Multiple Systems Estimation for Modern Slavery: Robustness of List Omission and Combination

Serveh Sharifi Far[*], Ruth King[*], Sheila Bird[*] [†], Antony Overstall[‡],

Hannah Worthington[§] and Nicholas P. Jewell[¶]

Abstract

Performing censuses on stigmatised or vulnerable populations is challenging, however, for such populations partial enumeration is often possible using different lists or sources. If the sources overlap then multiple systems estimation (MSE) methods can be applied to obtain an estimate of the total population. These are typically expressed by a log-linear model which permits positive/negative dependencies between lists. This paper considers issues that arise for the application of MSE to modern slavery where there is little to no overlap of individuals across lists. We investigate the robustness of MSE in terms of the importance of each list and the impact of combining lists on the estimation process. We undertake a simulation study and consider real national modern slavery data from the UK and Romania.

Key words: Combining sources; Estimate stability; Generalised linear models; List omission.

[*] University of Edinburgh, Edinburgh, UK

[†] MRC Biostatistics Unit, University of Cambridge, Cambridge, UK

[‡] University of Southampton, Southampton, UK

[§] University of St Andrews, St Andrews, UK

[¶] The London School of Hygiene and Tropical Medicine, London, UK

Corresponding author

Serveh Sharifi Far, University Teacher in Statistics, School of Mathematics, University of Edinburgh. JCMB, Peter Guthrie Tait Rd, Edinburgh, EH9 3FD, UK.
Email: serveh.sharifi@ed.ac.uk

Author Biographies

Serveh Sharifi Far is a University Teacher in Statistics in the School of Mathematics at the University of Edinburgh. Her research interests include parameter redundancy and analysis of categorical data motivated by social science and biological data.
Email: serveh.sharifi@ed.ac.uk; Tel: 44-131-6505051.

Ruth King is the Thomas Bayes' Chair of Statistics in the School of Mathematics at the University of Edinburgh. Her research interests include statistical modelling, capture-recapture data, multiple systems estimation and missing data applied to problems in ecology, epidemiology and healthcare.
Email: ruth.king@ed.ac.uk; Tel: 44-131-6505947.

Sheila M. Bird is an honorary professor at College of Medicine and Veterinary Medicine, University of Edinburgh and visiting scientist at MRC Biostatistics Unit, University of Cambridge, CB2 0SR. She is a biostatistician, formerly Programme Leader at MRC Biostatistics Unit, Cambridge.
Email: sheila.bird@mrc-bsu.cam.ac.uk; Tel: 44-7800-639269.

Antony Overstall is an Associate Professor of Statistics in the Southampton Statistical Sciences Research Institute at the University of Southampton. His research interests are statistical modelling and computation.
Email: a.m.overstall@soton.ac.uk; Tel: 44-238-0592724.

Hannah Worthington is a lecturer in Statistics in the School of Mathematics and Statistics at the University of St Andrews. Her research interests include hidden Markov models

applied to problems in ecology, capture-recapture data, incorporating individual heterogeneity and multi-state modelling.
Email: hw233@st-andrews.ac.uk; Tel: 44-1334-461806

Nicholas P. Jewell is Chair of Biostatistics and Epidemiology at the London School of Hygiene and Tropical Medicine, after a long career as Professor of Biostatistics and Statistics at the University of California, Berkeley. His research interests include statistical issues associated with infectious diseases and epidemiology and counting challenges in human rights arenas.
Email: nicholas.jewell@lshtm.ac.uk; Tel: 44-20-76368636

1 Introduction

Modern forms of slavery persist in the 21st century despite the legislative successes of 19th century reformers in having predominantly abolished traditional slavery. Documenting and quantifying the prevalence of modern slavery is a challenging task for many reasons but not least due to the hidden nature of individuals who would be classed in this category and how victims of modern slavery are defined. Further, the nature of modern slavery means that international boundaries may be crossed with many modern slavery victims also victims of illegal trafficking (see for example Cruyff *et al* (2017); van Dijk *et al* (2017) for the context to human trafficking). However, the problem is significantly wider that the exploitation of illegal immigrants - for example, 16% of the UK's identified potential victims of modern slavery are its own citizens. The own-citizen percentage was higher still at 32% for the 2121 potential victims in 2017 who were children (Home Office, 2018). Major other countries-of-origin for UK-identified victims include Albania and Vietnam but these two, together with the UK itself, may have a different representation within the totality of victims (non-identified as well as identified) of modern slavery in the UK. Hence, policy initiatives for the prevention of human trafficking that have been directed at Albania and Vietnam might need re-orientation when UK's unidentified victims are estimated by where they originated from.

In the UK, all police forces report identified victims of modern slavery to the National Crime Agency (NCA). Support, ranging in duration from 7 to 13 weeks, is available for "probable-cause" victims unless or until their final-status is determined otherwise. Overlaps between the list held by UK's NCA and those of other service providers arise both because of the support on offer to probable-cause victims, or because these services may have referred identified potential victims to NCA for appraisal of their eligibility for support, or because police action could rescue further victims. It is this overlap of individuals observed by the different sources that permits the use of multiple systems estimation (MSE) for estimating the difficult to obtain total prevalence and associated measure of the problem within society. See Bird and King (2018) for a review of multiple systems estimation and their application to different populations; Jewell *et al* (2013) for an application to estimating nonmilitary deaths in conflict; and Silverman (2019), and references therein, for discussion of their application to modern slavery.

Complexities can occur for modern slavery data as the term covers a range of different types of modern slavery, including for example, domestic/physical labour and forced prostitution. The characteristics of the type of victimisation typically varies by gender (for example, physical versus domestic labour) and age-group (child versus adult female prostitution); and is also likely to determine how many other victims belong to the same cluster as the listed victim, for example, many adult males engaged in physical labour, may be co-located and controlled by a gang-master; solo female domestic slave; or a clutch of sex workers who travel between premises in different towns and may include children in their number. Professionals in different capacities may report suspect activity to authorities. For example, doctors who are made aware that a child is at risk of prostitution, or that victims of human trafficking are held at a specific location, may (or be required to) inform the relevant authorities so that a rescue can be attempted by the police. Considerations pertain to non-governmental voluntary organizations including those which might, in less extreme circumstances, be unwilling to cross-refer leading to minimal overlap between different lists, for example, in relation to voluntary organizations giving refuge to escapee women versus males, or to adults versus children.

We focus on the common issue of limited or minimal overlap (where relatively few individuals are observed across the different lists used) within modern slavery application of MSE. Multiple lists with limited or minimal overlap can occur for numerous reasons, and affect different subsets of the population. For example, as discussed above, this may be the case for lists that are held by different non-governmental voluntary organizations. This, in turn, can lead to a number of different issues when applying a MSE approach, including models being unidentifiable with inestimable parameters (Sharifi Far *et al*, 2019) and potentially unstable estimation of the total population size. Further, demographic information or contextual data, such as type of victimization that victims of modern slavery are subjected to and whether drug dependent, may be important determinants of capture-propensity on some but not all lists, or the interaction between different lists. If such information is available, MSE can be extended to directly incorporate such factors (see for example, King *et al* (2005), in the case of MSE applied to injecting drug users). However, this leads to a further reduction in the overlaps observed between the different lists, potentially exacerbating the issues further, and introducing a greater number of parameters to estimate. Thus, within this paper, we do not consider such characteristics further, and focus on the standard cross-classification of individuals across the different lists.

Our aim in this paper is to investigate, by simulation and empirically, the impact of lists with minimal overlaps for capture-recapture estimation of victims of modern slavery, and methods to combat effects of such phenomena on population size estimation.

2 Methods

We consider standard log-linear models for MSE, where we are able to explicitly account for dependencies between lists via associated log-linear interaction terms (Fienberg, 1972). We investigate the effect on population size estimation where there is limited overlap between the lists relating to the two specific methods of (i) list omission; and (ii) list combination. In particular, we shall consider an approach where we assess the influence of the lists on the estimation process by removing each list in turn from the analysis; and the impact of combining two lists where there is limited overlap between the lists. We begin by defining the models and associated MSE approach.

2.1 Multiple systems estimation (MSE)

We begin by describing the general framework for MSE. Let K denote the total number of lists available in the dataset, we label the individual lists $k = 1, \dots, K$ (with a minimum of $K = 2$ lists). We construct an incomplete 2^K contingency table where each element of the table corresponds to the number of individuals observed by the given list combination. The table is incomplete since we do not observe the number of individuals not observed by any of the K lists, and hence taking the total population size to be equal to the total number of observed individuals will lead to an underestimate of the total population size.

Mathematically, each cell is indexed in the form $\boldsymbol{k} \in \{0,1\}^K$, where the 1/0 correspond to the given list observing/not observing an individual, respectively. For example, when $K = 4$ the cell $\boldsymbol{k} = \{0,1,1,0\}$ corresponds to being observed by lists 2 and 3 but not lists 1 and 4. The cell $\boldsymbol{k} = \{0\}^K$ corresponds to not being observed by any of the lists.

Let $n_{\boldsymbol{k}}$ denote the number of individuals in cell $\boldsymbol{k} \in \{0,1\}^K$ of the contingency table; and $\mu_{\boldsymbol{k}}$ correspond to the mean cell count for cell $\boldsymbol{k}$. We specify the model as a generalised linear model, with Poisson error and log-link function, such that,

$$n_{\boldsymbol{k}}|\mu_{\boldsymbol{k}} \overset{ind}{\sim} Poisson(\mu_{\boldsymbol{k}}), \qquad \text{for } \boldsymbol{k} \in \{0,1\}^K. \tag{1}$$

Letting $\boldsymbol{\mu}$ denote the column vector of the mean cell counts, $\mu_{\boldsymbol{k}}$, we can write,

$$\log \boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\theta},$$

where $\boldsymbol{\theta}$ denotes the column vector of log-linear parameters and $\boldsymbol{X}$ is the associated design matrix describing the relationship between the (log) of the expected cell counts and the parameters. In general, $\boldsymbol{\theta}$ contains an intercept term (associated with the mean cell count), main effect terms for each list (associated with the propensity of being observed by a given list) and interaction terms (associated with dependencies between the different lists). Due to the incompleteness of the contingency table, we cannot estimate the K-way interaction for hierarchical log-linear models.

This modelling structure permits the estimation of the total population size as follows: the log-linear parameters, $\boldsymbol{\theta}$, can be estimated from the observed cell counts; given these estimates we are able to obtain the associated maximum likelihood estimate (MLE) and associated uncertainty of the unobserved cell, via the model specified in Equation (1). The associated uncertainty is described via a 95% confidence interval (CI), using the standard asymptotic normality assumption and estimated standard error calculated via the Hessian matrix evaluated at the MLE of the parameters. However, we note that the estimate of the total population size (and 95% CI) is in general, dependent on the model specified in terms of the interactions present within the model. This typically leads to a two-step process, (i) identify the "best" model in terms of the interactions present in the model; then (ii) obtain an estimate of the total population size given the specified model.

To discriminate between competing models and conduct the model selection step, it is conventional to use Akaike's information criterion, AIC (Akaike, 1974), where,

$$AIC = -2l(\widehat{\boldsymbol{\theta}}; \boldsymbol{n}) + 2p,$$

such that $l(\widehat{\boldsymbol{\theta}}; \boldsymbol{n})$ denotes the log-likelihood of the model evaluated at the MLEs of the parameters denoted $\widehat{\boldsymbol{\theta}}$, and p denotes the number of parameters in the model i.e. $p = |\boldsymbol{\theta}|$. The likelihood in this case simply corresponds to a product over independent Poisson terms. The AIC criterion is easily interpreted as a trade-off between the fit of the model to the data and the complexity of the model. The model with the smallest AIC statistic is deemed to be the "best" of the models considered, in this respect AIC assesses the relative performance of the competing models. See, for example, Coumans *et al* (2017); Silverman (2014); Van der Heijden *et al* (2012) for the use of the AIC statistic within the MSE context for modern slavery and other related populations; and Davison (2003) for discussion of alternative model selection tools.

In practice, it may not be feasible to fit every possible model (including/excluding interaction terms) to the data. If the dataset features many sources, the number of possible models becomes prohibitive and so a model search algorithm is typically implemented. For example, adding/removing interaction terms in a systematic manner until no improvement in the model is detected. In this paper, we use a model selection procedure using the AIC statistic and estimate the total population size from the single "best" model in order to investigate the issues of combining and omitting lists without the additional confounding with model-averaging issues. In particular, we are interested in the influence of each individual list on the total population estimate.

2.2 List influence

The pattern of the observed data, in terms of the number of individuals observed in the cross-classification across different lists is the underpinning principle permitting the estimation of the total population size via MSE. In general, situations can arise whereby, for example, there is a dominant list where a substantial proportion of individuals are observed by this single source (see, for example, Cormack (2000)); there is substantial dependence between lists (either positive or negative: see, for example, Jones *et al* (2014)); or limited overlap across lists leading to sparse contingency tables, i.e. tables with a large number of zero counts (see, for example, Chan *et al* (2019); Sharifi Far (2017)). We focus on this last case of minimal overlap between the different lists. Issues encountered in this scenario include model fitting complexity, including for example, model identifiability and parameter redundancy (Chan *et al*, 2019; Fienberg and Rinaldo, 2012; Sharifi Far *et al*, 2019; Silverman, 2019; Vincent *et al*, 2019).

To investigate the influence of the different lists on the statistical analysis, and focussing in particular on the estimation of the total population size, we consider both a (i) “leave-one-out” approach and (ii) combining lists approach.

2.2.1 Leave-one-out approach

The leave-one-out approach involves cycling through each possible list, removing the given list, constructing the reduced incomplete contingency table from the remaining sources before conducting the statistical analysis to obtain the total population size estimate as described above. In particular, we obtain the MLE of the total population size for the model deemed optimal via the AIC statistic and an associated 95% CI. When there are K lists in general, this means conducting K leave-one-out contingency table analyses. We note that for each leave-one out analysis, the total number of observed individuals is reduced (assuming that all lists observe at least one unique individual not observed by any other source). The estimates of population size from each of the K leave-one-out analyses can be compared with each other and also with the estimate of the total population size using all K sources. In the simulation study in Section 4, we can also compare the estimates with the (known) true population size.

2.2.2 Combining lists approach

In some cases, we may wish to combine two lists into a single list prior to analysing the contingency table. For example, we focus on the particular case where we may wish to do this due to the limited (or even lack of any) overlap between two (or more) of the sources used within the analysis. The new list then essentially corresponds to individuals observed by source A, say, or source B (or both). In the case of there being no individuals observed by both these two sources, the interaction between these sources is also not estimable (shown for a saturated model by Sharifi Far (2017)). Combining the two sources automatically removes the issue of identifiability of the interaction between these two sources as this parameter is no longer present. Further, unlike the leave-one-out approach, this approach does not reduce the number of individuals observed within the new revised contingency table; however, the number of lists is reduced by one. Once again the estimates of total population size can be compared using the original all-list data and then the reduced (combined list) contingency table. For the simulation study the estimate can also be compared to the (known) true population size from which the data are simulated.

3 Case Studies

We consider two case studies relating to data from the UK and Romania, both with 5 sources. Both of these cases have minimal overlap between some of the sources. For the Romanian data, one of the lists is dominant and contains the majority of the observations.

3.1 UK data

We consider the data presented by Silverman (2014) relating to modern slavery in the UK. The data contains 5 different sources corresponding to: Local Authority (LA); Non-Government organisations (NG); Police Force and/or National Crime Agency (PF); Government Organisations (GO); and the General Public (GP). For further information, including discussion of combining the police force and National Crime Agency as a single list, see Silverman (2014). The data are presented in Table 1. We note that there is no overlap between the lists LA and GP, i.e. no individuals are recorded by both of these sources, and, in general, there is very little overlap between GP and the other remaining lists. Given these data, it can be shown that the interaction between LA and GP (and all higher order interactions) cannot be estimated (Sharifi Far, 2017). In our analyses, due to the sparsity of the contingency table, we restrict the interactions to only two-way interactions between lists. When modelling the 5 lists, all the two-way interactions, except the LA and GP interaction, are estimable.

LA					LA	LA	LA							LA	LA		LA
	NG				NG			NG	NG	NG				NG	NG	NG	NG
		PF				PF		PF			PF	PF		PF		PF	PF
			GO				GO		GO		GO		GO		GO	GO	GO
				GP						GP		GP	GP				
54	463	995	695	316	15	19	3	62	19	1	76	11	8	1	1	4	1

Table 1: UK modern slavery data of non-zero contingency table cell entries. The five lists are: LA = local authority; PF = police force and/or National Crime Agency; GO = government organisation; NG = non-government organisation; GP = general public.

3.1.1 Full data analysis

We initially analyse the full 5-list dataset, in order to consider the robustness of the total population estimate when investigating the two issues of (i) removing each list in turn; and (ii) combining GP with each of the remaining lists in turn. We consider a model search algorithm using the AIC statistic to compare competing models. We restrict the set of models to those with two-way interactions, omitting the LA×GP interaction. The model identified as optimal has the following six two-way interactions and associated direction of the interaction (+ve = positive interaction and -ve = negative interaction): LA×NG (+ve); LA×PF (+ve); NG×GO (-ve); NG×GP (-ve); PF×GP (-ve); GO×GP (-ve). All interactions identified relating to either GO or GP correspond to negative interactions (so being identified by either of these sources leads to a lower chance of being observed by the other data source where there is an interaction). Conversely, interactions that involve only the lists LA, NG or PF have positive relationships. Given the above model, the corresponding MLE for the total population is 11313 with the 95% CI (9750, 12876).

3.1.2 Omitting lists: “leave-one-out”

We consider the influence of each list on the estimate of the population size, by omitting each list in turn. The estimates and 95% CIs are presented in Table 2, along with the sign of the included interaction terms. The population size estimates are highly variable for the different omitted lists. Identifying structured patterns within the output is non-trivial: omitting lists masks patterns in the cell entries, for example, a previous overlap between two lists becomes an observation in a single list when one of the sources is left-out, and different models (and interactions) will be identified given these changes. In all cases where interactions are chosen in both the full 5-list and reduced 4-list dataset analyses, the direction of the interactions remains consistent (except for NG×GO interaction which is positive when PF is removed). In every instance, the reduced dataset includes interactions in common with those identified for the full 5-list dataset. When omitting lists LA, NG and PF (that exhibit positive interactions between them in the full analysis) the reduced datasets lead to different set of interactions to the full dataset (but with some common interactions). The comparison when removing lists GO and GP is more straightforward: the model identified is simply the reduced model from the 5-list analysis, omitting the interaction terms associated with the omitted list. For these latter two cases, the estimate of the population size is similar to the estimate from the 5-list analysis. List PF has the largest number of observations - removing this list provides an estimate where the 95% CI does not include the population estimate obtained in the 5-list analysis.

3.1.3 Combining lists

The GP list has very little overlap with the other lists and no overlap with LA. Therefore, we combine this list with each of the other lists in turn and estimate the total population size from the reduced contingency table. The corresponding MLEs of the population size, 95% CIs and selected interaction terms are given in Table 3. To clearly denote which lists have been combined we “dot product” the list names, for example, the combination of GP and LA is denoted by “GP.LA”. The largest deviation from the population size estimate of the 5-list dataset is observed when LA and GP are combined. These two lists

Omitted list	Population estimate	95% confidence interval	Model
-	11313	(9750, 12876)	LA×NG (+ve); LA×PF (+ve); NG×GO (-ve); NG×GP (-ve); PF×GP (-ve); GO×GP (-ve)
LA	18945	(11740, 26150)	NG×PF (+ve); NG×GP (-ve); PF×GO (+ve); PF×GP (-ve); GO×GP (-ve)
NG	31118	(18893, 43343)	LA×PF (+ve); PF×GO (+ve)
PF	32042	(13781, 50304)	LA×NG (+ve); NG×GO (+ve); NG×GP (-ve);
GO	10202	(8061, 12343)	LA×NG (+ve); LA×PF (+ve); NG×GP (-ve); PF×GP (-ve)
GP	11015	(9447, 12583)	LA×NG (+ve); LA×PF (+ve); NG×GO (-ve)

Table 2: MLEs and associated 95% CIs for the total population size for the UK data, and corresponding model selected in terms of interaction terms present with associated estimated sign of the interaction. The first row (denoted by a "-") gives the results of the complete 5-list analysis; the remaining rows are the results of omitting each list in turn.

have no overlap and their interactions with the other lists are in opposite directions. This appears to have resulted in some interactions cancelling each other out. For instance, in the 5-list analysis GP×NG has a negative interaction whilst LA×NG has a positive interaction, once combined GP.LA has no interaction with NG. This has further impact on the remaining interactions between the non-combined lists with clear changes in the selected interaction terms. For the combinations of GP with NG and GO the interactions for the combined model appear more predictable: where the uncombined lists displayed interactions, the combined lists share those same interactions. The combination GP.PF lies somewhere in between the above cases: the majority of interactions can be anticipated from the original interactions, but there are also some changes in the interactions of the uncombined lists. Overall, when compared to the leave-one-out method there appears to be less variability in the range of estimates.

3.2 Romania data

We consider data collected for Romania in 2015. Five lists are included corresponding to: Police/agency against trafficking in persons and border police (PF), International organization for Migration (IM), Non-Governmental organisations (NG), Foreign Authorities (FA) and Other (OT). 879 individuals are observed, with the majority of these obtained by list PF (a total of 806 individuals are observed by PF; of these 758 are only identified by PF). Thus, PF dominates the other lists. IM observes a total of 48 individuals (1 individual is unique to IM); NG observes 25 individuals (19 of these are observed by at least one other list); FA observes 72 individuals (all these individuals are observed by at least one other list); and OT has 66 individuals (with 34 only observed by OT).

3.2.1 Full data analysis

We conduct an analysis of the full 5-list dataset. We restrict the model search to those including two-way interactions, and use the AIC statistic to determine the interactions present. The model selected as "best" had interactions: PF×IM (-ve); PF×NG (-ve); PF×FA (+ve); PF×OT (-ve); IM×FA (+ve); NG×FA (+ve); NG×OT (-ve); FA×OT

Combined lists	Population estimate	95% confidence interval	Model
-	11313	(9750, 12876)	LA×NG (+ve); LA×PF (+ve); NG×GO (-ve); NG×GP (-ve); PF×GP (-ve); GO×GP (-ve)
GP.LA	16071	(12661, 19481)	GP.LA×GO (-ve); NG×PF (+ve) PF×GO (+ve)
GP.NG	12661	(10920, 14403)	LA×GP.NG (+ve); LA×PF (+ve); GP.NG×GO (-ve)
GP.PF	13180	(11343, 15017)	LA×NG (+ve); LA×GP.PF (+ve); NG×GO (-ve)
GP.GO	14394	(11862, 16926)	LA×NG (+ve); LA×PF (+ve); NG×PF (+ve); NG×GP.GO (-ve)

Table 3: MLEs and 95% CIs of the population size for the UK data given the model selected, and corresponding model selected in terms of interaction terms present (estimated sign). The first row (denoted by a "-") gives the results of the complete 5-list analysis; the remaining rows are the results of combining list GP with each of the other lists. For the model description we denote the combined lists by the combined "dotted" abbreviations.

(+ve). The associated estimate of the population size is 921, with 95% CI (879*, 993). We truncate the lower limit of the 95% CI to the observed number of individuals (indicated by *). We will use this estimate as a baseline to investigate the impact of removing each of the lists in turn and secondly combining PF with each of the other lists in turn (chosen since PF has the smallest percentage overlap with each of the other lists).

3.2.2 Omitting lists: "leave-one-out"

The population size estimates, 95% CIs and selected model when each list is omitted in turn, are given in Table 4. Removing the dominant list PF (of which 86% of the individuals on this list are only seen on this list) leads to a substantial decrease in the estimate of the total population. This is unsurprising given the dominance of this list in observing individuals. In particular, this source alone records 74% of all individuals observed; and 68% of all individuals observed are only observed by this list. Omitting the other lists leads to estimates similar to the estimate obtained when using all 5 lists. We note that removing the OT list leads to a larger and highly imprecise estimate of the population size. In line with the observations from the UK data, there is generally agreement across the different omissions in the interactions identified: where an interaction is identified to be present, the sign of the interaction remains consistent whenever the interaction is detected. On removing lists that have a negative interaction with the dominant list PF (i.e. IM, NG and OT) the interaction terms identified are typically those identified by the 5-list analysis with those featuring the omitted list removed. For list FA which originally displayed a positive interaction with the dominant list PF, and contains no unique individuals, the selection of interactions is somewhat different amongst the remaining lists.

3.2.3 Combining lists

For the Romanian data, PF has minimal overlap with the other lists: only 48 individuals observed by list PF are observed by another list, which corresponds to only 6% of individuals observed by PF. We investigate the effect of combining PF with each of the other lists. Whilst this approach is similar to that of the UK data (combining with a minimally

Omitted lists	Population estimate	95% confidence interval	Model
-	921	(879*, 993)	PF×IM (-ve); PF×NG (-ve); PF×FA (+ve); PF×OT (-ve); IM×FA (+ve); NG×FA (+ve); NG×OT (-ve); FA×OT (+ve)
PF	258	(142, 374)	IM×FA (+ve); IM×OT (+ve); NG×FA (+ve)
IM	971	(742, 1200)	PF×NG (-ve); PF×FA (+ve); PF×OT (-ve); NG×FA (+ve); NG×OT (-ve); FA×OT (+ve)
NG	923	(842, 1005)	PF×IM (-ve); PF×FA (+ve); PF×OT (-ve); IM×FA (+ve); FA×OT (+ve)
FA	1035	(895, 1175)	PF×NG (-ve); PF×OT (-ve); IM×NG (+ve); IM×OT (+ve)
OT	2915	(845*, 5638)	PF×IM (-ve); PF×FA (+ve); IM×FA (+ve); NG×FA (+ve)

Table 4: Results for the Romanian data in terms of the MLEs and associated 95% CIs for the total population size given the model selected, and corresponding model selected in terms of interaction terms present with associated estimated sign of the interaction. The first row (denoted by a "-") gives the results of the complete 5-list analysis; the remaining rows are the results of omitting each list in turn. When the lower bound of the confidence interval was truncated to the number of observed individuals, it is indicated by *.

overlapping list), here there is a structural difference in that the list also accounts for the majority of observations. The corresponding results are given in Table 5. The estimates obtained in each of the combined list analyses are reasonably consistent with substantially overlapping 95% CIs (compared with the estimate using all 5 lists). The largest discrepancy arises when combining list PF with list FA. This is potentially due to the complex relationship between these two lists: of the 8 interactions identified in the 5-list analysis, 7 feature PF, FA or both. Once again the interactions identified (and associated sign) are remain fairly consistent across analyses. As for the UK data, the combining of lists leads to less variable estimates of population size compared to omitting lists.

The case studies suggest that analyses should be conducted with some caution in the presence of minimally overlapping sources. In particular, omitting sources with limited overlap can lead to different behaviours in the estimate of the population size. Alternatively, combining a list with limited overlap to another list appears to provide less variable estimates. Thus, how we deal with such sources can have a significant impact on the population size estimate - and some sensitivity of the analyses should be conducted. To investigate the impact further where the observed contingency tables are more “controlled”, we conduct a simulation study, motivated by the larger UK data.

4 Simulation Study

The simulation study is motivated by the UK dataset with 5 sources, which represents a common structure between the victims-of-slavery sources - in particular when there are two sources for which no individuals are observed in common (sources LA and GP). We use the fitted model to the full 5-list data analysis (so that there are 6 interaction terms with non-zero effects) as the generating model within the simulation study and use the same list names for simplicity. We set the true population size to be equal to 11313. We generate 500 datasets from the given (conditional Multinomial) model. Only the cell count corresponding to cell $\boldsymbol{k} = \{0, 0, 0, 0, 0\}$ is unknown. For each simulated dataset,

Combined lists	Population estimate	95% confidence interval	Model
-	921	(879*, 993)	PF×IM (-ve); PF×NG (-ve); PF×FA (+ve); PF×OT (-ve); IM×FA (+ve); NG×FA (+ve); NG×OT (-ve); FA×OT (+ve)
PF. IM	1087	(879*, 1400)	PF. IM×NG (-ve); PF. IM×FA (+ve); PF. IM×OT (-ve); NG×FA (+ve); FA×OT (+ve)
PF. NG	904	(879*, 1647)	PF. NG×IM (-ve); PF. NG×FA (+ve); PF. NG×OT (-ve); IM×FA (+ve); FA×OT (+ve)
PF. FA	1679	(912, 2446)	PF. FA×IM (+ve); PF. FA×OT (-ve); IM×OT (+ve); IM×NG (+ve)
PF. OT	1139	(879*, 1585)	PF. OT×NG (-ve); PF. OT×FA (+ve); IM×FA (+ve); NG×FA (+ve)

Table 5: Results for the Romanian data in terms of the MLEs and associated 95% CIs for the total population size given the model selected, and corresponding model selected in terms of interaction terms present with associated estimated sign of the interaction. The first row (denoted by a "-") gives the results of the complete 5-list analysis; the remaining rows are the results of combining list PF with each of the other lists. For the model description we denote the combined lists by the combined "dotted" abbreviations. When the lower bound of the confidence interval was truncated to the number of observed individuals, it is indicated by *.

we repeat the model search algorithm to identify the model deemed optimal using the AIC statistic, and estimate the associated population size and associated 95% CI. We then remove each list in turn and repeat the analysis; before combining list GP (which has the smallest expected overlap) with each of the other lists and again repeat the model-fitting process. Finally, within the simulation study to consider the impact of the model selection process we also fit the generating model, or an alternative form of the model when a list is omitted or lists are combined. When omitting lists, the alternative model corresponds to the generating model but with all interactions involving the omitted list removed; for combined lists for the alternative model we include all possible two-way interactions (there are six in total). Note that we only use the simulated datasets for which we do not observe any potential identifiability problems to remove any possible confounding errors entering the simulation study. Thus, 30 percent of the simulated models in removing lists, and 55 percent of models in combining lists are used. For further discussion on identifiability, see for example, Vincent *et al* (2019).

4.1 Omitting lists: "leave-one-out"

For each simulated dataset we calculate the ratio of both the population estimate omitting the given source to the estimated total using all 5 lists; and the true population size (11313). We plot these estimates against two further statistics corresponding to (i) the proportion of the total number of observed individuals by the source that is subsequently omitted; and (ii) the proportion of overlap for the given list that is omitted (i.e. the proportion of individuals observed by the given list that are also observed by at least one other list). These results are plotted in Figure 1, where the left-hand plots, (a) and (c), correspond to the associated population size ratio for the estimate using all 5 lists plotted against (i); and the right-hand plots (b) and (d), correspond to the population size ratio for the estimate with the true simulated population size plotted against (ii). The black dots show the same quantities for the original UK data.

The relationships observed in the plots are similar when considering the true population size; or the estimated population size using all 5 lists (i.e. the columns in the figure are similar): although, the variability appears to be slightly greater when using the true population size. In general, the greater the proportion of individuals observed by a given list, then omitting that list leads to a greater variability in the estimate of the population size. Further, within this simulation study the variability of the estimates appears to be more dependent on the number of individuals observed by the given list that is omitted, as opposed to the proportion of overlap for that list - this is demonstrated by relatively similar estimates for LA and GP which observe the smallest number of individuals but have very different overlap patterns. Finally, we comment that there does not appear to be any structural over or under estimate of the population size when omitting any of the particular lists. However, we do note that underestimates do have a lower bound (i.e. the total number of individuals observed by the sources); whereas overestimates have no such bound and thus overestimates may be larger in magnitude.

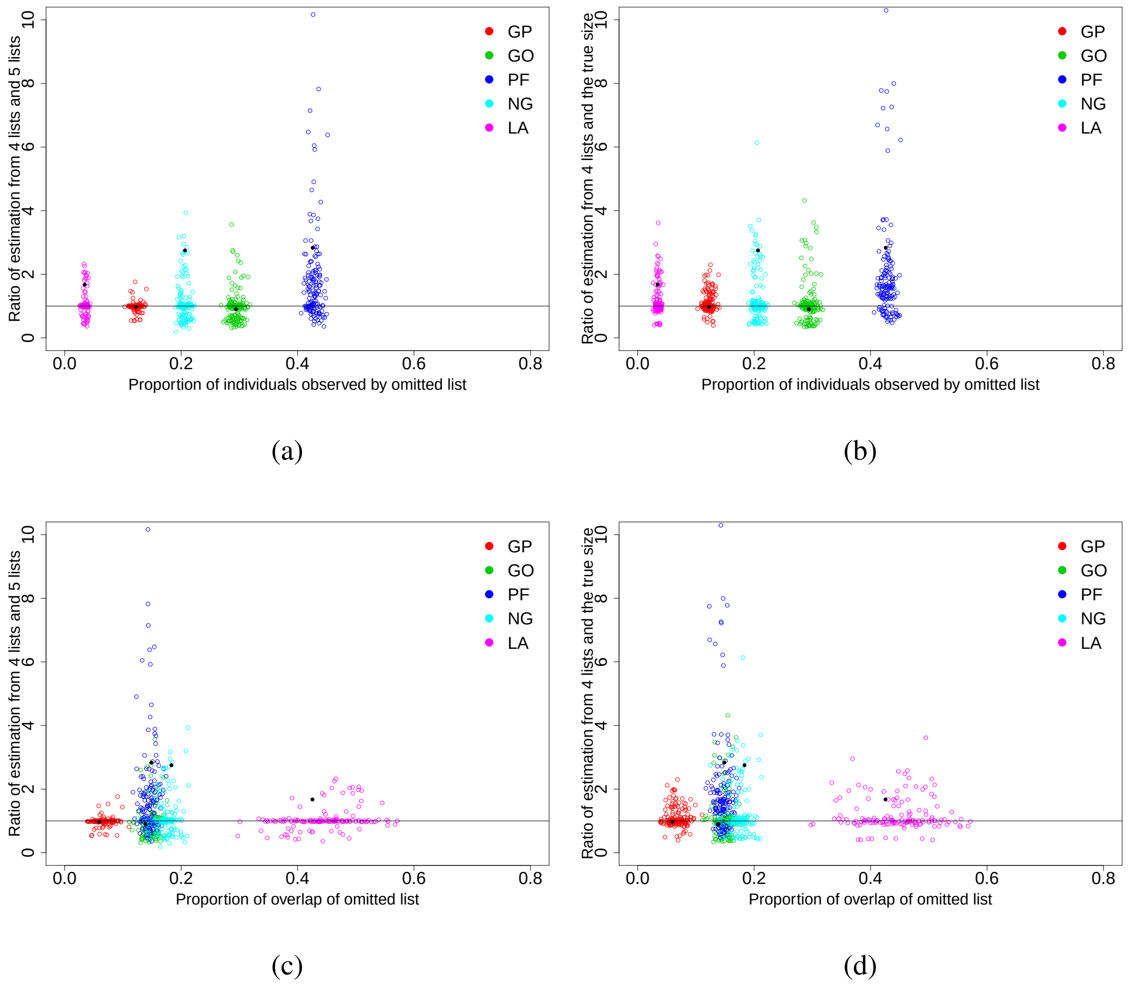
To investigate further the performance of the estimates, we consider the 95% CIs of the estimated population sizes and compare these with the true simulated population size. When considering all 5 lists, 69% of the 95% CIs contained the true value of the parameter. This is less than the nominal 95% level that we would expect and would indicate perhaps that there are further potential issues (for example, relating to model selection; see below for further discussion). However, we are primarily concerned with the impact of omitting each list, and thus we use this 69% as a comparison when we subsequently omit each list. The coverage probabilities in each case correspond to: 69%, 63%, 41%, 58% and 63% when removing GP, GO, PF, NG and LA, respectively. Further, the median of the length of these CIs for the models with 5 lists is 3341. Similarly, the median of the length of the 95% CIs after removing GP, GO, PF, NG, LA is respectively 3451, 5076, 12058, 4277, 3402. Thus, omitting the list GP leads to very similar performance as the full 5 lists (in terms of coverage and precision of the estimate) and suggests that the additional information that this list provides is minimal. Omitting lists GO, LA and NG leads to a relatively similar reduction in performance in terms of reduced coverage probabilities and precision. However, omitting list PF leads to a significant decrease in performance - this list also corresponds to the list that observes the greatest number of individuals.

Finally, we consider the impact of the model selection process by simply using the generating model or alternative form of the model. For the generated and alternative models for the reduced 4-list sources when omitting a list, the coverage probabilities were significantly higher and equal to 97% with using the five lists and 97%, 93%, 100%, 96% and 91% when removing GP, GO, PF, NG and LA, respectively. Further, the median of the length of the 95% CIs are increased to: 14483 with 5 lists, 15049, 22105, 55400, 31362, 20781 after removing GP, GO, PF, NG and LA. Thus model selection has a significant impact on the performance of the MSE approach - we return to this issue in Section 5.

4.2 Combining lists

For each simulated dataset, the list GP is combined with each of the other four lists in turn and the associated total population size is estimated. Figure 2 provides the corresponding plots of the ratio of the estimated population size using the combined lists compared to the estimated total using all 5 lists (in the left-hand plots); and the true population size used to simulate the data (in the right-hand plots), compared to the percentage of overlap of the source GP with the list it is combined with. The black dots show the same quantities for the original UK data. As for the above case of omitting the lists there is greater variability in the ratio of the estimated population size with the true value, compared to the case when we consider the estimated value using all 5 lists. Interestingly, within this simulation study there appears to be a clear and consistent overestimate of the total population size when we combine the GP list with the LA list - for which in the real UK data there was no overlap observed. However, combining the list GP with the other lists (GO, PF and NG) appears to provide less biased estimates of the total population size, and a reduced level of variability in the ratios. There also appears to be a slight decrease in the variability of the estimated ratio as the proportion of overlap of the combined lists increases.

For the set of retained datasets, the 95% CIs for the estimated population size using all the 5 lists, include the true value of the population size in 67% of the simulated datasets, with a median length of 3281. After combining GP with LA, NG, PF, GO this coverage probability is reduced to 23%, 51%, 33,% and 54%, respectively. The median length of



(a) (b)

(c) (d)

Figure 1: Ratio of estimated total population size using only 4 of the lists to the estimate obtained using all 5 lists plotted against proportion of individuals observed by omitted list (a) or proportion of overlap of omitted list (c); and similar plot for the ratio of estimated total population size against true simulated value plotted against proportion of individuals observed by omitted list (b) or proportion of overlap of omitted list (d). The black dots show the same quantities for the original UK data.

the 95% CIs were 7480, 4033, 3814, 4058, respectively after combining GP with LA, NG, PF and GO. Thus combining GP with each of the other sources leads to substantially worse performance in terms of coverage probabilities, particularly for LA and PF. With regard to LA (for which this has very small overlap across the simulations), not only does combining the GP list with the LA list lead to poor estimation of the total population size (i.e. a general overestimate and substantially reduced confidence interval performance) but the uncertainty of the estimate is also relatively large. Finally, to provide some insight into the impact of model selection within the analyses we also consider the generating and associated alternative models. In these cases the coverage probability are significantly increased to 90% (for the generating model) when using the five lists and 92%, 97%, 94% and 93% after combining GP with LA, NG, PF, and GO, respectively, for the reduced model. The corresponding median length of the 95% CIs are also increased to 13310 when using all 5 lists, and 24055, 15702, 15868, 15183 after combining GP with LA, NG, PF, and GO, respectively. This is a similar observation as for the case of omitting lists but without any a large increase in the size and variability of the length of the CIs.

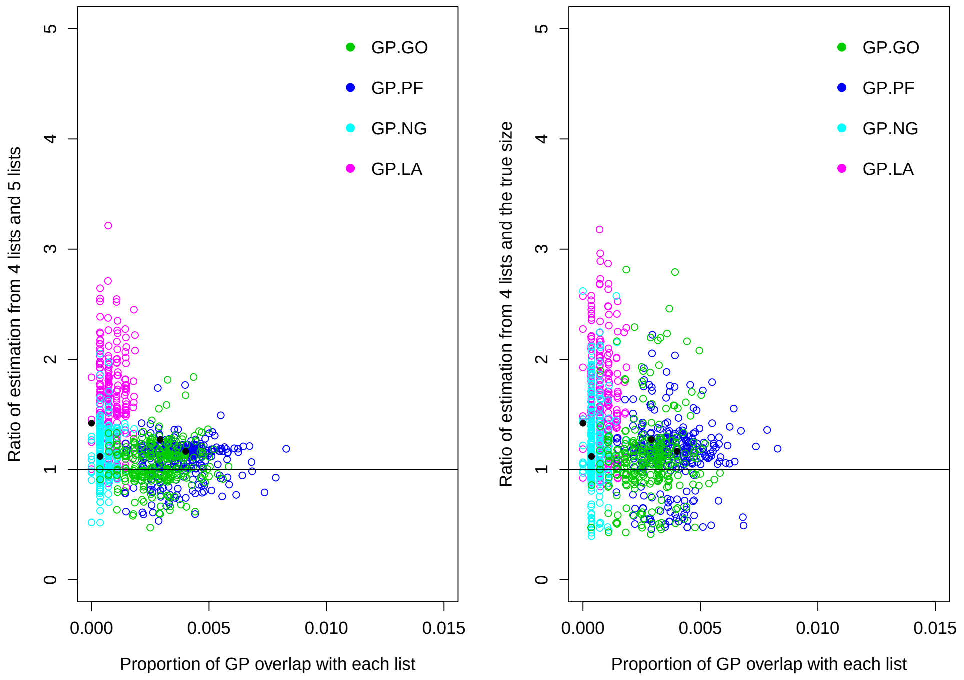


Figure 2: Ratio of estimated population size using 4 lists with GP combined with each source in turn with (i) the estimated population size using all 5 sources (on the left-hand side); and (ii) the true population size used to simulate the data, compared with the percentage of the list GP overlap with the given list it is combined with. The black dots show the same quantities for the original UK data.

The simulation studies suggest that the population size estimates can be sensitive to a number of different factors, including the number of sources that we include in the analysis and how we define a single source (i.e. a combined source). In general, assuming that we fit the generating model or the alternative version of this model (when we omit or combine a list) the corresponding population size estimates appear to be reasonable with generally good coverage probabilities. However, when adding the associated model search algorithm (using the AIC statistic as the criteria) the performance drops significantly and also appears to overestimate the precision of the resulting estimates.

5 Discussion

Collecting data from the different (and potentially diverse) sources and conducting the collation across the different lists requires resources. These resources may be limited, for example, in terms of person time or money. Thus, understanding the importance of different lists can have a direct impact on future data collection, and allocation of resources. Questions may particularly be raised in relation to sources that, for example, observe only a relatively small number of individuals, or those which have minimal overlap with other sources, since MSE relies on overlap between courses in order to estimate the total population size. This latter situation is very common in modern slavery applications - in this paper we considered the robustness of MSE in the case of small overlap between sources.

To reduce the minimal overlap between sources two approaches can be adopted: remove a source; or combine a source with another. This latter step may be done prior to any analysis being conducted, as may be done not only where there is minimal overlap but also in the opposite case where the overlap is substantial as was the case with UK data where the police force data was combined with the National Crime Agency data (Silverman, 2014). The analyses conducted within this paper suggests a note of caution with regard to the application of MSE to modern slavery data. In particular changes to the lists (omitting a list or combining two lists) could potentially have a significant impact on the total population size estimate - although combining lists appeared to have a lesser effect than simply omitting a list. Overall the model selection algorithm implemented - and in particular the use of the AIC statistic commonly used within MSE approaches Coumans *et al* (2017); Silverman (2014); Van der Heijden *et al* (2012) - had a significant effect on the performance of the MSE. It is possible that several competing models may be regarded to fit the observed data equally well but yet have very different estimates for the population size. These observations lead us to make the following minimum recommendations when implementing an MSE approach:

1. Fit multiple models to the data to investigate the sensitivity of the estimates to the different models - this would particularly include "similar" (i.e. neighbouring) models;

2. Investigate the robustness of the estimate by omitting each source in turn and repeating the analysis;

3. Combine pairs of sources together and again investigate the robustness of the parameter estimates; and

4. Conduct a simulation study to gain an understanding of the performance of the analyses (for example, using the MLEs of the fitted model as the generating model, as for the simulation study conducted within this paper based on the UK data).

The above aim to provide a greater understanding of the particular dataset and analysis. If similar estimates are obtained under the different scenarios there is some reassurance in the approach being robust. However, deviations may indicate some particular interesting aspect of the data. For example in our case for PF from the Romanian data lead to a significant decrease in the population estimate - and on inspection this was most likely due to the large number of (unique) individuals observed by this source. This in turn may be investigated to understand why so many individuals are only observed by PF, for example.

The simulation study suggests that the use of the AIC statistic as the model selection criteria may not be optimal, leading to poor coverage estimates of the true population size and over-confidence in the estimate. Alternative criteria exist, such as the Bayesian information criterion (BIC; Schwarz (1978)) and Focused information criterion (FIC; Claeskens and Hjort (2003)). Thus these different criteria could also be investigated within the exploratory analyses and added to the list of recommendations above. Further, with regard to model selection, an additional approach to consider includes a model-averaging approach, and thus removing the reliance on a single model. A weighted average over the set of plausible models can be calculated, so that the population size estimate includes both parameter *and* model uncertainty. See for example, Buckland *et al* (1997) in the classical framework and Hoeting *et al* (1999); King and Brooks (2001); Madigan and York (1997) in the Bayesian framework. If the set of plausible models all provide similar estimates of the total population size, then so too will the model-averaged estimate; however if the estimates differ between models the model-averaged approach will provide a weighted point estimate but typically have an associated significantly larger uncertainty interval to convey this additional uncertainty. In this latter circumstance it is useful to not only provide a single model-averaged estimate but also the set of most likely models and their associated estimate.

Another issue that we have not considered within this data analyses but may arise relates to cross-referrals, where one (or more lists) may refer individuals to other agencies but not vice-versa leading to asymmetry. For example, cross-referrals by another list to police may almost always be made when a child is or has been at risk. Cross-referral is also more likely when there is the prospect that an intelligence-led police raid could lead to the rescue of a clutch of other victims of modern slavery (Bird *et al*, 2019). Reports on MSE estimation of modern slavery, such as in Serbia and Ireland (23 Romanian men exploited in a waste recycling plant), for the United Nations Office on Drugs and Crime mention the context of annual counts for rescued victims being inflated by a particularly successful police operation. More generally, we acknowledge that MSE needs to evolve to take into account the underlying networks by which victims came to be listed. For example, a rescued victim may provide information leading to the rescue of other individuals, so that individuals are not independent of each other. Hence, in addition to list-membership and (selective) cross-referrals, consideration may also need to be given to the size and context of the rescued victim-network that selective cross-referral gives rise to. The current presentation of the data in terms of simply the presence of individuals on different lists discards the temporal information, so that it is not possible to take into account (or estimate) referrals between lists; or possible relationships between identifications. Worthington *et al* (2019) discuss similarities with ecological capture-recapture data where such temporal information is available and could provide insight/motivation for extended MSE models if such temporal information is available for modern slavery data. The challenges of modern slavery motivates further developments of MSE to incorporate the above particular complexities of the different processes acting on the and between the different lists used to identify victims.

References

Akaike, H. (1974) A New Look at the Statistical Model Identification, *IEEE Transactions on Automatic Control*, 19, 716–723.

Buckland, S. T., Burnham, K. P. and Augustin, N. H. (1997) Model Selection: An Integral Part of Inference. *Biometrics*, **53**, 603-618.

Bird, S. M. and King, R. (2018) Multiple Systems Estimation (or Capture-recapture Estimation) to Inform Public Policy. *Annual Review of Statistics and its Application*, **5**, 95-118.

Birds, S. M. *et al*. (2019) Public Health Perspective on UK-identified Victims of Modern Slavery. *Technical Report*

Chan, L., Silverman, B. and Vincent, K. (2019) Multiple Systems Estimation for Sparse Capture Data: Inferential Challenges when there are Non-Overlapping Lists. *ArXiv preprint: 1902.05156.*

Claeskens, G. and Hjort, N. L. (2003) The Focused Information Criterion (with discussion). *Journal of the American Statistical Association*, **98**, 879–899.

Cormack, R. M., Chang, Y-F., Smith, G.S. (2000) Estimating Deaths from Industrial Injury by Capture-recapture: A Cautionary Tale. *International Journal of Epidemiology*, **29**, 1053-1059.

Cruyff, M., J. van Dijk, and P. G. M. van der Heijden (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.

Coumans, A. M., Cruyff, M., Van der Heijden, P. G. M., Wolf, J. and Schmeets, H. (2017). Estimating Homelessness in the Netherlands Using a Capture-Recapture Approach. *Social Indicators Research*, **130(1)**, 189-212.

Davison, A. C. (2003) *Statistical Models*. Cambridge University Press.

Dellaportas, P. and Forster, J. J. (1999) Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. *Biometrika*, **88**, 317-336.

Fienberg, S. E. (1972) The Multiple Recapture Census for Closed Populations and Incomplete 2^k Contingency Tables. *Biometrika*, **59**, 591–603.

Fienberg, S. E. and Rinaldo, A. (2012) Maximum likelihood estimation in log-linear models. *Annals of Statistics*, **40**, 996-1023.

Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. (1999) Bayesian Model Averaging: A Tutorial. *Statistical Science*, **14**, 382-401.

Home Office (2018). 2018 UK annual report on modern slavery. `https://www.gov.uk/government/publications/2018-uk-annual-report-on-modern-slavery`.

Jewell, N. P., Spagat, M. and Jewell, B. L. (2013) Multiple systems estimation and casualty counts: Assumptions, interpretations and challenges. In Seybolt, T., Aronson, J. and Fischoff, B., (Eds.), *Counting Civilian Casualties: An Introduction to Recording and Estimating Nonmilitary Deaths in Conflict*. Oxford University Press, 185-211.

Jones, H. E., Hickman, M., Welton, N. J., De Angelis, D., Harris, R. J. and Ades, A. E. (2014) Recapture or Precapture? Fallibility of Standard Capture-Recapture Methods in the Presence of Referrals Between Sources. *American Journal of Epidemiology*, **179**, 1383-1393.

King, R., Bird, S. M., Brooks, S. P., Hutchinson, S. J. and Hays, G. (2005) Prior information in behavioural capture-recapture methods: demographic influences on drug injectors' propensity to be listed in data sources and their drugs-related mortality. *American Journal of Epidemiology*, **162**, 1-10.

King, R. and Brooks, S. P. (2001) On the Bayesian Analysis of Population Size. *Biometrika*, **86**, 615–633.

Madigan, D. and York, J. C. (1997) Bayesian Methods for Estimation of the Size of a Closed Population. *Biometrika*, **84**, 19–31.

Schwarz, G. E. (1978), Estimating the Dimension of a Model. *Annals of Statistics*, **6**, 461-464.

Sharifi Far, S. (2017) Parameter Redundancy in Log-linear Models. *PhD Thesis*, University of St Andrews.

Sharifi Far, S., Papathomas, M. and King, R. (2019) Parameter Redundancy and the Existence of Maximum Likelihood Estimates in Log-linear Models. *Arxiv preprint: 1902.10009*.

Silverman, B. (2014). Modern Slavery: An Application of Multiple Systems Estimation. `https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/386841/Modern_Slavery_an_application_of_MSE_revised.pdf`.

Silverman, B. (2019). Model fitting in Multiple Systems Analysis for the quantification of Modern Slavery: Classical and Bayesian approaches. *Journal of the Royal Statistical Society: Series A*, in press.

Van der Heijden, P. G. M., Whittaker, J., Cruyff, M. J. L. F., Bakker, B. and Van der Vliet, R. (2012). People born in the Middle East but residing in the Netherlands: Invariant population size estimates and the role of active and passive covariates. *Annals of Applied Statistics*, **6(3)**, 831-852.

van Dijk, J. J., M. Cruyff, P. G. M. van der Heijden, and S. L. J. Kragten-Heerdink (2017). Monitoring Target 16.2 of the United Nations Sustainable Development Goals; a multiple systems estimation of the numbers of presumed human trafficking victims in the Netherlands in 2010-2015 by year, age, gender, form of exploitation and nationality. United Nations Office on Drugs and Crime. *Research Brief, available at* `https://tinyurl.com/y9mpkach`

Vincent, K., Sharifi Far, S., and Papathomas, M. (2019) Common Methodological Challenges Encountered with Multiple Systems Estimation Studies. *Technical Report.*

Worthington, H., McCrea, R. M., King, R. and Vincent, K. (2019) How ideas from ecological capture-recapture models may improve multiple systems estimation analyses. *Technical Report.*