviation leads to a further small increase in posterior uncertainty. In panel (d), conditional independence of list inclusion is assumed for the female group, Jeffreys' prior is adopted for the total female population, a truncated normal prior tightly centred on the true value is adopted for the sex-ratio and a* $\mathrm{TNorm}(1, 2.19, 0.001, 9.999)$ *prior is adopted for the male dependence odds ratio. Under this model age-specific estimates for males appear badly biased with none of the age-specific credible intervals including the true value.*

sizes. It seems that dependence odds ratios within sub-populations for which good prior information on population size is available can be reliably estimated. However, modest amounts of prior uncertainty in the aggregate totals leads to considerably more posterior uncertainty in population estimates than is seen for the independence model.

7.4. Summarising what can be done to relax the independence assumption

The non-identifiability of the association between inclusion on the two lists is an inherent feature of the structure of the dual systems estimation problem. Absent knowledge of population size, we can never directly compare the proportion of List 2 individuals captured by List 1 with the proportion of List 2 non-captures captured by List 1 and vice versa. We can, however, place a prior on some measure of association between inclusion on the two lists. In principle, Bayesian inference then proceeds as usual to produce a posterior distribution (Lindley, 1972, p. 46, footnote 34). However, our examples indicate that unless a very tight prior is placed on the association parameter, inferences are markedly less precise than under the conditional independence model. In addition, working with an unidentified likelihood appears to cause numerical issues, even with a moderately informative prior, and some care has to be taken with model-fitting. Nevertheless, if there is uncertainty about the assumption of conditional independence either a sensitivity analysis in which the value of an association parameter is varied in a series of analyses, or a fully Bayesian analysis with an informative prior on the association parameter, provide methods for obtaining a realistic representation of uncertainty concerning population estimates.

If external information is available on the dependence odds ratios, as in Brown, Abbott and Diamond (2006), it can be used to inform the prior for the dependence odds ratio. Alternatively, if strong priors on the population size at some levels of aggregation of covariate combinations can be assumed it is possible to estimate dependence odds ratios that are assumed to be homogeneous within those levels of aggregation. Nevertheless, inferences remain sensitive to the priors for sub-population totals. Our analysis of the model that assumes conditional independence for females but allows the dependence odds ratio for males to be estimated illustrates the danger of erroneous prior assumptions of conditional independence, which can lead to substantial bias. When reliable prior information on sub-population totals is available, our example suggests it will often be safer to include this information in dual systems estimation and allow dependence to be estimated, than to assume conditional independence.

Ultimately, however, if list dependence is a major concern, moving beyond dual systems estimation to a multiple list approach is likely to reduce sensitivity to prior assumptions and allow the most flexibility in modelling list dependence. Though this lies outside the scope of this paper, we note that, with minor changes in notation, the likelihood derivations in Section 3 hold for the multiple list case. If we let $\mathbf{0}$ denote the cell in the multiway cross tabulation of list inclusion indicators corresponding to being missed by all lists and let

$$p_{\mathbf{0}}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \int \Pr(Y = \mathbf{0} | \mathbf{X} = x, \boldsymbol{\beta})\, p_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta})\, \mathrm{d}\mathbf{x} \tag{58}$$

denote the population averaged probability of that event, then (14) and (23) hold for the multiple list case, once the sample space of the inclusion cell, Y, is expanded to accommodate inclusion combinations appropriate for the multiple

list case: for example, with three lists the possible inclusion combinations are $\mathcal{Y} = \{(1,1,1),(1,1,0),(1,0,1),(0,1,1),(1,0,0),(0,1,0),(0,0,1),(0,0,0)\}$, with the last combination corresponding to the **0** cell defined above.

Directly extending the approach to incorporating dependence in dual systems estimation considered in Sections 7.2 and 7.3 above, to the multiple list case is difficult because, with multiple lists, a distribution over the cells of a multiway table may not exist for all combinations of marginal probabilities and odds ratios (Lee, 1997). However, by introducing a latent multivariate normal variable from which binary list inclusion indicators are derived as exceedances of a threshold (or cut-point), copula-style formulations whereby, joint models are constructed from marginal models for list coverage and models for dependence may be possible. Zwane and van der Heijden (2005) take a different route to modelling dependence in a multiple list situations by modelling the joint distribution of inclusion cells using a particular formulation of the multinomial-logistic regression model. However, in this approach, the marginal list coverage probabilities are not directly modelled.

A promising alternative for modelling dependence in multiple list settings is the latent class formulation considered in Manrique-Vallier (2016), whereby independence is assumed conditional on a latent variable. Combined with a Dirichlet Process prior on the latent class variable, which avoids the need to pre-specify the number of latent classes, this produces a flexible framework for modelling dependence.

We also note the log-linear modelling approach to dual systems estimation that we discuss in Section 8 extends very naturally to the multiple list situations.

8. Connections with log-linear modelling of capture-recapture data

Following introductions by Fienberg (1972) and Cormack (1989), log-linear modelling has proved a popular approach to population estimation from capture-recapture data, particularly when multiple population listings are available. In that case, log-linear models permit flexible modelling of the dependence between inclusion on lists. In dual systems estimation, there is less scope for modelling list dependence, but when all covariates are categorical, log-linear modelling can be used to model the joint distribution of covariates and list inclusion indicators (Tilling and Sterne, 1999; van der Heijden et al., 2018). This represents a different modelling strategy to the approach described so far, which separately models the joint covariate distribution and the conditional distribution of list inclusion indicators given the covariates. Henceforth, in order to distinguish the modelling approach developed in previous sections from the log-linear models discussed in this section, we refer to the former as the "logistic-multinomial" model, reflecting the logistic models for the marginal list inclusion probabilities, the multinomial model for the joint distribution of the list inclusion indicators conditional on covariates and, in the case of categorical covariates, the multinomial model for the covariate distribution.

There are close connections between the logistic-multinomial model developed in sections 2 to 7 and log-linear modelling for capture-recapture data, and we

now consider these connections, by extending the discussion of Huggins and Hwang (2011, p.3883) to include covariates and considering log-linear modelling from a Bayesian perspective.

Although originally proposed for the analysis of capture-recapture studies in the context of a multinomial model (Fienberg, 1972), it has become common to fit log-linear models under an assumption of independent Poisson cell counts (Cormack, 1989; Leclerc et al., 2014; van der Heijden et al., 2018, 2022). As noted by several authors (Sandland and Cormack, 1984; Cormack and Jupp, 1991; Huggins and Hwang, 2011; Tilling and Sterne, 1999), the two formulations are closely related, with the most important difference being the greater variance of the total population estimate under the Poisson model (Sandland and Cormack, 1984; Cormack and Jupp, 1991). In frequentist applications of log-linear modelling to capture-recapture data, parameter and population point estimates are sometimes obtained by maximising a Poisson log-likelihood, while confidence intervals are obtained under an assumption of multinomial sampling (Baillargeon et al., 2007; van der Heijden et al., 2012; Cormack and Jupp, 1991, p. 914). At first glance, such manoeuvres would appear to have no parallel in Bayesian inference since Bayesian inference follows directly from the posterior distribution of unknowns obtained under a specified model for the data. However, in Section 8.2 we discuss a Bayesian analogue of this procedure, suggested by well-known connections between Poisson and multinomial models (Gelman et al., 2014, p. 426).

8.1. Multinomial log-linear models

Using the notation established in Section 3.2, for the case where all covariates are categorical, suppose there are K possible covariate combinations indexed by $k \in \{1, \ldots, K\}$ with the k^{th} combination defined by $\mathbf{x}_{[k]}$. The count of people in the population with $\mathbf{X} = \mathbf{x}_{[k]}$ and list inclusion cell $Y = y$ is denoted by $M_{k,y}$. We let $\mathbf{M}_{.,(0,0)} = \{M_{k,(0,0)}, k \in \{1, \ldots, K\}\}$ denote the unobserved counts for the group missed by both lists, by covariate combination, $\mathbf{D}^{\text{obs}} = \{M_{k,y}, k \in \{1, \ldots, K\}, y \neq (0,0)\}$, the observed counts by covariate and list inclusion combination, and $\mathbf{M}^{\text{com}} = \{M_{k,y}, k \in \{1, \ldots, K\}, y \in \mathcal{Y}\}$ the complete vector of counts for all $4 \times K$ covariate and list inclusion combinations. It is also useful to recall that $n_{\text{obs}} = \sum_{k,y \neq (0,0)} M_{k,y}$ and $N - n_{\text{obs}} = \sum_k M_{k,(0,0)}$. Let $\eta_{k,y}$ denote the probability that an individual in the target population has covariate combination $\mathbf{x}_{[k]}$ and list inclusion cell y; that is, $\eta_{k,y} = \Pr(\mathbf{X} = \mathbf{x}_{[k]}, Y = y|\boldsymbol{\eta})$ where $\boldsymbol{\eta} = \{\eta_{k,y},\ k \in \{1, \ldots, K\}, y \in \mathcal{Y}\}$. A standard formulation for a multinomial log-linear model for complete data, $\mathbf{M}^{\text{com}}$, is

$$
\begin{aligned}
&[\mathbf{M}^{\text{com}}|N, \boldsymbol{\eta}] \sim \text{Multinomial}(N, \boldsymbol{\eta}) \\
&\log(\eta_{k,y}) = \lambda_0 + \lambda_k^{\mathbf{X}} + \lambda_{l_1(y)}^{L_1} + \lambda_{l_2(y)}^{L_2} + \lambda_{k,l_1(y)}^{\mathbf{X}L_1} + \lambda_{k,l_2(y)}^{\mathbf{X}L_2};\ k \in \{1, \ldots, K\}, y \in \mathcal{Y}
\end{aligned}
\tag{59}
$$

where the superscripts are simply notational devices to indicate parameters corresponding to particular covariates or list inclusion indicators or combinations

thereof. The L_1 and L_2 superscripts refer to List 1 and List 2 inclusion, respectively, and $l_1(y)$ and $l_2(y)$ denote specific values of the two list indicators; we have used a functional notation to emphasize their dependence on the list inclusion cell y, for instance, $l_1(y) = 1$, when $y \in \{(1,1),(1,0)\}$, and $l_1(y) = 0$, otherwise. The parameter $\lambda_k^{\mathbf{X}}$ is a linear combination of parameters associated with the values of the variables that define the k^{th} covariate category. $\lambda_k^{\mathbf{X}}$ will usually include product terms representing interactions. As an illustration, in the case of two covariates, X_1 and X_2, modelled with an interaction, $\lambda_k^{\mathbf{X}}$ would have the form

$$\lambda_k^{\mathbf{X}} = \lambda_{k[X_1]}^{X_1} + \lambda_{k[X_2]}^{X_2} + \lambda_{k[X_1X_2]}^{X_1X_2},$$

where the notation $k[.]$ is used to refer to the specific category of the indicated variable or variables that correspond to the covariate combination k. For example, if k refers to the covariate combination $X_1 = 1$ and $X_2 = 2$, then $k[X_1] = 1, k[X_2] = 2$, and $k[X_1X_2] = 1, 2$. Similarly, the parameters $\lambda_{k,l_1(y)}^{\mathbf{X}L_1}$ and $\lambda_{k,l_2(y)}^{\mathbf{X}L_2}$ expand to a linear combination of specific covariate-list inclusion interactions. In (59), the parameter λ_0 is not a free parameter but is used to ensure the sum of probabilities over the covariate and list inclusion combinations sum to one (Schafer, 1997, chapter 8).

For identifiability, we let

$$\lambda_0^{L1} = \lambda_0^{L2} = \lambda_{k,0}^{\mathbf{X}L1} = \lambda_{k,0}^{\mathbf{X}L2} = 0. \tag{60}$$

Some identifiability restrictions are also required for the components of $\lambda_k^{\mathbf{X}}, k \in \{1,\ldots,K\}$. These restrictions may set the parameter value associated with some reference category of a covariate to zero or require that the parameters sum to zero over the indexes of each covariate.

In (59), notice that there are no superscripts involving both L_1 and L_2, indicating that the model includes no interaction terms involving L_1 and L_2. This is the log-linear model equivalent of the assumption of conditionally independent list inclusion. It could be weakened by including an L_1, L_2 interaction, such as $\lambda_{l_1(y),l_2(y)}^{L_1L_2}$, however, a model including an L_1, L_2 interaction would give rise to the same identifiability issues encountered by the dependent logistic-multinomial model, considered in Section 7, because no cells with $L_1 = 0$ and $L_2 = 0$ are observed.

It may be desirable to model the effects of covariates with a large number of levels (such as small area geography, or single year of age) hierarchically.[1] For example, parameters indicating small geographic areas could be modelled as draws from a normal model with mean dependent on area-level covariates. Priors would then need to be specified for parameters of the hierarchical model.

The likelihood for the multinomial log-linear model is obtained by marginalising the complete data likelihood over the unobserved counts $\mathbf{M}_{.,(0,0)}$, The

[1] We use the term hierarchical in the multilevel modelling sense. In the log-linear modelling literature, the term hierarchical models often refers to single-level models which respect the constraint that whenever an interaction term is included in the model, the main effect terms for the effects involved in the interaction are also included.

complete data likelihood implied by (59), is

$$p(\mathbf{M}^{\text{com}}|N,\boldsymbol{\eta}) = \left[\frac{N!}{\prod_{k,y\neq(0,0)} M_{k,y}!}\prod_{k,y\neq(0,0)}\eta_{k,y}^{M_{k,y}}\right]\left[\frac{1}{\prod_k M_{k,(0,0)}!}\prod_k \eta_{k,(0,0)}^{M_{k,(0,0)}}\right]. \tag{61}$$

Only the second square-bracketed term in (61) involves $\mathbf{M}_{.,(0,0)}$ and this component of the complete data likelihood can be written as

$$\frac{1}{\prod_k M_{k,(0,0)!}}\prod_k \eta_k^{M_{k,(0,0)}} = \frac{(p^{\text{M}}_{(0,0)}(\boldsymbol{\eta}))^{(N-n_{\text{obs}})}}{(N-n_{\text{obs}})!}\left[\frac{(N-n_{\text{obs}})!}{\prod_k M_{k,(0,0)}!}\prod_k\left(\frac{\eta_{k,(0,0)}}{p^{\text{M}}_{(0,0)}(\boldsymbol{\eta})}\right)^{M_{k,(0,0)}}\right], \tag{62}$$

where $p^{\text{M}}_{(0,0)}(\boldsymbol{\eta}) = \sum_k \eta_{k,(0,0)}$ is the marginal probability of not being recorded on at least one list, under the multinomial log-linear model. Since the square bracketed term in (62) is the multinomial probability mass function for $M_{.,(0,0)}$, with size parameter $(N-n_{\text{obs}})$ and k^{th} element of the probability vector given by $\eta_{k,(0,0)}/p^{\text{M}}_{(0,0)}(\boldsymbol{\eta})$, marginalising (61) over the unobserved counts, $\mathbf{M}_{.,(0,0)}$, to obtain the observed data likelihood for the multinomial log-linear model gives

$$p(\mathbf{D}^{\text{obs}}|N,\boldsymbol{\eta}) = \frac{N!}{(N-n_{\text{obs}})!}p^{\text{M}}_{(0,0)}(\boldsymbol{\eta})^{(N-n_{\text{obs}})}\frac{1}{\prod_{k,y\neq(0,0)} M_{k,y}!}\prod_{k,y\neq(0,0)}^{M_{k,y}}\eta_{k,y}^{M_{k,y}}$$

$$= \left\{\left[\binom{N}{n_{\text{obs}}}(1-p^{\text{M}}_{(0,0)}(\boldsymbol{\eta}))^{n_{\text{obs}}}p^{\text{M}}_{(0,0)}(\boldsymbol{\eta})^{(N-n_{\text{obs}})}\right]\times \left[\frac{n_{\text{obs}}!}{\prod_{k,y\neq(0,0)} M_{k,y}!}\prod_{k,y\neq(0,0)}\left(\frac{\eta_{k,y}}{(1-p^{\text{M}}_{(0,0)}(\boldsymbol{\eta}))}\right)^{M_{k,y}}\right]\right. \tag{63}$$

$$\propto \left\{\left[\binom{N}{n_{\text{obs}}}(1-p^{\text{M}}_{(0,0)}(\boldsymbol{\eta}))^{n_{\text{obs}}}p^{\text{M}}_{(0,0)}(\boldsymbol{\eta})^{(N-n_{\text{obs}})}\right]\times \left[\prod_{k,y\neq(0,0)}\left(\frac{\eta_{k,y}}{(1-p^{\text{M}}_{(0,0)}(\boldsymbol{\eta}))}\right)^{M_{k,y}}\right]\right. . \tag{64}$$

Equation (64) shows the likelihood for the multinomial log-linear model is the product of a binomial probability for the number of individuals recorded on at least one list and a conditional multinomial probability for the distribution of observable counts over covariate and list inclusion cell combinations, conditional on being observed. Therefore, the likelihood for the multinomial log-linear dual systems model is proportional to the product of a binomial probability for n_{obs} and a conditional likelihood based only on the data from individuals observed on at least one list. Since the logistic-multinomial model for an aggregated data structure implies

$$\begin{aligned}\eta_{k,y} &= \Pr(Y=y|\mathbf{X}=\mathbf{x}_{[k]},\boldsymbol{\beta})\Pr(\mathbf{X}=\mathbf{x}_{[k]}|\boldsymbol{\theta})\\ &= \phi_y(\mathbf{x}_{[k]},\boldsymbol{\beta})\,\theta_k;\ ;k\in\{1,\ldots,K\},y\in\mathcal{Y},\end{aligned} \tag{65}$$

it is easily verified that substituting for $\eta_{k,y}$ in (63) and (64) reproduces the corresponding forms of the logistic-multinomial likelihood given by (24) and (25).

Following the approach of Huggins and Hwang (2011), the multinomial log-linear model and the logistic-multinomial model can be further related by noting the latter implies

$$\begin{aligned}\eta_{k,y} &= \begin{cases} \theta_k \left(\tilde{\phi}_1(\mathbf{x}_{[k]}, \boldsymbol{\beta}_1)\right)^{l_1(y)} \left(\tilde{\phi}_2(\mathbf{x}_{[k]}, \boldsymbol{\beta}_2)\right)^{l_2(y)} \times \\ \quad \left(1 - \tilde{\phi}_1(\mathbf{x}_{[k]}, \boldsymbol{\beta}_1)\right)^{(1-l_1(y))} \left(1 - \tilde{\phi}_2(\mathbf{x}_{[k]}, \boldsymbol{\beta}_2)\right)^{(1-l_2(y))}, \end{cases} \\ &= \begin{cases} \theta_k \left(1 - \tilde{\phi}_1(\mathbf{x}_{[k]}, \boldsymbol{\beta}_1)\right) \left(1 - \tilde{\phi}_2(\mathbf{x}_{[k]}, \boldsymbol{\beta})\right) \times \\ \left(\dfrac{\tilde{\phi}_1(\mathbf{x}_{[k]}, \boldsymbol{\beta}_1)}{1 - \tilde{\phi}_1(\mathbf{x}_{[k]}, \boldsymbol{\beta}_1)}\right)^{l_1(y)} \left(\dfrac{\tilde{\phi}_2(\mathbf{x}_{[k]}, \boldsymbol{\beta}_2)}{1 - \tilde{\phi}_2(\mathbf{x}_{[k]}, \boldsymbol{\beta}_2)}\right)^{l_2(y)} \end{cases}, \; k \in \{1, \ldots, K\}. \end{aligned}$$

Consequently,

$$\log(\eta_{k,y}) = \gamma_k + l_1(y) \times \log\left(\frac{\tilde{\phi}_1(\mathbf{x}_{[k]}, \boldsymbol{\beta}_1)}{1 - \tilde{\phi}_1(\mathbf{x}_{[k]}, \boldsymbol{\beta}_1)}\right) + l_2(y) \times \log\left(\frac{\tilde{\phi}_2(\mathbf{x}_{[k]}, \boldsymbol{\beta}_2)}{1 - \tilde{\phi}_2(\mathbf{x}_{[k]}, \boldsymbol{\beta}_2)}\right), \tag{66}$$

where $\gamma_k = \log(\theta_k) + \log\left(1 - \tilde{\phi}_1(\mathbf{x}_{[k]}, \boldsymbol{\beta}_1)\right) + \log\left(1 - \tilde{\phi}_2(\mathbf{x}_{[k]}, \boldsymbol{\beta}_2)\right)$ is the log of the joint probability that an individual in the target population has covariate combination $\mathbf{x}_{[k]}$ and is missed by both lists. In the notation of (59), with the identifiability constraints (60), $\gamma_k = \lambda_0 + \lambda_k^{\mathbf{X}}$. Note also that in the logistic-multinomial parameterisation

$$\log\left(\frac{\tilde{\phi}_j(\mathbf{x}_{[k]}, \boldsymbol{\beta}_j)}{1 - \tilde{\phi}_j(\mathbf{x}_{[k]}, \boldsymbol{\beta}_j)}\right) = \mathbf{x}'_{[k]}\boldsymbol{\beta}_j, \; j \in \{1, 2\},$$

is the logistic coverage model for inclusion on List j, where $\mathbf{x}_{[k]}$ is defined to include an intercept term and interaction and non-linear terms as required.

Comparing (66) with (59) and equating γ_k with $\lambda_0 + \lambda_k^{\mathbf{X}}$, and $\mathbf{x}'_{[k]}\boldsymbol{\beta}_j$ with $\lambda_1^{L_j} + \lambda_{k,1}^{\mathbf{X}L_j}$, for $j \in \{1, 2\}$, it is clear that the logistic coverage models can be recovered from the log-linear model. For example, $\lambda_1^{L_j}$ can be interpreted as the intercept of the logistic coverage model for List j. The logistic-multinomial model developed in preceding sections places no restriction on $\boldsymbol{\theta}$ except that it's components must sum to one. In the log-linear modelling context, this corresponds to a saturated model for the covariate distribution, meaning that terms for the highest order interaction between covariates and for all lower order interactions are included in $\lambda_k^{\mathbf{X}}$. Simpler log-linear models could be considered, and, analogously, the covariate probabilities, $\boldsymbol{\theta}$, in the logistic-multinomial model could also be modelled. In view of the connections between the multinomial log-linear and logistic-multinomial model structures, and the similarity of the likelihoods, it seems the multinomial log-linear model can be viewed as a re-parameterisation of the logistic-multinomial model for an aggregated data structure.

Despite the close connection between the multinomial log-linear and logistic-multinomial models, posterior inferences obtained under the two models are not guaranteed to be identical because, in view of the different parameterisations employed, prior specifications for the two models may differ. For the logistic-multinomial model, we assumed *a priori* independence for the three parameter blocks pertaining to the population total, covariate distribution and list inclusion models, that is $p(N, \boldsymbol{\theta}, \boldsymbol{\beta}) = p(N)\, p(\boldsymbol{\theta})\, p(\boldsymbol{\beta})$. While this structure could, in principle, be used to induce a prior on the cell probabilities of the multinomial log-linear model, using the relation $\eta_{k,y} = \phi_y(\mathbf{x}_{[k]}, \boldsymbol{\beta})\, \theta_k$, which, in turn, induces a prior on the log-linear model parameters $\boldsymbol{\lambda}$, a more natural approach from a log-linear modelling viewpoint is to specify a prior on $\boldsymbol{\lambda}$ directly. A multivariate normal is one candidate. An alternative strategy is to specify a constrained Dirichlet prior distribution for the cell probabilities, where the constraint is that the cell probabilities must satisfy the log-linear model. Since the log-linear model for the cell probabilities (59), can be written in the form

$$\log(\boldsymbol{\eta}) = \mathbf{W}\boldsymbol{\lambda}, \tag{67}$$

where $\boldsymbol{\lambda}$ is the full vector of log-linear model parameters and $\mathbf{W}$ is a design matrix, the constrained Dirichlet prior can be specified as

$$p(\boldsymbol{\eta}) \propto \begin{cases} \prod_{k,y} \eta_{k,y}^{(\alpha_{k,y}-1)}; & \log(\boldsymbol{\eta}) = \mathbf{W}\boldsymbol{\lambda}, \text{ for some } \boldsymbol{\lambda} \\ 0; & \log(\boldsymbol{\eta}) \neq \mathbf{W}\boldsymbol{\lambda}, \text{ for any } \boldsymbol{\lambda}. \end{cases} \tag{68}$$

The use of this prior for Bayesian log-linear modelling is discussed in Gelman et al. (2014, pp. 428–431) and Schafer (1997, Chapter 8). In complete data applications, a constrained Dirichlet prior for $\boldsymbol{\eta}$ implies the posterior for $\boldsymbol{\eta}$ is also a constrained Dirichlet distribution. A recent Bayesian application of multinomial log-linear modelling of multiple list capture-recapture data, employing the constrained-Dirichlet prior is given by Di Cecco, Di Zio and Liseo (2020b).

8.2. Poisson log-linear models

Under the Poisson log-linear model cell counts are assumed to be conditionally independent Poisson random variables, given the model parameters. Thus, a Poisson log-linear model for small domain dual systems estimation can be written as

$$\begin{aligned} &[M_{k,y}|\boldsymbol{\mu}] \overset{\text{indep}}{\sim} \text{Poisson}(\mu_{k,y}); \quad k \in \{1, \ldots, K\}, y \in \mathcal{Y} \\ &\log(\mu_{k,y}) = \xi_0 + \xi_k^{\mathbf{X}} + \xi_{l_1(y)}^{L_1} + \xi_{l_2(y)}^{L2} + \xi_{k,l_1(y)}^{\mathbf{X}L_1} + \xi_{k,l_2(y)}^{\mathbf{X}L_2}; \quad k \in \{1, \ldots, K\}, y \in \mathcal{Y} \end{aligned} \tag{69}$$

where $\boldsymbol{\mu} = \{\mu_{k,y},\, k \in \{1, \ldots, K\}, y \in \mathcal{Y}\}$, and similar notational conventions to those employed for the multinomial log-linear model (59) are adopted in (69). Some useful functions of the cell means are the expected total population size $\mu_{++} = \sum_{k,y} \mu_{k,y}$, the expected number of people missed by both lists $\mu_{+,(0,0)} =$

$\sum_k \mu_{k,(0,0)}$, and the expected number of people recorded on at least one list, $\mu_{\text{obs}} = \sum_{k,y\neq(0,0)} \mu_{k,y} = \mu_{++} - \mu_{+,(0,0)}$. The Poisson log-linear model implies the marginal probability of being missed by both lists is $p^P_{(0,0)}(\boldsymbol{\mu}) = \mu_{+,(0,0)}/\mu_{++}$.

In contrast to the multinomial log-linear model, the intercept, ξ_0, in the Poisson log-linear model is a free parameter. However, the number of parameters in the multinomial and Poisson log-linear models is the same because the total population size is not an explicit parameter in the standard Poisson log-linear model formulation. We note, however, that from standard properties of the Poisson distribution (Gelman et al., 2014, p. 585), the model of independent Poisson counts implies the total population is Poisson distributed with expectation μ_{++}. Thus, while the Poisson log-linear model does not explicitly condition on the total population size, it implies a Poisson prior for N, conditional on the expected cell counts. Therefore, the unconditional prior for the total population size implied by the Poisson log-linear model is $p(N) = \int \text{Poisson}(N|\mu_{++})\, p(\boldsymbol{\mu})\, \mathrm{d}\boldsymbol{\mu}$, and $p(\boldsymbol{\mu})$ may be specified directly or implied by a prior on the vector of parameters of the log-linear model, $\boldsymbol{\xi}$. Gelman et al. (2014, pp. 428–431) note that, analogously to the multinomial log-linear model, a constrained generalised Dirichlet prior is a potential prior for the parameters of a Poisson log-linear model, where the constraint is that expected cell counts conform to the specified log-linear model. We use the term generalised Dirichlet distribution to refer to a distribution with a density of the same form as the Dirichlet but without the restriction that the elements of the random vector sum to one. Analogously to the multinomial log-linear model, the posterior for the Poisson log-linear parameters, under a constrained generalised Dirichlet prior is also a constrained generalised Dirichlet distribution. Other prior specifications are, of course, possible; a multivariate normal model for $\boldsymbol{\xi}$ is a popular and convenient choice.

The observed data likelihood for the Poisson log-linear model is easily obtained because the assumption of independent cell counts which implies

$$p(\mathbf{M}^{\text{com}}|\boldsymbol{\mu}) = \prod_{k,y} \text{Poisson}(M_{k,y}|\mu_{k,y})$$

and marginalising over the unobservable counts for group missed by both lists ($y = (0,0)$), for each covariate combination gives the likelihood for the observed data:

$$\begin{aligned} p(\mathbf{D}^{\text{obs}}|\boldsymbol{\mu}) &= \prod_{k,y\neq(0,0)} \text{Poisson}(M_{k,y}|\mu_{k,y}) \\ &= \prod_{k,y\neq(0,0)} \frac{1}{M_{k,y}!} \exp(-\mu_{k,y})\, \mu_{k,y}^{M_{k,y}} \\ &= \exp(-\mu_{\text{obs}}) \prod_{k,y\neq(0,0)} \frac{1}{M_{k,y}!}\, \mu_{k,y}^{M_{k,y}} \\ &= \left[\frac{1}{n_{\text{obs}}!} \exp(-\mu_{\text{obs}})\, \mu_{\text{obs}}^{n_{\text{obs}}}\right] \left[\frac{n_{\text{obs}}!}{\prod_{k,y\neq(0,0)} M_{k,y}!} \prod_{k,y\neq(0,0)} \left(\frac{\mu_{k,y}}{\mu_{\text{obs}}}\right)^{M_{k,y}}\right] \end{aligned} \tag{70}$$

$$\propto \left[\frac{1}{n_{\text{obs}}!}\exp(-\mu_{\text{obs}})\,\mu_{\text{obs}}^{n_{\text{obs}}}\right]\left[\prod_{k,y\neq(0,0)}\left(\frac{\mu_{k,y}}{\mu_{\text{obs}}}\right)^{M_{k,y}}\right]. \tag{71}$$

From (70) it is clear that the Poisson log-linear model log-likelihood factors as the product of a Poisson probability for n_{obs}, with expectation μ_{obs}, and multinomial probability for the conditional distribution of the cell counts over the observed cells, and is, therefore, proportional to a Poisson probability for n_{obs} and a conditional likelihood for the log-linear model parameters. Noting that the logistic-multinomial model implies

$$\mu_{k,y} = N\,\phi_y(\mathbf{x}_{[k]},\boldsymbol{\beta})\,\theta_k, \tag{72}$$

it is easily verified that substituting for $\mu_{k,y}$ in (70) and (71) leads to the aggregate data logistic-multinomial model likelihood (25), except that the binomial probability for n_{obs} in (25) is replaced by the Poisson probability in (71). Thus, the logistic-multinomial, multinomial log-linear, and Poisson log-linear models differ only in the probability model for n_{obs}. Cormack and Jupp (1991) and Huggins and Hwang (2011) have previously noted the similarity of the multinomial log-linear and Poisson log-linear model likelihoods in the case with no covariates, and Cormack and Jupp (1991) shows that the maximum likelihood estimators for the coverage model parameters are identical for the Poisson and multinomial models in that case.

Analogously to the multinomial log-linear model, the Poisson log-linear model and logistic-multinomial model can be further related by using (72) to rewrite the model for expected cell counts as

$$\log(\mu_{k,y}) = \gamma_k^P + l_1(y)\log\left(\frac{\tilde{\phi}_1(\mathbf{x}_{[k]},\boldsymbol{\beta}_1)}{1-\tilde{\phi}_1(\mathbf{x}_{[k]},\boldsymbol{\beta}_1)}\right) + l_2(y)\log\left(\frac{\tilde{\phi}_2(\mathbf{x}_{[k]},\boldsymbol{\beta}_2)}{1-\tilde{\phi}_2(\mathbf{x}_{[k]},\boldsymbol{\beta}_2)}\right) \tag{73}$$

where $\gamma_k^P = \log(N) + \log(\theta_k) + \log\left(1-\tilde{\phi}_1(\mathbf{x}_{[k]},\boldsymbol{\beta}_1)\right) + \log\left(1-\tilde{\phi}_2(\mathbf{x}_{[k]},\boldsymbol{\beta}_2)\right)$ is the log of the expected number of people with covariate combination $\mathbf{x}_{[k]}$ that are missed by both lists. Equating $\log\left(\frac{\tilde{\phi}_j(\mathbf{x}_{[k]},\boldsymbol{\beta}_j)}{1-\tilde{\phi}_j(\mathbf{x}_{[k]},\boldsymbol{\beta}_j)}\right)$ in (73) with $\xi_{l_j(y)}^{L_j} + \xi_{k,l_j(y)}^{\mathbf{X}L_j}$, for $j\in\{1,2\}$ in (69), it can be seen that the logistic coverage models of the logistic-multinomial model can be recovered from the Poisson log-linear model, however, because of the (slight) difference in likelihoods and different prior specifications, inferences are not guaranteed to be identical under the two models. Similarly to the multinomial log-linear model, the agreement between the logistic-multinomial and the Poisson log-linear models for expected cell counts is exact when a saturated model for the covariates is adopted for $\xi_k^{\mathbf{X}}$.

To obtain the posterior predictive distribution for the number missed by both lists, at each covariate level, and the total population size, under the Poisson log-linear model, there appear to be two possible approaches. The most obvious approach is to exploit the assumption of conditionally independent cell counts which implies

$$p(\mathbf{M}_{\cdot,(0,0)}|\mathbf{D}^{\text{obs}}) = \int p(\mathbf{M}_{\cdot,(0,0)}|\mathbf{D}^{\text{obs}},\boldsymbol{\mu})\,p(\boldsymbol{\mu}|\mathbf{D}^{\text{obs}})\,d\boldsymbol{\mu}$$

Algorithm 4 Posterior predictive sampling for the number missed by both lists under the Poisson log-linear model

1: **for** t in $\{1 \dots T\}$ **do**
2: draw $\boldsymbol{\xi}^{(t)}$ from $p(\boldsymbol{\xi}|\mathbf{D}^{\text{obs}})$
3: **for** k in $\{1 \dots K\}$ **do**
4: set $\mu^{(t)}_{k,(0,0)} = \exp(\xi^{(t)}_0 + \xi^{\mathbf{X},(t)}_k)$
5: draw $M^{(t)}_{k,(00)} \sim \text{Poisson}(\mu^{(t)}_{k,(0,0)})$.
6: **end for**
7: combine $\mathbf{M}^{(t)}_{.,(0,0)} = \{M^{(t)}_{k,(0,0)}, k \in \{1, \dots, K\}$ with $\mathbf{D}^{\text{obs}}$ to create $\mathbf{M}^{\text{com},(t)}$.
8: **end for**

$$= \int \prod_k \text{Poisson}(M_{k,(0,0)}|\mu_{k,(0,0)})\; p(\boldsymbol{\mu}|\mathbf{D}^{\text{obs}})\; d\boldsymbol{\mu}, \tag{74}$$

where

$$\mu_{k,(0,0)} = \xi_0 + \xi^{\mathbf{X}}_k \;\text{ for } k \in \{1, \dots, K\}. \tag{75}$$

Given a draw from the posterior for the log-linear model parameters $\boldsymbol{\xi}$, we can obtain a draw from the posterior for the cell means $\{\mu_{k,(0,0)},\, k \in \{1, \dots, K\}\}$, using (75) and a draw from the posterior for the counts of people missed by both lists, at each covariate combination, by drawing independent Poisson random variates with expected values equal to these simulated cell means. Repeating these steps for a sample from the posterior for $\boldsymbol{\xi}$ gives an approximation of the posterior predictive distribution (74); this approach is described in Algorithm 4.

The posterior for the population counts for aggregations of interest follow straightforwardly from the observed and the generated counts of the number missed by both lists. Thus, whereas for the logistic-multinomial or multinomial log-linear model the total population size is estimated directly as a parameter of the model, and the corresponding number of people missed by both lists are distributed across the covariate levels in accordance with the estimated covariate distribution and coverage models, the estimate for the total population estimates under the Poisson log-linear model is built up from covariate-specific estimates.

An alternative approach to obtain the posterior predictive distribution for the unobserved counts for the group missed by both lists, exploits the idea that the Poisson model has implications for the total population size. Even though N is not a parameter of the Poisson log-linear model, it is still an unknown that we are interested in, so it is reasonable to consider the joint posterior predictive distribution

$$p(\mathbf{M}_{.,(0,0)}, N|\mathbf{D}^{\text{obs}}) = \int p(\mathbf{M}_{.,(0,0)}|N, \boldsymbol{\mu}, \mathbf{D}^{\text{obs}})\; p(N|\boldsymbol{\mu}, \mathbf{D}^{\text{obs}})\; p(\boldsymbol{\mu}|\mathbf{D}^{\text{obs}})\; d\boldsymbol{\mu}. \tag{76}$$

The posterior for $\boldsymbol{\mu}$ is obtained, unconditionally on N, directly from the Poisson model, i.e. $p(\boldsymbol{\mu}|\mathbf{D}^{\text{obs}}) \propto p(\mathbf{D}^{\text{obs}}|\boldsymbol{\mu})\; p(\boldsymbol{\mu})$, and, of course, $\boldsymbol{\mu}$ is completely deter-

mined by the log-linear model parameters $\boldsymbol{\xi}$, so we could work just as easily with the posterior for $\boldsymbol{\xi}$.

The conditional posterior for N is given by

$$p(N|\mathbf{D}^{\text{obs}}, \boldsymbol{\mu}) \propto p(\mathbf{D}^{\text{obs}}|N, \boldsymbol{\mu})\; p(N|\boldsymbol{\mu}). \tag{77}$$

From the well-known relationship between the Poisson and multinomial models (Gelman et al. (2014, p.426)), conditionally on N, the distribution of the complete vector of observed and unobserved cell counts, $\mathbf{M}^{\text{com}}$, is multinomial with size parameter N and cell probabilities given by $\mu^{(*)}_{k,y} = \mu_{k,y}/\mu_{++}$. Noting that, in this notation, $p^P_{(0,0)}(\boldsymbol{\mu}) = \sum_k \mu^{(*)}_{k,(0,0)}$, it follows from the derivation of the likelihood for the multinomial log-linear model (64) in Section 8.1 that

$$p(\mathbf{D}^{\text{obs}}|N, \boldsymbol{\mu}) \propto \frac{N!}{(N - n_{\text{obs}})!} \prod_{k,y \neq (0,0)} (\mu^{(*)}_{k,y})^{M_{k,y}} \left(p^P_{(0,0)}(\boldsymbol{\mu})\right)^{(N - n_{\text{obs}})}, \tag{78}$$

and consequently,

$$\begin{aligned} p(N|\mathbf{D}^{\text{obs}}, \boldsymbol{\mu}) &\propto \text{Poisson}(N|\mu_{++})\; \frac{N!}{(N - n_{\text{obs}})!} \left(p^P_{(0,0)}(\boldsymbol{\mu})\right)^{(N - n_{\text{obs}})} \\ &\propto \frac{1}{(N - n_{\text{obs}})!} \left(\mu_{++}\; p^P_{(0,0)}(\boldsymbol{\mu})\right)^{N - n_{\text{obs}}} \mu_{++}^{n_{\text{obs}}} \\ &\propto \frac{1}{(N - n_{\text{obs}})!} \left(\mu_{++}\; p^P_{(0,0)}(\boldsymbol{\mu})\right)^{N - n_{\text{obs}}}. \end{aligned} \tag{79}$$

The right hand side of (79) is proportional to a Poisson probability mass function with expectation $\mu_{++}\; p^P_{(0,0)}(\boldsymbol{\mu})$, which is the expected number of people missed by both lists. Thus, to generate a draw from the conditional posterior for N, we draw $N_{(0,0)} \sim \text{Poisson}(\mu_{++}\; p^P_{(0,0)}(\boldsymbol{\mu}))$, and set $N = n_{\text{obs}} + N_{(0,0)}$.

The final step in obtaining the joint posterior for $(M_{\cdot,(0,0)}, N, \boldsymbol{\mu})$ using the decomposition in (76) is to compute $p(M_{\cdot,(0,0)}|N, \boldsymbol{\mu}, \mathbf{D}^{\text{obs}})$. From (74), it follows that $p(M_{\cdot,(0,0)}|\boldsymbol{\mu}, \mathbf{D}^{\text{obs}})$ is the product of independent Poisson probabilities. Conditioning on N in addition to $\mathbf{D}^{\text{obs}}$ implies conditioning on $N - n_{\text{obs}} = \sum_k M_{k,(0,0)}$, since n_{obs} is just the sum of the observed counts. Consequently, $p(M_{\cdot,(0,0)}|N, \boldsymbol{\mu}, \mathbf{D}^{\text{obs}})$ is a multinomial distribution with size parameter $(N - n_{\text{obs}})$ and probability vector $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_K)'$, where $\nu_k = \mu_{k,(0,0)}/p^P_{(0,0)}(\boldsymbol{\mu}), k \in \{1, \ldots, K\}$.

By introducing N as an additional unknown of interest, it becomes clear that the posterior predictive distribution for the completed population counts $\mathbf{M}^{\text{com}}$ can be simulated using a version of Algorithm 1, presented, in Appendix E of the Supplementary Material (Graham et al., 2023). This further highlights the similarity of inference under Poisson log-linear and logistic-multinomial models, though the impact of the implied restriction to a conditional Poisson prior for N, imposed by the Poisson log-linear model is apparent in the derivation of the conditional posterior for N in (79).

Given the simplicity of drawing independent Poisson variates for the group missed by both lists, by covariate combination, as described in Algorithm 4, the alternative involving explicitly generating the total population size described above is primarily of theoretical interest but is useful in relating the Poisson log-linear and logistic-multinomial modelling approaches to small domain dual systems estimation.

8.3. Comparing estimates under Poisson log-linear model and logistic-multinomial models

Initially we considered fitting a Poisson log-linear model that mimicked the logistic-multinomial model fitted to the area (or region) by sex by age data, comprising 11,929 covariate combinations, discussed in Section 6. To construct a log-linear model with an equivalent covariate structure to the logistic-multinomial model, we formulate a saturated model for the covariates. In log-linear model notation of (69) this implies

$$\xi_k^{\mathbf{X}} = \xi_{k[S]}^{\mathrm{S}} + \xi_{k[A]}^{\mathrm{A}} + \xi_{k[R]}^{\mathrm{R}} + \xi_{k[SA]}^{\mathrm{SA}} + \xi_{k[SR]}^{\mathrm{SR}} + \xi_{k[AR]}^{\mathrm{AR}} + \xi_{k[SAR]}^{\mathrm{SAR}} \tag{80}$$

where $\xi_{k[R]}^{\mathrm{R}}$ represents the area-specific term (nominally representing local government regions in the simulated data), and the superscripts SA, SR, AR, SAR denote sex by age, sex by area, age by area, and sex by age by area interaction terms, respectively. The analogous component of the logistic-multinomial model is the unstructured multinomial distribution for the covariate distribution.

Components of the Poisson log-linear model involving the list inclusion indicators were modelled using:

$$\xi_{k,l_j(y)}^{\mathbf{X}L_j} = \xi_{k[S],l_j(y)}^{SL_j} + \xi_{k[A],l_j(y)}^{AL_j} + \xi_{k[SA],l_j(y)}^{SAL_j} + \xi_{k[R],l_j(y)}^{RL_j}, \quad j \in \{1,2\}, \tag{81}$$

where the area effects $\xi_{r,1}^{RL_j}$ were modelled as draws from a Normal distribution

$$\xi_{r,1}^{RL_j} \overset{\text{indep}}{\sim} \text{Normal}(\xi_1^{L_j}, \sigma^2); \quad r \in \{1,\ldots,K_R\}, j \in \{1,2\}, \tag{82}$$

where K_R is the number of areas (67 in our example). In (81), parameters with $l_j(y) = 0$ are set to zero, and the superscripts SL_j, AL_j, SAL_j, and RL_j correspond to parameters representing sex, age, sex by age interaction, and area effect on List j coverage, for $j \in \{1,2\}$. Note also that the parameter in (69) that corresponds to the intercept of the List j coverage model, $\xi_1^{L_j}$, has been moved to centre the model for the area effects on coverage in (82). The parameters $\boldsymbol{\xi}_{a,l_j(y)}^{AL_j}$ in (81) represent spline models for age effects. Letting $\mathbf{Z_a}$ denote a vector with entries representing the basis function values for a spline representation of the age a, $\boldsymbol{\xi}_{a,l_j(y)}^{AL_j} = l_j(y) \times \mathbf{Z_a}'\boldsymbol{\psi}_j$, for some parameter vector $\boldsymbol{\psi}_j$, for $j \in \{1,2\}$. We used the same spline representation of age discussed in Section 6, i.e. a cubic spline with internal knots set at 10 years intervals. The spline representation of age is also used to construct the terms, $\xi_{k[SA],l_j(y)}^{SAL_j}$, $j \in \{1,2\}$, that correspond

to the interaction effect of sex and age on list inclusion. Thus, for the sex and age group denoted (s, a), with $s = 0$ and $s = 1$, denoting males and females, respectively, $\xi^{SAL_j}_{s,a,l_j(y)} = l_j(y) \times s \times \mathbf{Z_a}'\tilde{\boldsymbol{\psi}}_j$, $j \in \{1, 2\}$.

It proved impractical to fit the Poisson log-linear model defined by (80)–(81) and (82) to the area by sex by age data due to memory and computation time issues. Consequently, we considered a simplified version of the full Poisson log-linear model that omitted the age by area and sex by age by area terms from the model for the covariates, described in (80), and omitted the sex by age interaction effects in the model for coverage effects, in (81). However, computation times for fitting this simplified model in Stan remained impractically long. For example, 1000 iterations (of three parallel chains) took over 29 hours and remained far from convergence after discarding the first 500 iterations as burn-in (the largest $\hat{R}$ statistic was 4.77). We note that the model implementation in Stan took advantage of the Q-R decomposition, as recommended in the Stan user guide (Stan Development Team, 2021, `https://mc-stan.org/docs/2_29/stan-users-guide-2_29.pdf`), and this lead to a substantial improvement in computing time. A comparable logistic-multinomial model took 7.5 hours to complete 4,000 iterations with strong evidence of convergence after discarding the first 3000 iterations (largest $\hat{R} < 1.02$).

In view of the computational issues encountered in the Poisson log-linear modelling of the area by age by sex data, we based our comparison of inference under Poisson log-linear and logistic-multinomial models on a simpler data structure, similar to that used to explore dependent dual systems estimation in Section 7, but generated under an assumption of conditionally independent list inclusion. Thus, our comparison is based on a simulated target population of 1,000,000 with the same distribution over 180 sex by age combinations as the data used in Section 7, and generated using the same marginal list inclusion models but assuming conditional independence, of list inclusion.

We fitted a single-level Poisson log-linear model. Letting $M_{sa,y}$ denote the count for list inclusion cell y, for the covariate combination $(\text{sex} = s, \text{age} = a)$, the model structure can be described as:

$$\begin{aligned}
&[M_{sa,y}|\boldsymbol{\mu}] \overset{\text{indep}}{\sim} \text{Poisson}(\mu_{sa,y}) \\
&\log(\mu_{sa,y}) = \xi_0 + \xi^{\text{S}}_s + \xi^{\text{A}}_a + \xi^{SA}_{sa} + \xi^{L_1}_{l_1(y)} + \xi^{SL_1}_{s,l_1(y)} + \boldsymbol{\xi}^{AL_1}_{a,l_1(y)} + \xi^{SAL_1}_{sa,l_1(y)} + \\
&\qquad \xi^{L_2}_{l_2(y)} + \xi^{SL_2}_{s,l_2(y)} + \xi^{AL_2}_{a,l_2(y)} + \xi^{SAL_2}_{sa,l_2(y)}.
\end{aligned} \tag{83}$$

We set $\xi^{L_j}_0 = 0$, for $j \in \{1, 2\}$. The $\xi^{L_j}_1$ parameters correspond to the intercept parameters of the logistic coverage models, for $j \in \{1, 2\}$. Similarly, we set $\xi^{SL_j}_0 = 0$, for $j \in \{1, 2\}$, so that $\xi^{SL_j}_1$ represents the effect of female sex on the probability of inclusion on List j. The parameters $\boldsymbol{\xi}^{AL_1}_{a,l_1(y)}$ and $\boldsymbol{\xi}^{AL_2}_{a,l_2(y)}$ represent spline models for age effects and the spline representation of age is also used to construct the sex by age by list interaction terms $\xi^{SAL_j}_{sa,l_j(y)}$, $j \in \{1, 2\}$, as described above for the Poisson log-linear model for the full area by sex by age data. The other parameters associated with age, ξ^A_a and ξ^{SA}_{sa} are not modelled

as spline functions, but as distinct parameters for each integer age value, though we adopt the convention that $\xi_0^A = \xi_{s,0}^{SA} = 0$. Moreover, we set $\xi_0^S = \xi_{0,a}^{SA} = 0$, for identifiability reasons. Note that the model (83) saturates the sex by age distribution. Including the intercept, ξ_0, there are 180 $(1 + 1 + 89 + 89)$ parameters for the 180 sex by age combinations in the data. We adopted independent Normal$(0, 3)$ priors for all model parameters and fitted the model in Stan to obtain a posterior sample for the log-linear model parameters. The posterior for the counts of people missed by both lists, by covariate combination, was obtained as described in Algorithm 4.

The logistic-multinomial model for the sex by age data includes the total population size, N, as an explicit parameter and can be written

$$[\mathbf{M}^{\text{com}}|N, \boldsymbol{\theta}] \sim \text{Multinomial}(N, \boldsymbol{\theta})$$

$$[M_{sa,y}|M_{sa,+}, \boldsymbol{\beta}] \overset{\text{indep}}{\sim} \text{Multinomial}\left(M_{sa,+}, \boldsymbol{\phi}(s, a, \boldsymbol{\beta})\right), \tag{84}$$

$$s \in \{0, 1\},\ a \in \{0, \ldots, 89\},\ y \in \mathcal{Y}, \tag{85}$$

where $M_{sa,+} = \sum_y M_{sa,y}$, $\mathbf{M}^{\text{com}} = \{M_{sa,+}, s \in \{0, 1\},\ a \in \{0, \ldots, 89\}\}$, and $\boldsymbol{\phi}(s, a, \boldsymbol{\beta}) = \big(\phi_{(1,1)}(s, a, \boldsymbol{\beta}), \phi_{(1,0)}(s, a, \boldsymbol{\beta}), \phi_{(0,1)}(s, a, \boldsymbol{\beta}), \phi_{(0,0)}(s, a, \boldsymbol{\beta})\big)$, and the usual conditional independence assumptions are invoked so

$$\begin{aligned}\phi_y(s, a, \boldsymbol{\beta}) = (\tilde{\phi}_1(s, a, \boldsymbol{\beta}_1))^{l_1(y)}\ (\tilde{\phi}_2(s, a, \boldsymbol{\beta}_2))^{l_2(y)} \\ (1 - \tilde{\phi}_1(s, a, \boldsymbol{\beta}_1))^{1-l_1(y)}\ (1 - \tilde{\phi}_2(s, a, \boldsymbol{\beta}_2))^{1-l_2(y)}.\end{aligned}$$

The list inclusion probabilities are modelled using the logistic models

$$\text{logit}\left(\tilde{\phi}_j(s, a, \boldsymbol{\beta}_j)\right) = \beta_{j,0} + s\beta_j^S + \mathbf{Z}'_a\boldsymbol{\beta}_j^A, + s\mathbf{Z}'_a\boldsymbol{\beta}_j^{SA},\ j \in \{1, 2\}. \tag{86}$$

We adopted Normal$(0, 3)$ priors for all the logistic regression parameters, a Dirichlet prior for $\boldsymbol{\theta}$—with all Dirichlet parameters set to 0.01—and the Jeffreys' prior for the total population size $(p(N) \propto 1/N)$. We fitted the logistic-multinomial model in Stan using the full likelihood, to obtain a posterior sample for $(N, \boldsymbol{\theta}, \boldsymbol{\beta})$. We then used (36) to obtain the distribution of the covariate values of the group missed by both lists, as per Algorithm 1.

The Poisson log-linear and logistic-multinomial models produced almost identical coverage model parameter estimates and population estimates in this example. Coverage model parameter estimates are compared in Tables 1 and 2, in Appendix E of Supplementary Material (Graham et al., 2023), and population estimates by age and sex are compared in Figure 14. Population estimates are presented as relative differences from true values, and it can be seen that the pattern of difference is virtually identical for the Poisson log-linear and logistic-multinomial models, and that both methods recover the underlying population structure. The 2.5%, 50% and 97.5% quantiles of the posterior distribution for the total population size N were (999,479, 1,011,030, 1,001,178) and (999,483, 1,000,323, 1,001,199) for the Poisson log-linear model and the logistic-multinomial model, respectively.

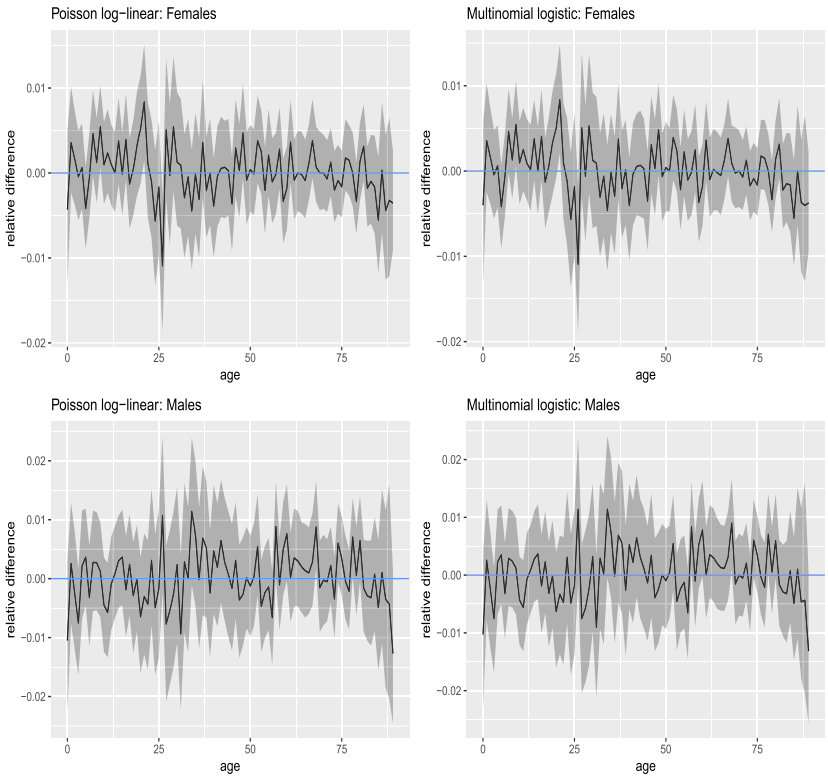

FIG 14. *Comparison of population estimates by sex and age, for Poisson log-linear (left column) and logistic-multinomial models (right column). Results are shown as relative difference from true values. Results for females are shown in the top row and results for males are shown in the bottom row. The shaded area represents an equal-tail-area* 95% *credible interval, and the solid black line represents the posterior median for the relative difference. Estimates are virtually identical for the Poisson and log-linear models and both models recover the underlying population structure.*

The biggest difference between the Poisson log-linear and logistic-multinomial models was the number of iterations required to achieve convergence and reasonable effective posterior sample sizes. For the logistic-multinomial model we ran five parallel chains of 2,500 iterations, discarded the first 1,000 iterations as burn-in and thinned the post-burn-in sample by five, to produce a nominal posterior sample size of 1,500. These settings produced strong evidence of convergence after 1,000 iterations with $\hat{R}$ statistics less than 1.01 for all parameters. Over all parameters, the minimum effective Monte Carlo sample size was 1,088 and the median was 1,500, equal to the nominal Monte Carlo sample size. The total computing time for the logistic-multinomial model was 173.7 seconds. With the same MCMC settings, the distribution of $\hat{R}$ convergence diagnostics for the Poisson log-linear model had a maximum of 1.042, third quartile of 1.025, median of 1.017 and lower quartile equal to 1.010. The distribution of effective

Monte Carlo sample sizes has a minimum of 41.0, lower quartile of 60.1, median of 157.9, third quartile of 220.8 and a maximum of 1,500. For many parameters of the Poisson log-linear model, the effective Monte Carlo sample size therefore seems too low for trustworthy posterior inference. The computing time of 91.7 seconds was, however, just over half that required for the logistic-multinomial model.

To obtain adequate effective posterior sample sizes for the Poisson log-linear models, we increased the number of iterations to 30,000 with a burn-in period of 3,000 and thinned by a factor of 90 to again yield a nominal posterior sample size of 1,500. These MCMC settings yielded a distribution of $\hat{R}$ statistics with maximum equal to 1.019, and distribution of effective sample sizes with minimum equal to 759.8, lower quartile equal to 947.5, median equal to 1411, and upper quartile equal to 1,500. Results reported above for the Poisson log-linear model are from this longer MCMC run for which the computing time was 524.2 seconds. Thus, although the Poisson log-linear model appears faster per iteration, it took approximately three times as long to produce effective Monte Carlo sample sizes comparable to (though slightly less than) the logistic-multinomial model.

Overall, evidence from this example suggests the logistic-multinomial model provides a more convenient parameterisation than the Poisson log-linear model for Stan to sample from. However, posterior inferences obtained under the two models were very similar, suggesting that the different prior specifications implied by the two models have little impact on the posterior, at least in the simple example considered. Additional details on fitting the Poisson log-linear model are give in Appendix E of Supplementary Material (Graham et al., 2023).

9. Discussion

Small domain population estimation is fundamentally about estimating the covariate distribution of the group not captured on at least one of the observed lists. The problem is, therefore, inherently a missing data problem to which a Bayesian approach is well-suited. This is perhaps most obvious in the application of Gibbs sampling to the population estimation. As discussed in Section 5.4, the Gibbs sampler alternates between imputing the covariates for the unobserved group, conditional on the most recent update of the dual systems model parameters, and updating model parameters by drawing from the conditional posterior that conditions on both the observed data and the most recently imputed set of covariate values for the unobserved group. While conceptually appealing, with advances in Bayesian computation, other approaches are possible, such as the sequential approach of Algorithm 1 whereby the posterior for the model parameters is first obtained using MCMC, followed by drawing in turn from the conditional posterior distributions for the total population size and the unobserved covariates.

The observation in Section 5.1, that under the Jeffreys' prior the marginal likelihood for the coverage model and covariate distribution parameters is the

conditional likelihood, provides an interesting connection between Bayesian and frequentist approaches, which often make use of a conditional likelihood, albeit conditioned on observed covariate values. The conditional likelihood is implemented in Stan reasonably easily and is a simple illustration of the idea of integrating out discrete parameters from the likelihood in order to make use of the HMC algorithm implemented in Stan. While the equivalence of the marginal likelihood for the coverage model and covariate distribution parameters under the Jeffreys' prior and the conditional likelihood may have been noted previously, at the time of writing we have been unable to locate a reference to this equivalence.

Much of the motivation for this paper came from the potential of administrative data to contribute to estimation of the size and distribution of human populations and the need for theory and methods for realising this potential. However, since administrative data may be prone to measurement error, if administrative data is to be a part of the future of small domain population estimation, further work is required to deal with the problems of measurement error on administrative lists. With the focus of much work on population estimation being on estimation of the total population size, the issue of covariate measurement error has not received much attention in the population estimation literature. However, van der Heijden et al. (2018) and van der Heijden et al. (2022) address the issue of measurement error, or differential reporting of covariates in different sources, in the context of frequentist log-linear modelling of multiple lists, using the E-M algorithm. There appears clear potential for incorporating these ideas into the Bayesian approach to population estimation.

Administrative data may also be prone to over-coverage, whereby an administrative list includes records for people not in the target population. For example, people who have emigrated may still have a presence in administrative data. Such over-coverage poses a serious challenge for dual systems estimation since lists with over-coverage are not subsets of the target population. Dual systems estimation cannot be expected to provide good estimates of the target population in the presence of list over-coverage, because dual systems estimation can only adjust for list under-coverage with respect to the population from which the lists are drawn. Thus, if list over-coverage is substantial, dual systems estimation is likely to result in over-estimation of the target population and, potentially, distorted covariate distributions, depending on the distribution of list over-coverage, by covariates. However, with good data on migration and accurate linkage between migration and the administrative list(s) included in dual systems estimation, the problem of over-coverage due to undetected out-migration can be minimised. When high quality migration data cannot be accessed, dual (and multiple) systems estimation methodology needs to be extended to estimate and adjust for over-coverage. Some work along these lines has been initiated, from both frequentist (Zhang, 2015, 2019) and Bayesian perspectives (Graham and Lin, 2020), however much remains to be done to build a generally applicable methodology to deal with over-coverage in population estimation, particularly when only two lists are available.

Another issue in using large administrative datasets for population estimation is the likely problem of linkage error. When links between individuals truly recorded on both lists are missed, individuals who should be recorded once in the $(1, 1)$ cell appear as two separate records, one in each of the $(1, 0)$ and $(0, 1)$ cells. This leads to over-estimation of the population. On the other hand, erroneous links between individuals recorded on only one of the lists, lead to underestimation of the population and may introduce covariate measurement error. However, as noted in Section 1 (Introduction), progress is being made on the problem of adjusting for linkage error in dual and multiple systems estimations (Ding and Fienberg, 1994; Di Consiglio and Tuoto, 2015, 2018; de Wolf, van der Laan and Zult, 2019; Sadinle, 2018), and on the related problem of uncertain identification of animals in ecological applications (Link et al., 2010; Schofield and Bonner, 2015; Zhang, Bravington and Fewster, 2019). Development of computationally tractable solutions to incorporating adjustment for linkage error into small domain population estimation is, however, likely to require continuing research. In general, further progress in Bayesian computation remains a priority for increasing the appeal of Bayesian approaches to population estimation and for unlocking the potential of Bayesian methods and large administrative datasets to improve the understanding of demographic and geographic distribution of human populations.

Supplementary Material

Supplement to "Bayesian Dual Systems Population Estimation for Small Domains"
(doi: 10.1214/23-SS146SUPP; .pdf). Likelihood derivations, extensions and additional results.

References

Alho, J. M. (1990). Logistic regression in capture-recapture models. *Biometrics* **46** 623–635. https://doi.org/10.2307/2532083. MR1085811

Baillargeon, S., Rivest, L.-P. et al. (2007). Rcapture: loglinear models for capture-recapture in R. *Journal of Statistical Software* **19** 1–31. https://doi.org/10.18637/jss.v019.i05. MR2432188

Bell, W. R. (1993). Using information from demographic analysis in dual-systems estimation. *Journal of the American Statistical Association* **88** 1106–1118. https://doi.org/10.1080/01621459.1993.10476381.

Brown, J., Abbott, O. and Diamond, I. (2006). Dependence in the 2001 one-number census project. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **169** 883–902. https://doi.org/10.1111/j.1467-985X.2006.00431.x. MR2291349

Carmona, C. and Nicholls, G. (2020). Semi-Modular Inference: Enhanced learning in multi-modular models by tempering the influence of components.

In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics* (S. Chiappa and R. Calandra, eds.). *Proceedings of Machine Learning Research* **108** 4226–4235. PMLR https://proceedings.mlr.press/v108/carmona20a.html. Accessed 7th November 2022.

Cormack, R. M. (1989). Log-linear models for capture-recapture. *Biometrics* **45** 395–413. https://doi.org/10.2307/2531485.

Cormack, R. and Jupp, P. (1991). Inference for Poisson and multinomial models for capture-recapture experiments. *Biometrika* **78** 911–916. https://doi.org/10.1093/biomet/78.4.911. MR1147028

Cruyff, M., van Dijk, J. and van der Heijden, P. G. (2017). The challenge of counting victims of human trafficking: Not on the record: A multiple systems estimation of the numbers of human trafficking victims in the Netherlands in 2010–2015 by year, age, gender, and type of exploitation. *Chance* **30** 41–49. https://doi.org/10.1080/09332480.2017.1383113.

de Wolf, P.-P., van der Laan, J. and Zult, D. (2019). Connecting correction methods for linkage error in capture-recapture. *Journal of Official Statistics* **35** 577–597. https://doi.org/10.2478/jos-2019-0024. MR3967717

Di Cecco, D. (2019). Estimating population size in multiple record systems with uncertainty of state identification. In *Analysis of Integrated Data* (L.-C. Zhang and R. L. Chambers, eds.) 169–196. Chapman and Hall.

Di Cecco, D., Di Zio, M. and Liseo, B. (2020a). Bayesian latent class models for capture–recapture in the presence of missing data. *Biometrical Journal* **62** 957–969. https://doi.org/10.1002/bimj.201900111. MR4122299

Di Cecco, D., Di Zio, M. and Liseo, B. (2020b). Population size estimation from incomplete multisource lists: A Bayesian perspective on latent class modelling. In *Proceedings of the 62nd ISI World Statistics Congress 2019: Special Topic Session: Volume 4* 65–69. International Statistics Institute. Department of Statistics, Malaysia, Kuala Lumpur. https://www.isi2019.org/isi-proceeding. Accessed 10th November 2022.

Di Consiglio, L. and Tuoto, T. (2015). Coverage evaluation on probabilistically linked data. *Journal of Official Statistics* **31** 415–429. http://dx.doi.org/10.1515/JOS-2015-0025.

Di Consiglio, L. and Tuoto, T. (2018). Population size estimation and linkage errors: The multiple lists case. *Journal of Official Statistics* **34** 889–908. http://dx.doi.org/10.2478/JOS-2018-0044.

Ding, Y. and Fienberg, S. E. (1994). Dual system estimation of census undercount in the prescence of matching error. *Survey Methodology* **20** 149–158. https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X199400214422. Accessed 10th November 2022.

Elliot, M. R. and Little, R. J. A. (2000). A Bayesian approach to combining information from a census, a coverage measurement survey and demographic analysis. *Journal of the American Statistical Association* **95** 351–362. https://doi.org/10.1080/01621459.2000.10474205.

Fellegi, I. and Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association* **64** 1183–2010. https://doi.org/10.1080/01621459.1969.10501049.

Fienberg, S. E. (1972). The multiple recapture census for closed populations and incomplete 2^k contingency tables. *Biometrika* **59** 591–603. https://doi.org/10.1093/biomet/59.3.591. MR0383619

Fienberg, S., Johnson, M. S. and Junker, B. W. (1999). Classical multilevel and Bayesian approaches to population size estimation using multiple lists. *Journal of the Royal Statistical Society, Series A* **162** 383–405. https://doi.org/10.1111/1467-985X.00143.

International Working Group for Disease Monitoring and Forecasting (1995). Capture—recapture and multiple-record systems estimation I: History and theoretical development. *American Journal of Epidemiology* **142** 1047–1058. https://doi.org/10.1093/oxfordjournals.aje.a117558.

Gelman, A., Jakulin, A., Pittau, M. G. and Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics* **2** 1360–1383. https://doi.org/10.1214/08-AOAS191. MR2655663

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B. and Vehtari, D. B. A. Rubin (2014). *Bayesian Data Analysis*, third edition. CRC Press, Boca Raton, FL. MR3235677

George, E. and Robert, C. (1992). Capture—recapture estimation via Gibbs sampling. *Biometrika* **79** 677–683. https://doi.org/10.1093/biomet/79.4.677. MR1209469

Graham, P. and Lin, A. (2020). Recent progress on implementing a Bayesian approach to population estimation from an administrative list subject to under and over-coverage. In *Proceedings of the 62nd ISI World Statistics Congress 2019, Special Topic Session* **4** 56–64. International Statistics Institute. Department of Statistics, Malaysia, Kuala Lumpur. https://www.isi2019.org/isi-proceeding. Accessed 10th November 2022.

Graham, P., Varn, L., Hendtlass, M., Green, R. and Richens, A. (2023). Supplement to "Bayesian dual systems population estimation for small domains". https://doi.org/10.1214/23-SS146SUPP

Huggins, R. M. (1989). On the statistical analysis of capture-recapture experiments. *Biometrika* **76** 133–140. https://doi.org/10.1093/biomet/76.1.133. MR0991431

Huggins, R. and Hwang, W.-H. (2011). A review of the use of conditional likelihood in capture-recapture experiments. *International Statistical Review* **79** 385–400. https://doi.org/10.1111/j.1751-5823.2011.00157.x.

King, R., McClintock, B. T., Kidney, D. and Borchers, D. (2016). Capture-recapture abundance estimation using a semi-complete data likelihood approach. *Annals of Applied Statistics* **10** 264–285. https://doi.org/10.1214/15-AOAS890. MR3480496

Leclerc, P., Vandal, A. C., Fall, A., Bruneau, J., Roy, É., Brissette, S., Archibald, C., Arruda, N. and Morissette, C. (2014). Estimating the size of the population of persons who inject drugs in the island of Montréal, Canada, using a six-source capture–recapture model. *Drug and Alcohol Dependence* **142** 174–180. https://doi.org/10.1016/j.drugalcdep.2014.06.022.

Lee, A. (1997). Some simple methods for generating correlated categorical variates. *Computational Statistics and Data Analysis* **26** 133–148. https://doi.org/10.1016/S0167-9473(97)00030-3.

Lindley, D. V. (1972). *Bayesian Statistiscs, a Review*. Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania. MR0329081

Link, W. A. and Barker, R. J. (2009). *Bayesian Inference: With Ecological Applications*. Academic Press, Amsterdam.

Link, W. A., Yoshizaki, J., Bailey, L. L. and Pollock, K. H. (2010). Uncovering a latent multinomial: analysis of mark–recapture data with misidentification. *Biometrics* **66** 178–185. https://doi.org/10.1111/j.1451-0420.2009.01244.x. MR2756704

Lum, K., Price, M. E. and Banks, D. (2013). Applications of multiple systems estimation in human rights research. *The American Statistician* **67** 191–200. https://doi.org/10.1080/00031305.2013.821093. MR3303809

Madigan, D. and York, J. C. (1997). Bayesian methods for estimation of the size of a closed population. *Biometrika* **84** 19–31. https://doi.org/10.1093/biomet/84.1.19. MR1450189

Manrique-Vallier, D. (2016). Bayesian population size estimation using Dirichlet process mixtures. *Biometrics* **72** 1246–1254. https://doi.org/10.1111/biom.12502. MR3591609

Manrique-Vallier, D., Ball, P. and Sulmont, D. (2019). Estimating the Number of Fatal Victims of the Peruvian Internal Armed Conflict, 1980–2000: an application of modern multi-list Capture-Recapture techniques. *arXiv preprint*. https://doi.org/10.48550/arXiv.1906.04763. Accessed 12th November 2022.

Manrique-Vallier, D., Price, M. E. and Gohdes, A. (2013). Multiple systems estimation techniques for estimating casualties in armed conflicts. In *Counting Civilian Casualties: An Introduction to Recording and Estimating Non-military Deaths in Conflict* (T. B. Seybolt, J. P. Aronson and B. Fischoff, eds.) 165–184. Oxford University Press https://hrdag.org/wp-content/uploads/2013/04/Manrique_Price_Gohdes_WorkingPaper.pdf. Accessed 12th November 2022.

Pledger, S. (2000). Unified maximum likelihood estimates for closed capture–recapture models using mixtures. *Biometrics* **56** 434–442. https://doi.org/10.1111/j.0006-341X.2000.00434.x.

Plummer, M. (2015). Cuts in Bayesian graphical models. *Statistics and Computing* **25** 37–43. https://doi.org/10.1007/s11222-014-9503-z. MR3304902

Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. *Annals of Statistics* **11** 416–431. https://doi.org/10.1214/aos/1176346150. MR0696056

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592. MR0455196

Rubin, D. B. (1981). The Bayesian bootstrap. *The Annals of Statistics* **9** 130–134. https://doi.org/10.1214/aos/1176345338. MR0600538

Sadinle, M. (2018). Bayesian propagation of record linkage uncertainty into

population size estimation of human rights violations. *The Annals of Applied Statistics* **12** 1013–1038. https://doi.org/10.1214/18-AOAS1178. MR3834293

Sanathanan, L. (1972). Estimating the size of a multinomial population. *Annals of Mathematical Statistics* **43** 142–152. https://doi.org/10.1214/aoms/1177692709. MR0298815

Sandland, R. and Cormack, R. (1984). Statistical inference for Poisson and multinomial models for capture-recapture experiments. *Biometrika* **71** 27–33. https://doi.org/10.1093/biomet/71.1.27. MR0738322

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. CRC press, Boca Raton, Florida. MR1692799

Schofield, M. R. and Bonner, S. J. (2015). Connecting the latent multinomial. *Biometrics* **71** 1070–1080. https://doi.org/10.1111/biom.12333. MR3436732

Seber, G. A. F. (1982). *The Estimation of Animal Abundance and Related Parameters*. Macmillan, New York. MR0686755

Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S. and Boatwright, P. (2005). A useful distribution for fitting discrete data: Revival of the Conway-Maxwell-Poisson distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **54** 127–142. https://doi.org/10.1111/j.1467-9876.2005.00474.x. MR2134602

Silverman, B. W. (2020). Multiple-systems analysis for the quantification of modern slavery: classical and Bayesian approaches. *Journal of the Royal Statistical Society, Series A* **183**. https://doi.org/10.1080/01621459.2019.1708748. MR4114463

Tancredi, A. and Liseo, B. (2012). A hierarchical Bayesian approach to record linkage and population size problems. *Annals of Applied Statistics* **5** 1553–1585. https://doi.org/10.1214/10-AOAS447. MR2849786

Tancredi, A., Steorts, R. and Liseo, B. (2020). A unified framework for de-duplication and population size estimation. *Bayesian Analysis* **15** 633–682. https://doi.org/10.1214/19-BA1146. MR4122517

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association* **82** 528–540. https://doi.org/10.1080/01621459.1987.10478458. MR0898357

Stan Development Team (2021). Stan Modeling Language Users Guide and Reference Manual, version 2.30 https://mc-stan.org. Accessed 12th November 2022.

Tilling, K. and Sterne, J. A. C. (1999). Capture-recapture models including covariate effects. *American Journal of Epidemiology* **149** 392–400. https://doi.org/10.1093/oxfordjournals.aje.a009825.

Tuoto, T., Di Cecco, D. and Tancredi, A. (2022). Bayesian analysis of one-inflated models for elusive population size estimation. *Biometrical Journal* **64** 912–933. https://doi.org/10.1002/bimj.202100187. MR4444857

van der Heijden, P. G. M., Whittaker, J., Cruyff, M., Bakker, B., van der Vliet, R. et al. (2012). People born in the Middle East but residing

in the Netherlands: Invariant population size estimates and the role of active and passive covariates. *The Annals of Applied Statistics* **6** 831–852. https://doi.org/10.1214/12-AOAS536. MR3012511

VAN DER HEIJDEN, P. G., SMITH, P. A., CRUYFF, M. and BAKKER, B. (2018). An overview of population size estimation where linking registers results in incomplete covariates, with an application to mode of transport of serious road casualties. *Journal of Official Statistics* **34** 239–263. http://dx.doi.org/10.1515/JOS-2018-0011.

VAN DER HEIJDEN, P. G., CRUYFF, M., SMITH, P. A., BYCROFT, C., GRAHAM, P. and MATHESON-DUNNING, N. (2022). Multiple system estimation using covariates having missing values and measurement error: Estimating the size of the Māori population in New Zealand. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **185** 156–177. https://doi.org/10.1111/rssa.12731. MR4384301

VAN DIJK, J., VAN DER HEIJDEN, P. G. and KRAGTEN-HEERDINK, S. L. (2016). Multiple systems estimation for estimating the number of victims of human trafficking across the world Technical Report, UNODC: United Nations Office on Drugs and Crime. https://eprints.soton.ac.uk/399731/1/UNODC_DNR_research_brief.pdf. Accessed 14th November 2022.

WOLTER, K. M. (1990). Capture-recapture estimation in the presence of a known sex ratio. *Biometrics* **46** 157–162. https://doi.org/10.2307/2531638. MR1059109

STATISTICS NEW ZEALAND (2019). *Dual System Estimation Combining Census Responses and an Admin Population*. Statistics New Zealand Tatauranga Aotearoa, Wellington, New Zealand. https://stats.govt.nz/methods/. Accessed 12th November 2022. MR1075417

ZHANG, L.-C. (2015). On modelling register coverage errors. *Journal of Official Statistics* **31** 381–396. https://doi.org/10.1515/jos-2015-0023.

ZHANG, L.-C. (2019). Log-linear models of erroneous list data. In *Analysis of Integrated Data* (L.-C. Zhang and R. L. Chambers, eds.) 197–218. Chapman and Hall, Boca Raton, Fl.

ZHANG, W., BRAVINGTON, M. and FEWSTER, R. (2019). Fast likelihood-based inference for latent count models using the saddlepoint approximation. *Biometrics* **75** 723–733. https://doi.org/10.1111/biom.13030. MR4012079

ZWANE, E. and VAN DER HEIJDEN, P. (2005). Population estimation using the multiple system estimator in the presence of continuous covariates. *Statistical Modelling* **5** 39–52. https://doi.org/10.1191/1471082X05st086oa. MR2133527