ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Statistics in Society
Series A

J. R. Statist. Soc. A (2020)

Multiple-systems analysis for the quantification of modern slavery: classical and Bayesian approaches

Bernard W. Silverman

University of Nottingham, UK

[*Read before* The Royal Statistical Society *on Wednesday, November 13th, 2019*, Professor R. Henderson *in the Chair*]

Summary. Multiple-systems estimation is a key approach for quantifying hidden populations such as the number of victims of modern slavery. The UK Government published an estimate of 10000–13000 victims, constructed by the present author, as part of the strategy leading to the Modern Slavery Act 2015. This estimate was obtained by a stepwise multiple-systems method based on six lists. Further investigation shows that a small proportion of the possible models give rather different answers, and that other model fitting approaches may choose one of these. Three data sets collected in the field of modern slavery, together with a data set about the death toll in the Kosovo conflict, are used to investigate the stability and robustness of various multiple-systems-estimate methods. The crucial aspect is the way that interactions between lists are modelled, because these can substantially affect the results. Model selection and Bayesian approaches are considered in detail, in particular to assess their stability and robustness when applied to real modern slavery data. A new Markov chain Monte Carlo Bayesian approach is developed; overall, this gives robust and stable results at least for the examples considered. The software and data sets are freely and publicly available to facilitate wider implementation and further research.

Keywords: Hidden populations; Human trafficking; Markov chain Monte Carlo methods; Public policy; Thresholding

1. Introduction

The original motivation for this work came from the estimation of the number of 'potential victims of human trafficking' in the UK, based on the National Crime Agency (NCA) strategic assessment of 2013. This was part of the strategy leading to the Modern Slavery Act 2015. See Silverman (2014) and Bales *et al.* (2015). The method used was multiple-systems estimation.

Quantifying modern slavery has crucial importance for policy. For example Cockayne (2015) has written

> 'without good data on where slaves are, how they become slaves and what happens to them, anti-slavery policy will remain guesswork'

and went on in this context to cite the use of multiple-systems approaches as a significant innovative approach in a field where good quantification is in its infancy. It is not just in narrow policy terms that good prevalence estimates are important; they also play a vital role in raising the public and political consciousness of modern slavery.

Multiple-systems estimation is a development of the classical capture–recapture approach and has been used in many contexts, such as counting casualties in armed conflicts (Manrique-

Address for correspondence: Bernard W. Silverman, University of Nottingham School of Politics and International Relations, Law and Social Sciences Building, University Park, Nottingham, NG7 2RD, UK.
E-mail: bernard.silverman@stats.ox.ac.uk

Vallier *et al.*, 2013) and numbers of injecting drug users (King *et al.*, 2013). Cases that come to light are recorded on a number of lists. By identifying cases across the various lists, the numbers that fall on each possible combination of lists are tabulated. Then a mathematical model is used to estimate the 'dark figure' of cases that have not come to attention and so are not recorded on any list survey, For an overall survey, see Bird and King (2018).

Crucial to this approach is the choice of model, in particular deciding which interactions or correlations to allow between the various lists. Some methods choose a particular model, whereas others seek a model averaging approach. This paper reviews several methods and investigates their performance on a range of real data sets. There is a deliberate focus on data collected in the area of modern slavery and human trafficking, because the primary aim of this paper is to develop methodology that is relevant to that area. In addition one of the data sets considered, drawn from the wider human rights area, relates to deaths in the Kosovo conflict in 1999. The choice of existing methods for discussion and review is again guided by our particular context, focusing on methods that have already been proposed for the multiple-systems analysis of human rights and modern slavery data.

The modern slavery context presents particular challenges for the use of multiple-systems analysis. No true prevalence or 'ground truth' is available to investigate the accuracy of any estimates, and so we need to assess other properties of estimation methods. For example, it is clearly desirable to have reasonable stability under operations such as combining or omitting lists with small counts or adjusting model parameters. Also, if multiple-systems estimation is to be used more widely to quantify modern slavery, it is important to consider the performance of the various possible approaches specifically on data sets of the kinds that are likely to be observed. Furthermore it may be important that there should be an agreed standard approach, at least as a starting point for more detailed investigation, and it is hoped that our detailed comparative study may contribute to that.

Another issue that must be borne in mind is the extremely sensitive nature of the data. Typically, much as we would like more details, such as covariate information, about the individuals observed in the study, these are not available to the statistical analyst. Without giving assurances of confidentiality to individual victims, for example, it would often not be ethical or even possible to collect their data. Collation of data between lists naturally involves sharing or matching information, but this is often done by a trusted individual who cannot reveal any details. Indeed, on some occasions all details of the lists themselves, and even of the type of organization that provided particular lists, must be obfuscated.

Our comparative study using real data sets and the methods so far proposed will demonstrate that, unfortunately, all the existing methods display instabilities of various kinds, sometimes dramatic, when tested on the real data sets. To address this issue, we introduce a Bayesian–thresholding approach that places prior distributions on the individual terms in the standard model.

In Section 2 of this paper, the various data sets are reviewed and tabulated. Section 3 sets out the standard Poisson model which underlies various possible approaches. Section 4 then examines frequentist approaches to model selection, including that used by Silverman (2014). Two other, rather different, Bayesian methods have been proposed and these are investigated in Section 5. In Section 6 our proposed Bayesian–thresholding method for the Poisson model is introduced. This casts the problem in a form where a standard Markov chain Monte Carlo (MCMC) package can be used to estimate the parameters, but there are some mathematical aspects that have to be taken into account for this to work. The method is demonstrated on the various data sets; it appears to avoid some of the gross instabilities that can arise with the existing methods but still requires care in its application. Finally, some conclusions are drawn in Section 7.

A key factor in developing a standard approach is the open accessibility of data and of methodology. All the data sets, together with R software to implement the methodology that is described in this paper, and to reproduce its results, are given in Silverman (2018a). For some additional remarks about the importance of open data and open research, see Silverman (2018b).

2. The data sets

The full data that were analysed by Silverman (2014), broken down into six lists, are summarized in Table 1.

Some of the methods that we consider do not deal with more than five lists, and so for some

Table 1. Potential victims of trafficking in the UK, 2013: numbers of cases on each possible combination of lists†

LA	*NG*	*PF*	*GO*	*GP*	*NCA*	*Count*
×						54
	×					463
		×				907
			×			695
				×		316
					×	57
×	×					15
×		×				19
×			×			3
	×	×				56
	×		×			19
	×			×		1
	×				×	3
		×	×			69
		×		×		10
		×			×	31
			×	×		8
			×		×	6
				×	×	1
×	×	×				1
×	×		×			1
	×	×	×			4
	×	×			×	3
		×	×		×	1
×	×	×	×			1

†LA, local authorities; NG, non-government organizations such as charities; PF, police forces, GO, government organizations such as the Border Force and the Gangmasters and Labour Abuse Authority; GP, general public, through various routes; NCA, National Crime Agency. For example there are 54 cases that appear only on the LA list, and 15 cases that appear on the overlap between LA and NG, but not on any others. There is one case that appears on all four of LA, NG, PF and GO but not on the other two. Those combinations of lists for which no cases were observed have been omitted from the table but are still taken into account in the analysis. From Bales *et al.* (2015).

purposes we shall combine the police force (PF) list with the NCA list to construct the 'UK five-list' data set. The NCA is not, strictly speaking, a police organization, but it has many powers and characteristics in common with police forces and so combining these two lists is the natural way to reduce to a smaller number.

In addition, the general public (GP) list raises issues because cases on this list may not always be specified in sufficient detail to allow for reliable matching with other lists. Therefore, at least to test for the robustness of any results, it will be helpful to consider, in addition to the full and five-list data sets, a 'UK four-list' data set constructed by omitting the GP list and combining the PF and NCA lists. The total number of observed cases is 2744 for the five- and six-list data, but only 2428 for the four-list data set.

A second important data set (van Dijk *et al.*, 2017; Cruyff *et al.*, 2017) comprises six lists for identified victims in the Netherlands for the period 2010–2015. The data are given in Table 2. For a five-list version of these data, we combine the two smallest lists I and O. The total number of observed cases in this data set is 8234.

Table 2. Victims of trafficking in the Netherlands: numbers of cases on each possible combination of lists, leaving out combinations for which no cases were observed†

I	*K*	*O*	*P*	*R*	*Z*	*Count*
×						352
	×					1299
		×				403
			×			4466
				×		650
					×	632
×		×				1
×			×			18
×				×		3
×					×	16
	×	×				1
	×		×			44
	×				×	4
		×	×			59
		×		×		2
		×			×	57
			×	×		82
			×		×	125
				×	×	2
×		×	×			4
×			×		×	4
		×	×	×		2
		×	×		×	7
			×	×	×	1

†The lists are as follows: P, National Police; K, Border Police; I, Inspectorate Ministerie Sociale Zaken en Werkgelegenheid (Ministry of Social Affairs and Employment); R, regional co-ordinators; O, residential treatment centres and shelters; Z, others (e.g. ambulatory care centres, organizations providing legal services and the Immigration and Naturalization Service). Constructed from van Dijk *et al.* (2017), Table 3.

The third example is constructed from data that were collected by eight agencies in the New Orleans–Metairie metropolitan statistical area (Greater New Orleans) and analysed by Bales *et al.* (2019). These include 185 individuals who interacted with law enforcement and service providers in Greater New Orleans during the year 2016. They are given in Table 3. The sensitivity among the various agencies, partly for legal reasons, means that it is not possible even to label the lists themselves informatively. No further information was available to the statistical analysis than the table itself, with lists labelled A–H. Where it is necessary to reduce the number of lists, a five-list data set is constructed by combining the lists with the four smallest counts into a single list BEFG.

Finally, we consider a data set from a different area of human rights: that of determining the numbers of victims of armed conflict. The data, due to Ball *et al.* (2002), relate to the numbers of those who were killed in Kosovo in a 3-month period in 1999. They are available within the R package `LCMCR` (Manrique-Vallier, 2017) and are reproduced in Table 4. This four-list data set, which includes 4400 known victims, displays high correlation between lists and has larger numbers in the higher order three-list and four-list overlaps than do the modern slavery examples. This is in the nature of the particular application and is highly unlikely to occur in any modern slavery data set.

3. Models and methods

In this section, we review the basic log-linear model as proposed by Cormack (1989). Suppose that we have K lists labelled $\{1, 2, \ldots, K\}$. For each subset A of $\{1, 2, \ldots, K\}$, let N_A be the

Table 3. Victims related to modern slavery and trafficking in New Orleans: numbers of cases on each possible combination of lists, leaving out combinations for which no cases were observed†

A	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*Count*
×								25
	×							5
		×						70
			×					33
				×				6
					×			6
						×		6
							×	21
×		×						1
×			×					2
×				×				1
	×				×			1
		×	×					1
		×		×				1
		×				×		1
			×	×				2
				×			×	1
×		×				×		1
×			×	×				1

†For confidentiality the lists are labelled uninformatively. From Bales *et al.* (2019).

Table 4. Killings in the Kosovo war from March 20th to June 22nd, 1999, grouped into four lists†

EXH	*ABA*	*OSCE*	*HRW*	*Count*
×				1131
	×			845
		×		936
			×	306
×	×			177
×		×		228
×			×	106
	×	×		217
	×		×	31
		×	×	123
×	×	×		181
×	×		×	18
×		×	×	42
	×	×	×	32
×	×	×	×	27

†All 15 observable combinations have a non-zero count. EXH, exhumations; ABA, American Bar Association Central and East European Law Initiative; OSCE, Organization for Security and Cooperation in Europe; HRW, Human Rights Watch. From Manrique-Vallier (2017).

number of cases that occur on all the lists in A but on no others. So, if $K = 6$ there are 64 possible subsets A, including the empty set $\emptyset$. The 'dark figure' is the number of cases $N_\emptyset$ that do not appear on any list.

Using the UK data as an illustrative example, Table 1 gives counts for only 26 subsets A, and the first step in the analysis is to reinstate all the rows in the table for which the observed count is 0, yielding 63 observations in all. There is no observed count for the dark figure.

The basic model is that each N_A has, independently, a Poisson distribution with parameter λ_A, with some structure on the λ_A. This is quite a strong assumption, because it assumes that the cases each behave independently of one another and obey the same probability laws of appearing on the various lists. Especially if there are observed covariates, the model will be only a jumping-off point for more detailed modelling, but it is at least a start. The model does not assume that the various lists are independent; interactions between the lists are allowed by appropriate modelling of the parameters λ_A.

Under the model, the dark figure $N_\emptyset \sim \text{Poiss}(\lambda_\emptyset)$. It is likely that the estimation error in $\lambda_\emptyset$ will be much larger than the Poisson variation, and so in practice the parameter estimate of $\lambda_\emptyset$ will be taken as the estimate of the dark figure, though if possible the Poisson variation should be taken into account as well. To obtain an estimate of the total population, the estimate of the dark figure is added to the total number of cases actually observed.

The Poisson model can also be seen as an approximation to a multinomial model where there is a fixed (unknown) total population size, and cases independently fall on the various lists or combinations of lists with probabilities proportional to the expected values under the Poisson model. Cormack (1992) provided a way of using the profile likelihood under the Poisson model set

out below, to obtain confidence intervals for the total population size under the multinomial model.

For the most part, the model that we shall investigate will be of the form

$$\log(\lambda_A) = \mu + \sum_{i \in A} \alpha_i + \sum_{\substack{i,j \in A \\ i<j}} \beta_{ij}. \tag{1}$$

For example, if $K = 6$ then there will be six main effects α_i and 15 two-list interactions β_{ij}, making 22 parameters altogether to be estimated from the 63 observable values N_A. Within this model, we have $\log(\lambda_\emptyset) = \mu$. Therefore the estimate of the dark figure is $\exp(\mu)$; we do not actually need estimates of the other parameters to estimate the dark figure.

There are basically two approaches to model fitting in this context. One is to use a model selection criterion to choose a particular set of parameters to fit, constraining all the others to 0. The other is to use some sort of model averaging approach, usually of a Bayesian nature.

4. Frequentist model selection

The package `Rcapture` (Baillargeon and Rivest, 2007) can be used as the basis of various approaches, which are explored in this section. The simplest is to set all interaction terms to 0, fitting main effects α_i only. Under this model, the lists themselves are independent, which is an assumption that may be unrealistic. Nevertheless this model may be a good reference point for more detailed analysis.

4.1. Adding parameters stepwise

In their original work on the UK data, Silverman (2014) and Bales *et al.* (2015) used a stepwise approach, starting with main effects only and then adding two-list interactions β_{ij} stepwise. At each step, the interaction that best improves the Akaike information criterion (AIC) is chosen. The process of adding interactions is stopped if the AIC cannot be improved by adding an interaction, or if the new interaction is not significant at some threshold. This variable-selection method is implemented within the R package `modslavmse` (Silverman, 2018a) and makes use of the package `Rcapture`.

Table 5 shows estimates and confidence limits by using main effects only, and the stepwise procedure with two different p-value thresholds, for the UK data summarized into six, five and four lists as set out in Section 2. The original work used the stepwise method with $p = 5\%$. Here and subsequently in this section, the confidence intervals are constructed from the profile likelihood using the approach of Cormack (1992) as implemented within `Rcapture`. Both the six- and the five-list data give a 95% confidence interval, conditionally on the model choice, of 10000–13000 in round terms. Using main effects only, or a more stringent criterion for adding parameters to the model, gives larger estimates. The results for the four-list case, in contrast, give smaller estimates, but none of these effects is dramatic.

4.2. Choosing from a large class of models by using an information criterion

The stepwise method is not the only possibility. Another approach is to fit all possible models, considering every subset of the interactions, and to choose between these by using some criterion. This can be done by using the routine `closedpMS.t` within the package `Rcapture`. If the full six-list data are used, then there are 2^{15} models even if only pairwise interactions are considered, which presents an excessive computational burden. To make the method computationally feasible in practice, the approach is applied to only the five-list data, allowing for

Table 5. Estimates and confidence intervals for the UK data, for the main effects model and for the stepwise AIC approach†

Data	*Estimates and confidence limits*				
	2.5%	*10%*	*Point estimate*	*90%*	*97.5%*
Main effects only					
UK 6 lists	11.0	11.4	12.2	13.1	13.6
UK 5 lists	12.0	12.5	13.4	14.5	15.2
UK 4 lists	9.5	9.9	10.7	11.6	12.1
Stepwise AIC, threshold p-value 0.1%					
UK 6 lists	12.6	13.1	14.2	15.4	16.1
UK 5 lists	12.6	13.1	14.2	15.4	16.1
UK 4 lists	10.5	11.0	12.0	13.1	13.8
Stepwise AIC, threshold p-value 5%					
UK 6 lists	10.0	10.4	11.4	12.5	13.2
UK 5 lists	*9.9*	*10.3*	*11.3*	*12.4*	*13.1*
UK 4 lists	9.6	10.0	11.0	12.1	12.8

†The figures are for the numbers of thousands of victims, rounded to the nearest 100. The row in italics corresponds to the analysis carried out by Silverman (2014).

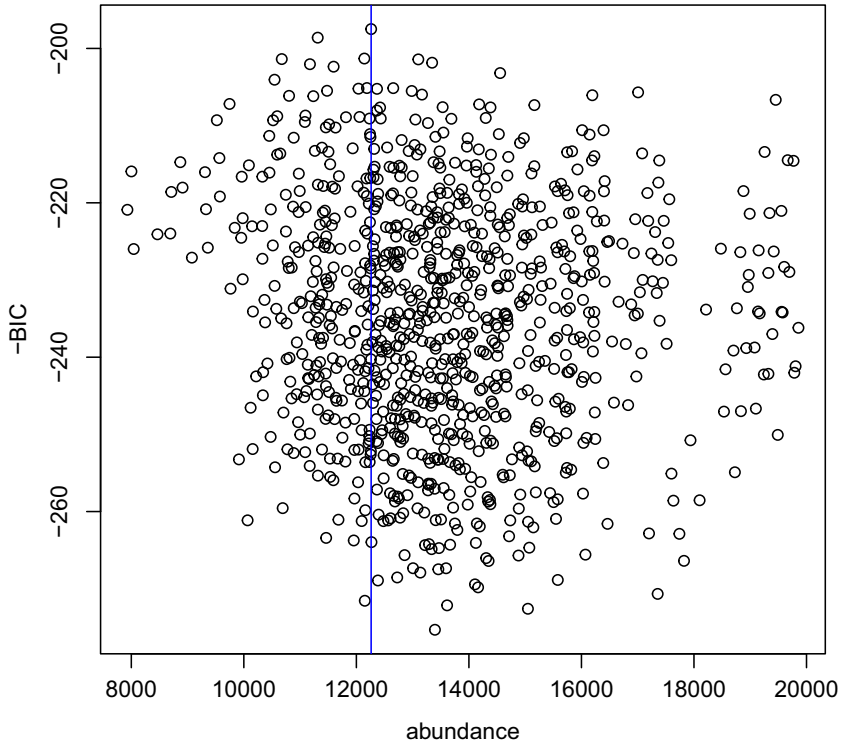

Fig. 1. Estimates of abundance plotted against the BIC, with outliers omitted, the default plot option

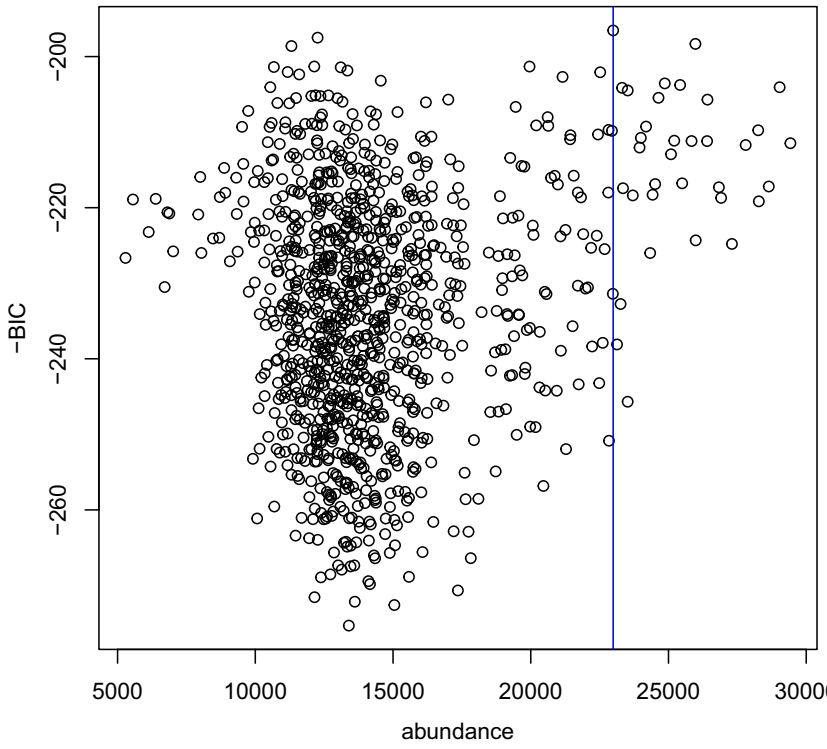

Fig. 2. Estimates of abundance plotted against the BIC, with outliers included

two-list interactions only, leaving 2^{10} models to be considered. The package `Rcapture` displays the results by using the Bayesian information criterion (BIC) rather than the AIC as the primary method of model choice. The BIC and AIC differ in the amount that they correct for parsimony of models, with the BIC having a heavier preference for more parsimonious models.

The default plot is shown in Fig. 1, with the vertical line showing the model with the lowest BIC (197.5). (The `Rcapture` routine plots −BIC and chooses the maximum of that.) The population size estimate for that model is 12262, with the estimates for other models clustering approximately around the Silverman (2014) estimate. However, setting the argument `omitOutliers = F` yields Fig. 2. There is a subsidiary cloud of results corresponding to a much larger estimate for the population size, and the estimate for the best BIC is actually within that cloud.

A closer examination of the top 10 models chosen by each of the BIC and AIC is instructive. There are no models in the top 10 for the AIC which yield estimates over 17000, and only one which yields an estimate that is much outside the range that was suggested in the original analysis. In contrast, the BIC chooses models yielding a much wider range of estimates. Overall, the results for the five-list data demonstrate that the estimate of the total population can vary considerably depending on the model that is chosen, and that even concentrating on well-fitting models, by some criterion, does not necessarily resolve this issue.

Because of the *caveats* about the GP list, the analysis was repeated for the four-list data. Fig. 3 shows all models and demonstrates that the cloud of points corresponding to the much larger estimate disappears altogether if the GP list is omitted.

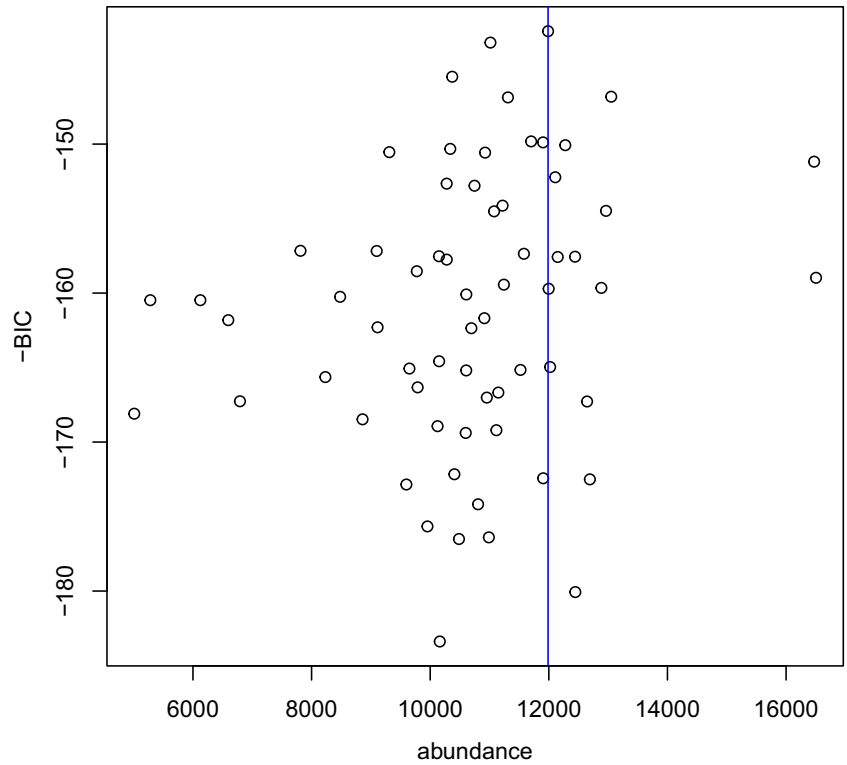

Fig. 3. Estimates of abundance plotted against the BIC, with the GP list excluded

4.3. Further examples

In Table 6 we present the results of applying the main effects only and the stepwise AIC choice methods to the other three example data sets. There is a somewhat alarming instability in the analysis of the Netherlands data; combining the two smallest lists more than doubles the stepwise estimates. There is no such instability if main effects only are fitted. Some further intuition may be gained from Fig. 4. There is a long tail of models with very large estimates. Although the globally optimal model according to the BIC is not in this group, the stepwise method is choosing one of these, indeed one yielding almost the largest estimate among all choices of model.

For the full New Orleans data with eight lists, the lower threshold for the p-value yields a very different estimate, indeed one where the profile likelihood does not allow an upper 97.5% confidence value, and a warning is generated by the routine within `Rcapture`. With a large number of lists and so many possible parameters to fit, it is not surprising that it should be inappropriate to use $p = 5\%$. All the other estimates are similar to estimates fitting main effects only, which are virtually unaffected by reducing to five lists.

The Kosovo data yield quite a different result if interactions are allowed. This is to be expected given the strong correlations that are evident in the data.

4.4. Identifiability and existence of estimates

The three data examples that were drawn from the study of human trafficking all give rise to contingency tables with some 0 cell counts. This raises issues discussed in generality by Fienberg and Rinaldo (2012a), and in our particular context by Chan *et al.* (2020).

Table 6. Estimates and confidence intervals for the Netherlands, New Orleans and Kosovo data†

Data	*Estimates and confidence limits*				
	2.5%	*10%*	*Point estimate*	*90%*	*97.5%*
Main effects only					
Netherlands	48.5	50.0	52.8	55.9	57.6
Netherlands 5 lists	48.6	50.0	52.9	56.0	57.8
New Orleans	0.7	0.7	1.0	1.4	1.7
New Orleans 5 lists	0.7	0.8	1.0	1.4	1.8
Kosovo	7.1	7.2	7.4	7.6	7.7
Stepwise AIC, threshold p-value 0.1%					
Netherlands	53.3	55.6	60.3	65.6	68.7
Netherlands 5 lists	119.4	127.8	146.0	167.8	181.0
New Orleans	0.7	0.7	1.0	1.4	1.7
New Orleans 5 lists	0.7	0.8	1.0	1.4	1.8
Kosovo	12.5	13.1	14.3	15.7	16.5
Stepwise AIC, threshold p-value 5%					
Netherlands	53.3	55.6	60.3	65.6	68.7
Netherlands 5 lists	119.4	127.8	146.0	167.8	181.0
New Orleans	1.4	1.8	3.4	7.2	∞
New Orleans 5 lists	0.7	0.8	1.0	1.4	1.8
Kosovo	12.5	13.1	14.3	15.7	16.5

†The figures are for the numbers of thousands of victims, rounded to the nearest 100.

One possibility is that there are no finite maximum likelihood estimates of all the parameters, but that the likelihood is maximized when one or more parameters tend to $-\infty$. This yields what Fienberg and Rinaldo (2012a) termed an *extended maximum likelihood estimate*, which gives a *bona fide* estimate, possibly 0, of each λ_A. This is handled within `Rcapture`, somewhat unsatisfactorily, by returning large negative estimates for some of the β_{ij}. These then give estimates for the resulting λ_A which are very close to 0. A consequence of this behaviour is that it is no longer possible to expand the log-likelihood as a quadratic approximation around the maximum, and hence the standard likelihood theory, including the justification of information-based criteria, breaks down. This breakdown is discussed further and illustrated in a small simulation example by Chan *et al.* (2020). Other aspects will be considered in Section 6.2 later.

There are two other estimability issues for maximum likelihood, neither of them addressed in `Rcapture`. One is that the extended maximum likelihood estimate does not exist; consideration of a small artificial example in Chan *et al.* (2020) shows that this may manifest itself as an infinite (or, numerically, very large) estimate of the dark figure. Fienberg and Rinaldo (2012b) showed in a very general context that non-existence of the extended maximum likelihood estimate can be checked by solving a linear programming problem, which is set out for our particular case in Chan *et al.* (2020). The other possibility is that, although the likelihood can be maximized, the parameters that attain this maximum are unidentifiable; this can be checked by finding the rank of a particular model matrix.

Chan *et al.* (2020) derive an efficient algorithm that can demonstrate (without actually checking every single model) whether either check would be failed by any choice of the set of interaction parameters β_{ij} to include in the model. For each of the data sets that are considered in

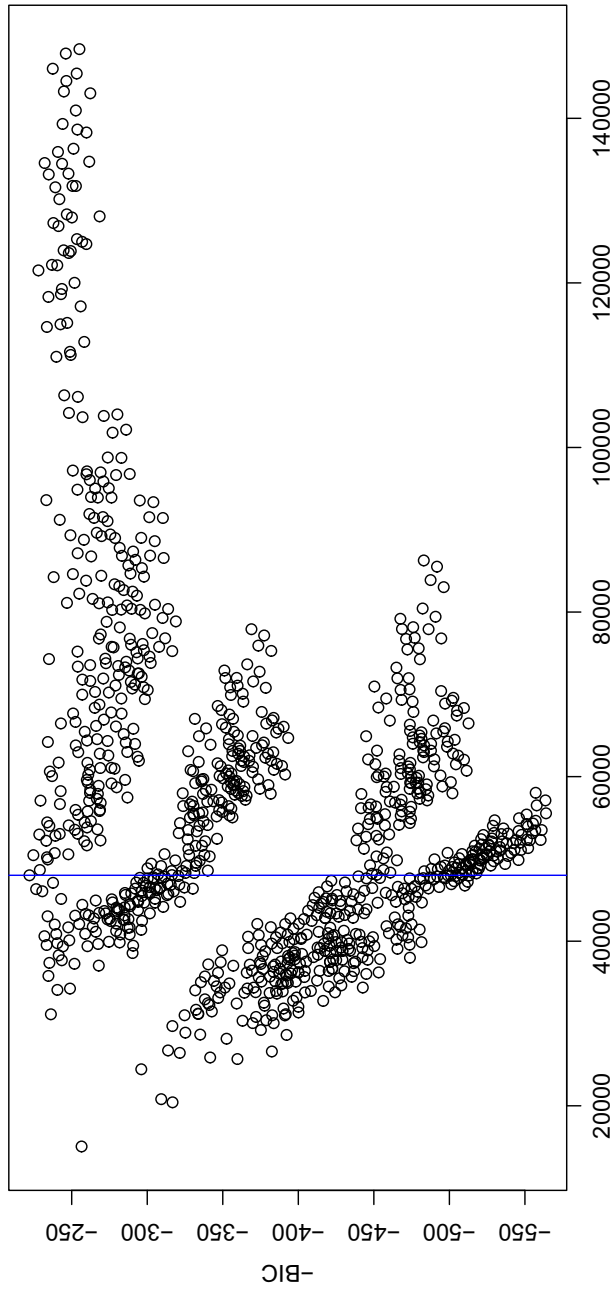

Fig. 4. Estimates of abundance plotted against the BIC, for the Netherlands data consolidated into five lists

this paper, every possible model passes the checks. Thus it seems unlikely that the instabilities and bimodalities in the estimation that are displayed in Section 4, and in some of the other methods discussed later in the paper, are due to the problems that were considered by Fienberg and Rinaldo (2012a).

5. Bayesian approaches

Two rather different Bayesian approaches have been proposed or developed specifically for human rights data. Their performance on our data sets is reviewed in this section. Unfortunately, neither method escapes the instabilities that have already been seen in the actual data sets.

5.1. Graphical models

A graphical models method was developed by Madigan and York (1997) and implemented in the package `dga` (Johndrow *et al.*, 2015). This uses every decomposable graph model of dependences between the various lists and obtains the joint posterior probabilities of the models and the total population size. The routine `bma.cr` which carries out the analysis requires an array of possible values of the dark figure. A reasonable standard range is from zero to 10 times the number of cases actually observed, but this will be discussed further below.

The routine is only fully implemented for three, four and five lists, where the numbers of possible models are 8, 61 and 822 respectively. The combinatorial burden becomes excessive if six or more lists are used. Therefore, the method is applied only on the five-list versions of the UK, Netherlands and New Orleans data, as well as on the Kosovo data and the four-list UK data.

For the UK data, initial application of the method on the five-list data showed a strong bimodal distribution which extended beyond the standard range, and so the calculation was repeated with the range for the total population extended to 40000. The results for both the five- and the four-list data are shown in Fig. 5. The dotted curves show the joint posterior probabilities of particular values for the total population size and individual models. The full curve is the sum of the dotted curves: in other words the marginal posterior distribution of the total population size. There are 822 dotted curves in Fig. 5(a) and 61 curves in Fig. 5(b). Most of the models have posterior probability very close to 0 for all values of the total population. The quantiles of the posterior distribution are given in Table 7, though of course in the case of the full five-list data these are not an adequate description of the bimodal distribution.

Now we turn to the Netherlands data, where the number of observed cases is 8234. Fig. 6(a) shows the posterior when calculated on the range of up to 10 times this figure for the dark figure. In contrast with the UK data, there is no suggestion of any second mode within this range. However, if the range is extended further, a noticeable mode appears, which has total posterior probability about 34%. The quantiles for the two estimates are given in Table 7.

Results are also given in Table 7 for the New Orleans and Kosovo data. In these cases the posterior distribution is definitely concentrated within the standard range. Interestingly, and in contrast with the other data that were considered, these two data sets illustrate two extremes of the method. For the New Orleans data, the largest posterior probability of any of the possible models is about 0.05, so no model is dominant, whereas, for the Kosovo data, one model has posterior probability nearly 0.99. The corresponding probabilities (for the extended ranges) are 0.44 for the UK data and 0.67 for the Netherlands data.

5.2. Dirichlet process mixtures

Another approach that has recently been proposed is a Bayesian latent class method (Manrique-

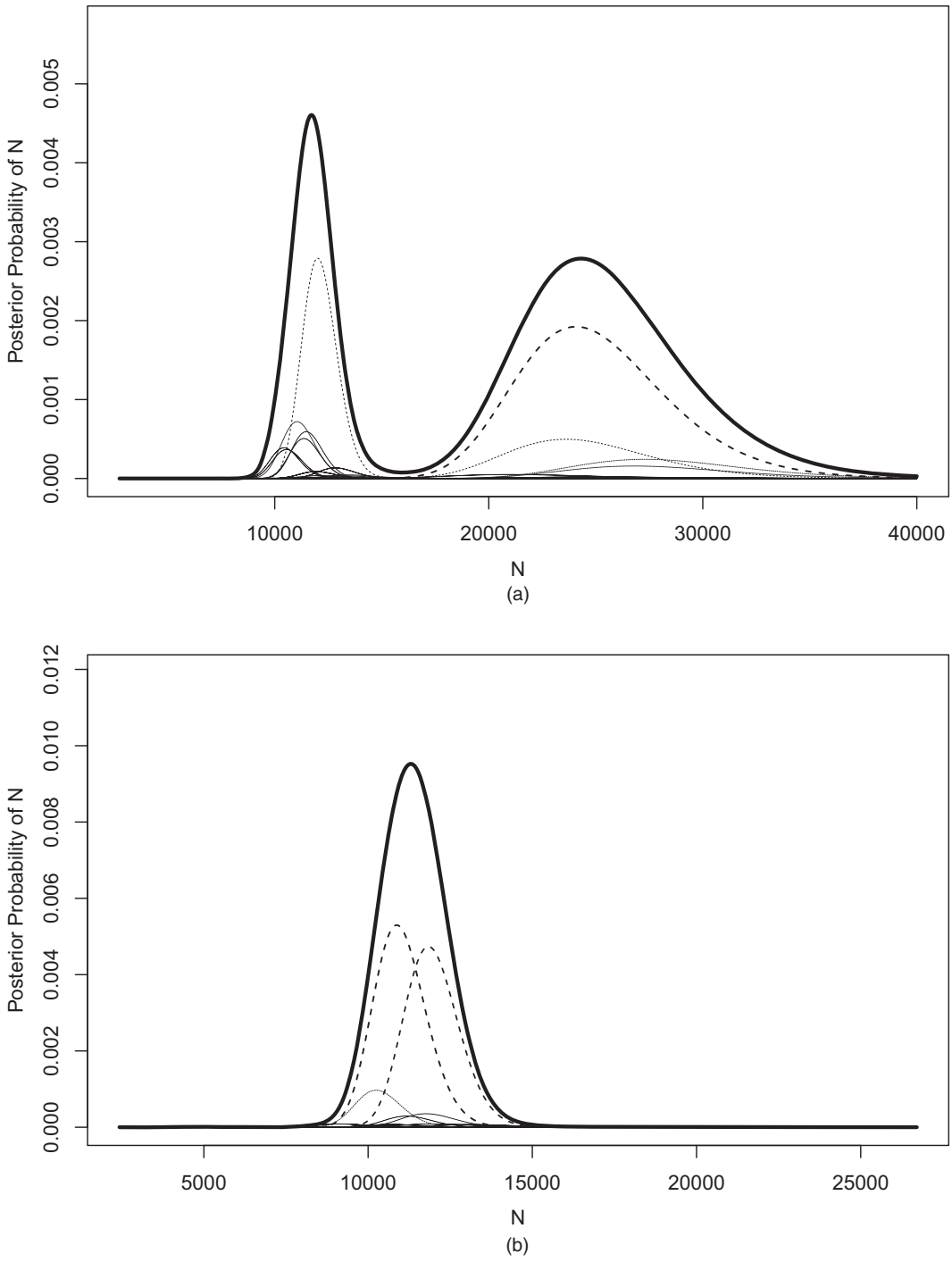

Fig. 5. Posterior distribution of the total population size for the UK data, for (a) the five-list and (b) four-list data, using the method of Madigan and York (1997): ———, averaged posterior probability; – – –, posterior probability by the model

Table 7. Quantiles of the posterior distribution by using the method of Madigan and York (1997)†

Data	*Maximum population*	*Quantiles of posterior*				
		2.5%	*10%*	*50%*	*90%*	*97.5%*
UK 5 lists	40000	10.4	11.3	23.0	29.6	33.3
UK 4 lists (GP omitted)		9.5	10.1	11.4	12.7	13.5
Netherlands 5 lists		40.8	43.5	47.9	52.6	55.4
Netherlands 5 lists	250000	41.5	44.3	50.3	177.4	202.8
New Orleans 5 lists		0.5	0.6	0.9	1.3	1.6
Kosovo		9.8	10.8	12.6	14.9	16.5

†Where the value of the maximum population is given in the table, the range of possible population estimates is extended to that value beyond the default.

Table 8. Quantiles of posterior distribution by using the Dirichlet process mixtures approach

Data	*Quantiles of posterior*				
	2.5%	*10%*	*50%*	*90%*	*97.5%*
UK six lists	17.2	18.8	23.0	29.5	34.2
UK five lists	15.1	17.4	22.0	28.8	35.2
UK four lists	10.1	10.7	12.0	13.6	14.5
Netherlands	115.7	126.1	150.3	189.9	250.3
Netherlands five lists	43.0	44.7	49.1	54.2	58.3
New Orleans	0.5	0.6	0.7	0.9	1.1
New Orleans five lists	0.6	0.6	0.8	1.1	1.3
Kosovo	8.5	9.4	10.4	12.3	14.6

Vallier, 2016). This is implemented in the R package `LCMCR` (Manrique-Vallier, 2017). It provides an MCMC estimate of the population size. In contrast with the method that was described in Section 5.1, there is no restriction on the number of lists. The results for the various data sets are shown in Table 8.

Because the output from the method is a Monte Carlo estimate, it is necessary to check whether there has been sufficient burn-in and also whether the output demonstrates sufficient mixing to be reliable. To ensure reproducibility the seed was set to 12345 rather than the default setting which yields different results each time. To ensure better mixing than the default, the parameter thinning was set to 100 and the burn-in value was set to 100000.

Comparing the two Bayesian methods of this section is instructive. For the UK data not omitting the GP list, the Dirichlet process approach essentially ignores the lower component of the posterior distribution that is found by the Madigan–York method and displayed in Fig. 5(a). Once the GP list has been omitted, the two methods give very similar results. For the Netherlands data, the Dirichlet approach homes in on the upper mode for the full data and the lower mode for the five-list case—the reverse of the behaviour of the AIC stepwise approach.

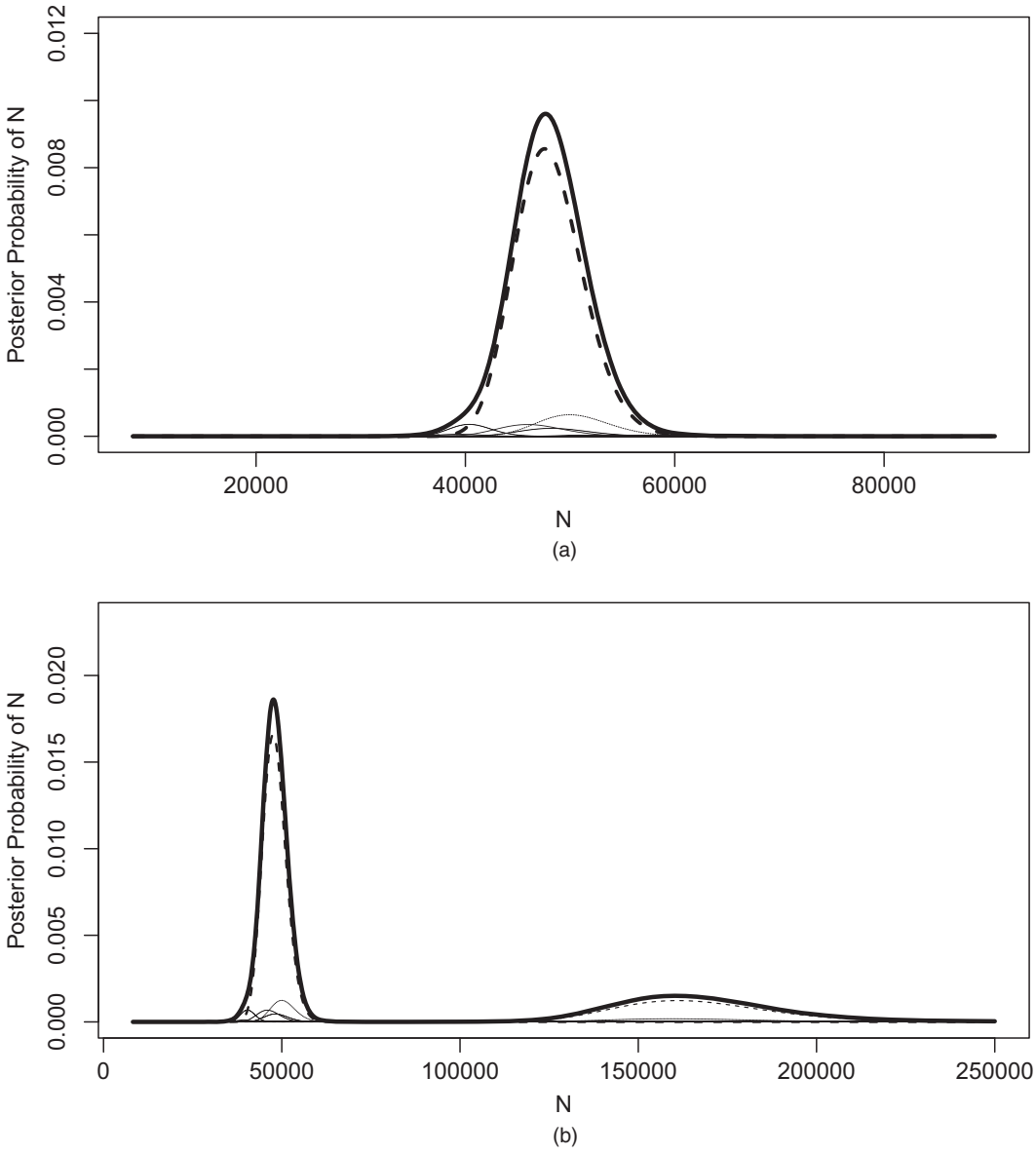

Fig. 6. Posterior distribution of the total population size for the Netherlands data, using the method of Madigan and York (1997) (———, averaged posterior probability; – – –, posterior probability by the model): (a) standard range of possible population size; (b) range extended to 250000

6. The Bayesian–threshold approach

6.1. Defining the prior and thresholding the results

In this section, we return to the Poisson log-linear model as specified in Section 3 and set out a Bayesian–threshold approach to fitting the model, dependent on two prior parameters, λ and τ. The first step of the model is to specify a prior which does not constrain the intercept parameter or the main effects, but allows for the prior to shrink the interaction parameters towards 0. In

the second step, those interactions for which there is no strong evidence that they are not 0 are dropped from the model, and the analysis repeated. The steps of the model are as follows.

Step 1: use a prior model under which

(a) the parameters μ, α_i and β_{ij} for all i and j are independent,
(b) μ and the α_i have uniform (improper) priors on $(-\infty, \infty)$ and
(c) the β_{ij} have a Gaussian prior with mean 0 and variance $1/\lambda$ for $\lambda \geqslant 0$. If $\lambda = 0$ this is interpreted as an improper uniform prior on $(-\infty, \infty)$.

In every case the R package `MCMCpack` (Martin *et al.*, 2011) and in particular the function `MCMCpoisson` enable MCMC sampling to be used to simulate from the posterior distribution. The improper uniform prior is the default for parameters within `MCMCpoisson`.
Step 2: constrain to 0 those β_{ij} for which the ratio of their posterior mean to their posterior standard deviation does not pass some threshold τ, and repeat the MCMC analysis with these β_{ij} omitted.

One justification for the thresholding step is that it is an approximation to a prior for the interactions which is a mixture of an atom of probability at zero and some other distribution, a prior which in other contexts leads to a thresholding approach; see, for example, Johnstone and Silverman (2004). The exact implementation of such a prior is a topic for future research. If $\tau = 0$ then no thresholding is carried out.

In broad terms, a case is $\exp(\beta_{ij})$ times more or less likely to be on both the lists i and j than if occurrence on the lists is independent. This interpretation makes it seem unlikely that values of β_{ij} that are much outside the range ± 1 should be contemplated, and so, if a Gaussian prior is used, the precision parameter λ might be chosen in the range 1–10.

Turning to the thresholding parameter, two different approaches will be investigated. The first is to take a 'liberal' view, to include interactions where they are not clearly spurious; this would suggest using a threshold parameter of something like 2. The other is to take a 'parsimonious' view, using a much larger threshold, so that interaction parameters will be included only if there is very strong evidence that they are not 0. For this approach we use a threshold of 5, admittedly chosen rather arbitrarily.

6.2. Implementation issues

Two implementation issues are taken into account in the package `modslavmse` (Silverman, 2018a). Firstly, the routine `MCMCpoisson` in `MCMCpack` does not appear to deal properly with the case where some of the parameters have an improper uniform distribution whereas others have finite variance, so if $\lambda > 0$ the calling routine `MCMCfit` in `modslavmse` gives the intercept and main effects a prior with large finite variance 10^4. Note, in passing, that a proper Bayesian approach will avoid the issues that were considered in Section 3.3.4 because there will necessarily be a well-defined posterior distribution for the parameters.

If an improper prior is used ($\lambda = 0$) for the interaction parameters, then some care is needed. Consider the UK data as in Table 1. No cases fall in both local authority (LA) and GP lists, whether or not in combination with other lists. If the improper uniform prior is used for the corresponding interaction parameter $\beta_{\mathrm{LA,GP}}$, then we show that the posterior distribution of $\beta_{\mathrm{LA,GP}}$ is concentrated at $-\infty$ and set out the way that the other parameters can be estimated by MCMC sampling. This is an instance where the maximum likelihood approach leads to an extended maximum likelihood estimate of the parameter; see Chan *et al.* (2020) for further discussion.

In the general multiple-systems estimation model, suppose that there is a pair of lists, without

loss of generality lists 1 and 2, which contain no case in common. Then $N_A = 0$ for all combinations A of lists containing both 1 and 2. To find the posterior distribution of β_{12}, for each combination B of lists, define

$$C_B = \exp\left(\mu + \sum_{i \in B} \alpha_i + \sum_{\substack{i,j \in B, i<j \\ (i,j) \neq (1,2)}} \beta_{ij}\right).$$

It follows that

$$N_B \sim \begin{cases} \text{Poiss}(C_B) & \text{if} \{1,2\} \not\subseteq B, \\ \text{Poiss}\{C_B \exp(\beta_{12})\} & \text{if} \{1,2\} \subseteq B. \end{cases}$$

Because no cases are observed in the overlap of lists 1 and 2, we shall have $N_B = 0$ for all $B \supseteq \{1,2\}$. So the conditional likelihood of β_{12} given all the other parameters satisfies

$$\log\{L(\beta_{12} | \text{ no cases in common between 1 and 2, all other parameters})\}$$
$$= - \sum_{B \supseteq \{1,2\}} C_B \exp(\beta_{12}) = -C \exp(\beta_{12}), \tag{2}$$

where $C > 0$ depends only on the parameters other than β_{ij}. Whatever the value of C, the log-likelihood (2) is maximized as $\beta_{12} \to -\infty$.

The posterior density of β_{12} is proportional to $\exp\{-C \exp(\beta)\}$. Although this appears at first sight to be an improper distribution, this function has the properties that, for all y,

$$\int_{-\infty}^{y} \exp\{-C \exp(\beta)\} \, \mathrm{d}\beta = \infty$$

and

$$\int_{y}^{\infty} \exp\{-C \exp(\beta) \, \mathrm{d}\beta < \infty$$

so $P(\beta_{12} > y)/P(\beta_{12} \leqslant y) = 0$. This corresponds to the distribution where $\beta_{12} = -\infty$ with probability 1. Since this is true conditionally on all the other parameters whatever their values, the unconditional posterior distribution is the same. Hence the posterior distribution of the Poisson parameter for every B that includes lists 1 and 2 is an atom of probability at 0. Given the value $-\infty$ for β_{12}, the distribution of every N_B for each $B \supseteq \{1,2\}$ is then Poisson with parameter 0, in other words the constant value 0, regardless of the other parameters, whereas, for all other B, $N_B \sim \text{Poiss}(\lambda_B)$ with λ_B defined as in equation (1) above. So, as asserted above, the likelihood of all the other parameters conditionally on $\beta_{12} = -\infty$ is then obtained by simply omitting all combinations of lists which contain 1 and 2.

Returning to the UK data example, where there are six lists and hence 63 observable combinations B, we omit the 16 N_B for which B included both LA and GP lists, leaving 47 observations from which to estimate the remaining 21 parameters. In fact there is a second pair of lists for which there is no overlap at all, namely LA and NCA, and by the same argument the parameter $\beta_{\text{LA,NCA}}$ is also estimated to be $-\infty$ with probability 1. Removing (from the 47 remaining combinations) all combinations of lists containing both LA and NCA lists leaves 39 observations from which to apply the MCMC approach to the remaining 20 parameters. Within the package `modslavmse`, the routine `removeemptyoverlaps`, which is called from `MCMCfit`, produces the relevant data matrix and also a list of those interaction parameters that take the value $-\infty$ in the posterior.

6.3. Results

In this section, the results for the three examples are presented, exploring the effects of using various priors and various thresholds.

The results for the UK data are given in Tables 9–11. The first row in each table shows the result of fitting the main effects only, with no β_{ij} considered. Once interactions have been considered, the results are not enormously sensitive to the prior, especially if a non-zero threshold is used. If there is no thresholding, so that all interactions are included within the model, then the posterior credible intervals are much larger, but the central estimate is similar.

Computationally, the uniform improper prior is the fastest, though the possibly more plausible prior with variance 1 gives much the same results. A model with every possible interaction is more complicated than the amount of data can bear, and the thresholding at a threshold of 2 is a liberal approach which nevertheless eliminates extraneous complication. This would tend to suggest a point estimate of about 12300 for the overall prevalence, with an 80% credible interval (rounding to the nearest 500) of about 11500–13500 and a 95% credible interval, in round terms, of 11000–14000. This is about 1000 more than the confidence interval that is obtained from the fixed model, but this is possibly because the model averaging takes some note of the second group of models exemplified by the model chosen by the BIC. Increasing the threshold to 5 makes little difference for the full data, and, as we see below, homes in on just a single interaction among the lists.

Now turn to the Netherlands data, the results for which are shown in Table 12. Again, and not surprisingly, if all interactions are considered in the model then the posterior intervals are much wider. However, if the thresholding procedure is used to restrict attention to a smaller number of interactions, then the width of the intervals is not dramatically different from the main effects model. Threshold 2 with variance 1 appears to be an exception; however, examination of the results shows that only five of the 15 two-factor interactions are thresholded out. As a check, the method was run on the five-list version of the data, with the two smallest lists consolidated. The results for the five-list data were, in general, slightly lower for thresholds 0 and 2 and slightly

Table 9. Quantiles of the posterior distribution of the total population size, including both the observed data and the dark figure, UK data with six lists

Prior	*Threshold*	*Quantiles of posterior*				
		2.5%	*10%*	*50%*	*90%*	*97.5%*
Main effects only		11.0	11.4	12.2	13.1	13.6
Uniform	0	6.2	7.5	10.9	15.3	18.4
Variance 10	0	7.0	8.5	10.3	15.4	20.4
Variance 1	0	8.5	9.3	13.0	15.5	17.1
Variance 0.1	0	10.2	10.3	11.9	13.5	14.3
Uniform	2	10.7	11.2	12.2	13.6	14.4
Variance 10	2	10.9	11.4	12.3	13.5	14.3
Variance 1	2	10.9	11.3	12.3	13.6	14.4
Variance 0.1	2	10.8	11.2	12.0	13.0	13.3
Uniform	5	10.9	11.2	12.1	13.0	13.6
Variance 10	5	11.1	11.5	12.3	13.1	13.8
Variance 1	5	11.4	11.9	12.7	13.8	14.2
Variance 0.1	5	11.1	11.5	12.3	13.1	13.8

Table 10. Quantiles of the posterior distribution of the total population size, including both the observed data and the dark figure, UK data with five lists

Prior	*Threshold*	*Quantiles of posterior*				
		2.5%	*10%*	*50%*	*90%*	*97.5%*
Main effects only		12.0	12.5	13.5	14.6	15.3
Uniform	0	5.9	7.7	11.1	18.4	22.2
Variance 10	0	6.2	7.8	13.0	19.7	24.5
Variance 1	0	9.4	10.8	13.8	19.1	23.0
Variance 0.1	0	11.4	12.3	14.5	17.2	18.5
Uniform	2	10.7	11.1	12.2	13.3	13.9
Variance 10	2	11.6	12.0	13.1	14.1	14.7
Variance 1	2	11.7	12.1	13.2	14.3	15.1
Variance 0.1	2	12.2	12.8	14.1	15.4	16.2
Uniform	5	12.0	12.4	13.3	14.4	15.1
Variance 10	5	12.0	12.5	13.5	14.6	15.3
Variance 1	5	12.6	13.1	14.1	15.3	16.0
Variance 0.1	5	12.0	12.5	13.5	14.6	15.3

Table 11. Quantiles of the posterior distribution of the total population size, including both the observed data and the dark figure, UK data with four lists (five-list data with GP omitted)

Prior	*Threshold*	*Quantiles of posterior*				
		2.5%	*10%*	*50%*	*90%*	*97.5%*
Main effects only		9.5	9.9	10.7	11.6	12.1
Uniform	0	6.4	8.1	12.1	18.5	23.2
Variance 10	0	6.1	7.7	11.3	16.7	21.5
Variance 1	0	6.7	7.6	10.1	14.3	17.8
Variance 0.1	0	7.5	8.2	9.6	11.4	12.6
Uniform	2	10.5	11.0	12.0	13.1	13.7
Variance 10	2	10.6	11.0	12.0	13.0	13.7
Variance 1	2	10.4	10.9	11.8	12.9	13.7
Variance 0.1	2	9.3	9.7	10.6	11.5	11.9
Uniform	5	9.5	9.9	10.7	11.6	12.1
Variance 10	5	9.5	9.9	10.7	11.6	12.1
Variance 1	5	9.5	9.9	10.7	11.6	12.1
Variance 0.1	5	9.5	9.9	10.7	11.6	12.1

higher for threshold 5. The only substantially different case was variance 1, threshold 2, where the five-list data results are about 70% of the result for the six-list data.

The New Orleans data are a smaller set of observations and also consist of eight lists with none of the overlap sets containing more than two cases. Therefore it does not seem appropriate to use more than the main effects model and that approach was adopted in the original analysis (Bales *et al.*, 2019). However, it is of interest to see what would happen if we use the Bayesian approach allowing for interactions. Some trials suggest that, if the full eight-list data are used, the MCMC algorithm requires both a long burn-in period and then a long run, and possibly other

Table 12. Quantiles of the posterior distribution of the total population size, including both the observed data and the dark figure, Netherlands data

Prior	*Threshold*	*Quantiles of posterior*				
		2.5%	*10%*	*50%*	*90%*	*97.5%*
Main effects only		48.7	49.9	52.6	55.9	57.9
Uniform	0	31.0	36.0	50.9	65.3	72.4
Variance 10	0	35.7	36.6	52.0	73.1	83.5
Variance 1	0	46.5	52.0	69.3	74.5	81.5
Variance 0.1	0	49.1	53.6	60.9	67.7	71.5
Uniform	2	42.4	44.4	47.6	52.2	54.7
Variance 10	2	43.5	44.2	47.3	52.2	53.5
Variance 1	2	60.6	64.0	73.0	85.6	93.0
Variance 0.1	2	56.9	59.2	66.2	74.1	78.7
Uniform	5	51.3	52.9	56.1	59.1	60.9
Variance 10	5	54.7	56.0	59.5	63.2	65.4
Variance 1	5	54.6	56.2	59.4	62.9	65.2
Variance 0.1	5	61.0	63.4	68.2	73.3	75.9

Table 13. Quantiles of the posterior distribution of the total population size, including both the observed data and the dark figure, New Orleans data consolidated into five lists

Prior	*Threshold*	*Quantiles of posterior*				
		2.5%	*10%*	*50%*	*90%*	*97.5%*
Main effects only		0.7	0.8	1.1	1.5	1.9
Uniform	0	0.4	0.8	4.1	17.4	38.7
Variance 10	0	0.6	1.2	2.8	10.5	24.2
Variance 1	0	0.6	0.8	1.3	2.4	2.9
Variance 0.1	0	0.6	0.8	1.0	1.5	1.8
Uniform	2	0.6	0.6	0.8	1.2	1.3
Variance 10	2	0.8	0.9	1.2	1.8	2.3
Variance 1	2	0.7	0.8	1.1	1.5	1.9
Variance 0.1	2	0.7	0.8	1.1	1.5	1.9
Uniform	5	0.6	0.6	0.8	1.2	1.3
Variance 10	5	0.7	0.8	1.1	1.5	1.9
Variance 1	5	0.7	0.8	1.1	1.5	1.9
Variance 0.1	5	0.7	0.8	1.1	1.5	1.9

adjustments to the control parameters, to give reasonable mixing in the posterior realizations. For simplicity, therefore, we analyse the five-list version, and the results are given in Table 13. The variance 1, threshold 2, model (and indeed some of the other models) gives results that are identical to the main-effects-only model, and closer examination of the estimates within the package shows that the thresholding step in fact removes all the interactions, leaving main effects only.

However, the uniform prior, even with strong thresholding, gives different estimates. To understand why, note that there are 10 two-factor interactions β_{ij} between the five lists. In three of these cases, the observed overlap between lists i and j is zero, and so the corresponding β_{ij} is estimated as $-\infty$ regardless of the thresholding. Even with a moderate threshold all the other

Table 14. Quantiles of the posterior distribution of the total population size, including both the observed data and the dark figure, Kosovo data

Prior	*Threshold*	*Quantiles of posterior*				
		2.5%	*10%*	*50%*	*90%*	*97.5%*
Main effects only		7.2	7.2	7.4	7.6	7.7
Uniform	0	12.5	13.1	14.4	15.9	16.7
Variance 10	0	12.5	13.1	14.3	15.9	16.8
Variance 1	0	12.3	12.9	14.1	15.6	16.4
Variance 0.1	0	10.7	11.1	12.1	13.2	13.8
Uniform	2	12.6	13.1	14.3	15.7	16.3
Variance 10	2	12.6	13.1	14.2	15.5	16.4
Variance 1	2	12.3	12.9	14.0	15.2	16.1
Variance 0.1	2	10.9	11.2	12.1	13.1	13.6
Uniform	5	12.6	13.1	14.3	15.7	16.3
Variance 10	5	12.6	13.1	14.2	15.5	16.4
Variance 1	5	12.3	12.9	14.0	15.2	16.1
Variance 0.1	5	10.9	11.2	12.1	13.1	13.6

interactions are thresholded out, but the model is fitted not just on the basis of the main effects only but with three of the interactions included and estimated to $-\infty$. If the original eight-list data are considered, then the effect is much stronger, with 18 of the 28 possible interaction parameters estimated as $-\infty$.

The Kosovo data are unusual in that all the models allowing for interactions give broadly similar results; Table 14. The thresholding has little or no effect, even at a threshold of 5, because most of the interactions are very strong.

6.4. Choosing the threshold for interactions

The Bayesian approach avoids the necessity of choosing a particular model, but it still contains tuneable prior parameters. The implausibility of very large positive or negative values for the interaction parameters suggests that a prior variance of 1 is a reasonable choice. The standard MCMC software does not allow for the mixed model with an atom of probability at zero for the parameters (a topic for future research), but the thresholding approach gives a simple alternative.

In Table 15 we see the interactions that exceed the threshold at the first stage for both thresholds considered. The three results for the UK data are entirely consistent with one another, given that the second data set is obtained by combining the PF and NCA lists and the third by omitting the GP list. For the Netherlands data, 10 of the possible 15 interactions survive a threshold of 2, and the results that were obtained are somewhat anomalous, both when compared with those for other parameter values and when compared with the effect of combining the two smallest lists. For threshold 5, the method picks out the LA:NG interaction only for the UK data and the O:Z interaction for the Netherlands data, with the same results in both cases if the two smallest lists are consolidated. Leaving aside prevalence estimation as such, an advantage of the more parsimonious approach is that it focuses in on those pairs where there is a very clear interaction, giving pointers to where to look particularly to gain a greater understanding of what is going on. However, it is intuitively clear in the modern slavery case that correlations between lists are not at all surprising, and the results demonstrated in Table 15 suggest that the less restrictive threshold 2 is probably to be preferred at least as a starting point. Interestingly, reducing the threshold in the Netherlands data to 4.5 yields a similar result to threshold 2.

Table 15. Interactions included in the variance 1, threshold 2, model†

Data set	*Lists*	*Interactions included*
UK	6	*LA:NG*, LA:PF, NG:GP, PF:GP, PF:NCA, GO:GP
UK	5	*LA:NG*, LA:PFNCA, NG:GP, PFNCA:GP
UK excluding GP	4	LA:NG, LA:PFNCA
Netherlands	6	I:K, I:Z, K:O, K:P, K:R, K:Z, O:P, *O:Z*, P:R, P:Z
New Orleans	8	No interactions at either threshold
Kosovo	4	*All except ABA:HRW*

†For threshold 5, only the effects shown in italics survive the thresholding step. For the four-list UK data, the effect LA:NG does survive up to thresholds of about 3.5.

For the New Orleans data, any reasonable level of thresholding leads back to the fitting of main effects only, which is probably the most realistic model given the number of lists and the numbers of cases in the various overlaps. In contrast, for the Kosovo data, only one of the interaction effects is thresholded out, even at the high threshold. This is not surprising since the data clearly demonstrate strong interlist correlations, and it is very reassuring that even the high threshold adapts well to data of this kind.

Overall, consideration of these examples suggests that the Bayesian–threshold model with variance 1 and threshold 2 adapts reasonably well to the characteristics of different data sets, although it is advisable not to apply the method completely blindly.

7. Conclusions

Estimating and keeping track of the numbers of victims is a crucial component of the fight against modern slavery. If multiple-systems estimation is to be used as one of the standard methods, then the stability and robustness of point and interval estimation are an important consideration. The most stable method would, of course, be to ignore the possibility of interactions and simply to fit main effects, but the Kosovo example shows that this would clearly be inadequate in some practical cases. It would also fail to take account of the correlations which are not unexpected between the lists that are obtained in the modern slavery context.

The Bayesian approach of this paper, with a threshold of 2 and a prior variance of 1 for the interaction parameters, is at least a candidate. On the data sets considered, it gives results which are stable and robust when smaller lists are combined, and it automatically rules out implausible secondary estimates which are almost certainly spurious. If it is desirable to obtain parsimonious explanations in cases where there may be interactions of particular interest, then the threshold can if necessary be increased. The approach adapts well between data such as the Kosovo data, with strong dependences between lists, and those situations where few, if any, interactions are clearly present in the data.

One contrast between the modern slavery data sets and the Kosovo data is that the modern slavery data are much sparser, in that not every combination of lists is observed at all. This is not a reflection of the quality of the data but is intrinsic to the field. In modern slavery, we shall often wish to quantify the number of victims in a fairly constrained geographical area over a reasonably short time period, and so the total population size may be quite small, as in the Greater New Orleans example. Even when we consider larger data sets, such as the Netherlands data, the number of cases that are actually observed may be only a relatively small proportion of the total population. Sparse data, and lists that do not overlap at all, are the norm rather than

the exception, and methods need to take account of that. Of course, it is to be hoped that, as public and political consciousness about modern slavery increases, a larger proportion of cases will actually come to light, but this is likely to be a long process. Further recent work on this aspect by the author and colleagues is reported in Chan *et al.* (2020). It should also be noted that when multiple-systems estimation is used for a census of an animal or an easy-to-count human population, attempts can be made to design the surveys or captures to be independent of one another and also to be sufficiently large to avoid the sparsity issues that are raised by the modern slavery data sets; however, in most human rights contexts, there is no such control over the way that lists arise.

The availability of real data has been an important contribution to the study that was carried out in this paper, because data on modern slavery and human trafficking will have specific characteristics which need to be taken into account. It is to be hoped that more data sets will be put into the public domain, of course in formats that preserve the privacy of individuals and do not hamper the primary task of rescuing and supporting victims, bringing perpetrators to justice, and discouraging modern slavery in the future.

Multiple-systems estimation is not a panacea, but part of the quest for better information and understanding. A key topic for discussion and for future research is how we can build on a whole range of information and methods to gain a deeper understanding of modern slavery. For example, a promising development is the typology that was developed by Cooper *et al.* (2017) and the associated case file coding template. More widely, the important role of research in fighting modern slavery is underlined by the research priorities that are set out in Her Majesty's Government (2018). A broad discussion of the actual and potential modes of measurement, and how these fit into the legal, definitional and historical background of modern slavery, is given by Landman (2020).

There are several avenues for future research on the multiple-systems methodology that is set out in this paper. For example, how can it be developed to handle concomitant information and segmentation of populations? What is the best approach when the aim is to discern whether the overall level is different between two time points or between two different sectors or geographical areas? Can the approach be easily extended to the case of fuzzy matching, where it is not quite clear whether cases on different lists are or are not the same? Perhaps most importantly, are there particular patterns in data sets drawn in the context of modern slavery and human trafficking, and can these, as well as the prevalence estimates themselves, contribute to a deeper understanding of the problem itself?

Acknowledgements

The author gratefully acknowledges correspondence and other help from Kevin Bales, Patrick Ball, Peter van der Heijden, Ella Kaye and Daniel Manrique-Vallier, and the very helpful comments of the referees. This work was supported by the Arts and Humanities Research Council and the Economic and Social Research Council grant ES/P001491/1, 'Modern slavery: meaning and measurement (PaCCS Transnational Organised Crime, University of Nottingham, 2016–18').

References

Baillargeon, S. and Rivest, L.-P. (2007) Rcapture: loglinear models for capture-recapture in R. *J. Statist. Softwr.*, **19**, no. 5, 1–31.

Bales, K. B., Hesketh, O. and Silverman, B. W. (2015) Modern slavery in the UK: how many victims? *Significance*, **12**, no. 3, 16–21.

Bales, K., Murphy, L. and Silverman, B. W. (2019) How many trafficked people are there in Greater New Orleans?: Lessons in measurement. *J. Hum. Traffckng*, to be published.
Ball, P., Betts, W., Scheuren, F., Dudukovich, J. and Asher, J. (2002) Killings and refugee flow in Kosovo March–June 1999. *Report*. American Association for the Advancement of Science, Washington DC.
Bird, S. M. and King, R. (2018) Multiple systems estimation (or capture-recapture estimation) to inform public policy. *A. Rev. Statist. Appl.*, **5**, 95–118.
Chan, L., Silverman, B. W. and Vincent, K. (2020) Multiple systems estimation for sparse capture data: inferential challenges when there are non-overlapping lists. *J. Am. Statist. Ass.*, to be published, doi 10.1080.01 621459.2019.1708748.
Cockayne, J. (2015) Unshackling development: why we need a global partnership to end modern slavery. Freedom Fund, London.
Cooper, C., Hesketh, O., Ellis, N. and Fair, A. (2017) A typology of modern slavery offences in the UK. *Research Report 93*. Home Office, London.
Cormack, R. M. (1989) Log-linear models for capture-recapture. *Biometrics*, **45**, 395–413.
Cormack, R. M. (1992) Interval estimation for mark-recapture studies of closed populations. *Biometrics*, **48**, 567–576.
Cruyff, M., van Dijk, J. and van der Heijden, P. G. M. (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.
van Dijk, J. J., Cruyff, M., van der Heijden, P. G. M. and Kragten-Heerdink, S. L. J. (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, Vienna.
Fienberg, S. E. and Rinaldo, A. (2012a) Maximum likelihood estimation in log-linear models. *Ann. Statist.*, **40**, 996–1023.
Fienberg, S. E. and Rinaldo, A. (2012b) Maximum likelihood estimation in log-linear models: supplementary material. *Technical Report*. Carnegie Mellon University, Pittsburgh.
Her Majesty's Government (2018) *2018 UK Annual Report on Modern Slavery*. London: Stationery Office. (Available from `data.parliament.uk/DepositedPapers/Files/DEP2018-1042/UK_Annual_Report_on_Modern_Slavery_2018.pdf`.)
Johndrow, J., Lum, K. and Ball, P. (2015) dga: Capture-recapture estimation using Bayesian model averaging. *R Package Version 1.2*.
Johnstone, I. M. and Silverman, B. W. (2004) Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences. *Ann. Statist.*, **32**, 1594–1649.
King, R., Bird, S. M., Overstall, A. M., Hay, G. and Hutchinson, S. J. (2013) Injecting drug users in Scotland, 2006: number, demography, and opiate-related death-rates. *Addictn Res. Theory*, **21**, 235–246.
Landman, T. (2020) Measuring modern slavery: law, human rights and new forms of data. *Hum. Rights Q.*, **42**, no. 2, in the press.
Madigan, D. and York, J. C. (1997) Bayesian methods for estimation of the size of a closed population. *Biometrika*, **84**, 19–31.
Manrique-Vallier, D. (2016) Bayesian population size estimation using Dirichlet process mixtures. *Biometrics*, **72**, 1246–1254.
Manrique-Vallier, D. (2017) LCMCR: Bayesian non-parametric latent-class capture-recapture. *R Package Version 0.4.3*.
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. *Preprint arXiv:1906.04763*.
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 Nonmilitary Deaths in Conflict* (eds T. Seybolt, B. Fischhoff and J. Aronson), pp. 77–93. New York: Oxford University Press.
Martin, A. D., Quinn, K. M. and Park, J. H. (2011) MCMCpack: Markov chain Monte Carlo in R. *J. Statist. Softwr.*, **42**, no. 9, 1–21.
Silverman, B. W. (2014) Modern slavery: an application of multiple systems estimation. Home Office, London. (Available from `https://www.gov.uk/government/publications/modern-slavery-an-application-of-multiple-systems-estimation`.)
Silverman, B. W. (2018a) modslavmse: multiple systems estimates for estimating the prevalence of modern slavery. *R Package*. University of Nottingham, Nottingham. (Available from `https://github.com/bernardsilverman/modslavmse`.)
Silverman, B. W. (2018b) Demonstrating risks is not the same as estimating prevalence. In *Proc. Delta 8.7 Modelling the Risk of Modern Slavery Symp.* New York: United Nations University Center for Policy Research. (Available from `https://delta87.org/2018/12/demonstrating-risk-not-same-estimating-prevalence/`.)

Discussion on the paper by Silverman

Dankmar Böhning (*University of Southampton*)

As the proposer of the vote of thanks I congratulate Professor Silverman for an excellent paper on the 'dark number' involved in modern slavery. The topic of the paper is of vital importance and I am delighted to see that the issue is addressed in such a prominent place. It is of the utmost difficulty to estimate the magnitude of modern slavery as there is only a certain amount visible but, because of the very nature of the subject, there is also much hidden activity. This is very similar to other sensitive subjects such as family violence, possession of firearms or drink-driving, all having illegal aspects involved. In the case that we have information from different sources or lists we can construct an estimate of the hidden number of units involved in the subject of interest. This is the topic of multiple-systems estimation. Given that we have k lists, we can construct a k-dimensional frequency table with 2^k frequencies where each cell corresponds to the frequency of units that have a certain combination of lists. The cell, corresponding to the combination of not being present on any of the lists, is empty and is the interest of multiple-system estimation. Professor Silverman provides four very detailed examples of such settings, which are one of the numerous strong points of the paper. The paper continues by highlighting various problems using log-linear models in modelling the frequency table including that similarly fitting models often show a wide range of estimates for the hidden number which occurs in the frequentist and also in the Bayesian approaches that have been used so far. The paper then suggests a thresholding Bayesian approach and it is indicated that this works very well, potentially, with future work investigating it in more detail.

In what follows I would like to comment on what has been called *internal validation*. The technique is mentioned and explained in Hook and Regal (1995). Great concern exists about the fact that seemingly similar fitting models can lead to rather different estimates of the hidden number. We shall explain the process at hand for one of the examples given in the paper by Professor Silverman, namely the casualties in the Kosovo war. I shall use the same notation as in the paper. Evidently, the model fitting problem would

Table 16. Models considered in the conditional analysis of the Kosovo data for the subtable of those identified by EXH

Model	*Effects involved*
1	ABA OSCE HRW
2	ABA OSCE HRW ABA∗OSCE
3	ABA OSCE HRW ABA∗HRW
4	ABA OSCE HRW OSCE∗HRW
5	ABA OSCE HRW ABA∗OSCE ABA∗HRW
6	ABA OSCE HRW ABA∗OSCE OSCE∗HRW
7	ABA OSCE HRW ABA∗HRW OSCE∗HRW
8	ABA OSCE HRW ABA∗OSCE ABA∗HRW OSCE∗HRW

Table 17. Goodness-of-fit analysis for the subtable in the Kosovo data created by those identified by list EXH

Model	*Goodness of fit*	*p-value*
1	192.10	0
2	13.14	0.011
3	191.86	0
4	177.06	0
5	12.85	0.005
6	0.79	0.853
7	177.06	0
8	0.75	0.688

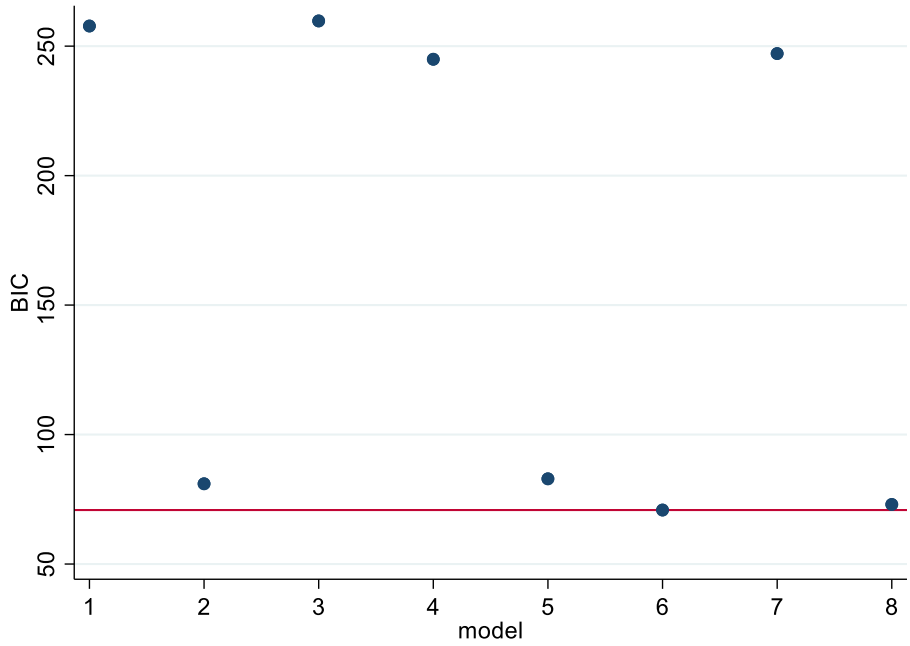

Fig. 7. BIC-values for the eight models using all eight frequencies

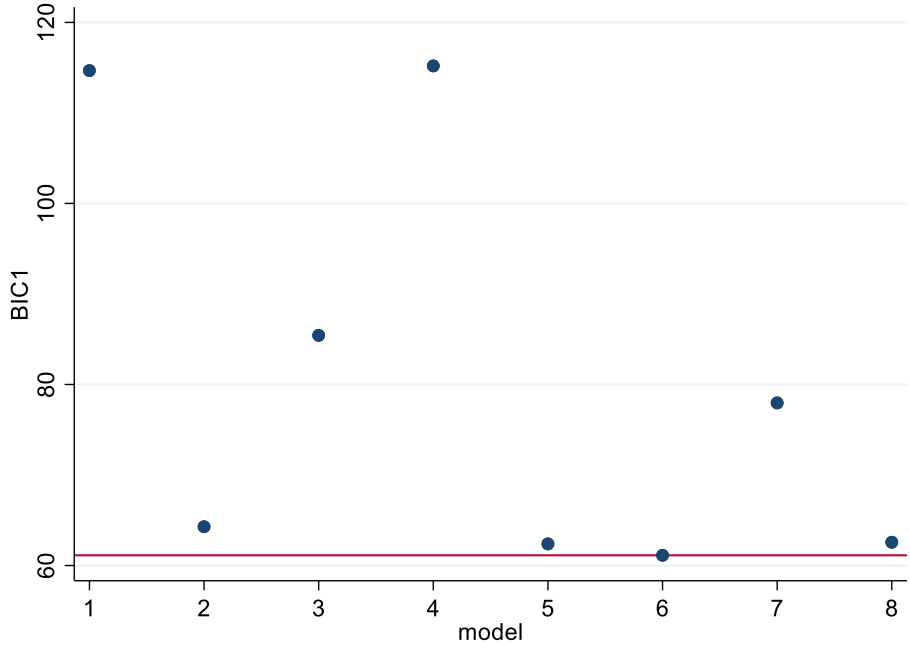

Fig. 8. BIC-values for the eight models using seven frequencies: the frequency of cell (0,0,0) is set to be missing

be much simpler if the missing frequency is available. This is, of course, not so as it is the objective of the exercise. However, one can do the following construction. We take the first list, called EXH (exhumations), say, and condition on those who are present in this list. For this subtable we have a frequency for those not present in any of the other three lists. Now, we can do the modelling for this table with all eight entries available. We are now looking at all main effects of ABA (American Bar Association), OSCE (Organization for Security and Cooperation in Europe) and HRW (Human Rights Watch) as well as any combinations of two-way interactions; the models are given in detail in Table 16.

The Bayesian information criterion values BIC for the eight models are presented in Fig. 7 with the clear winner being model 6 which has an even smaller BIC-value than model 8. In addition, it is the *only* model (except model 8) which has acceptable goodness of fit although there are several with similar BIC-values (Table 17).

In the next step, we set the (0,0,0) cell frequency (which here is 1131) to missing and do the modelling exercise again. The BIC-values against model are shown in Fig. 8. Again, model 6 is the winner but note that the rankings of models according to BIC have shifted. In addition, besides model 6, also model 2 and model 5 gain acceptable goodness of fit. The estimate of the missing cell frequency, using model 6, is 1042 which compares favourably with the observed 'missing' frequency 1131 whereas models 2 and 5 produce far-off estimates of 575 and 760 respectively.

In practice, we must choose between models 2, 5 and 6 with no additional information and it is then easily possible not to pick the right model (if there is one). So, internal validation, using full table information, can help to identify suitable models, in particular, if only lower order interactions are of interest. I very much hope that these comments can supplement the outstanding paper by Professor Silverman to which I feel honoured to propose the vote of thanks.

John Whitehead (*Lancaster University*)

It is a great pleasure to second the vote of thanks for this paper. You might wonder: why me? I am a clinical trials statistician—retired. But a couple of years ago, seeking to widen my academic horizons, I enrolled on a Master's course on international slavery studies at the University of Liverpool. There, during the first term in a module on modern slavery, I first came across the problem of estimating the number of victims of such exploitation and the paper of Bales *et al.* (2015) which introduced the multiple-systems estimation approach of which we have heard tonight. This appeared to be a welcome bridge between the linear modelling world of my former discipline and the problems that are inherent in my new one.

Tonight, Professor Silverman has expanded on that earlier work, considering Bayesian as well as frequentist approaches to analysis, and providing several further examples. Nonetheless, the essential features of Bales *et al.* (2015) remain. In particular, to fit the model defined by equation (1) of his paper, he makes two assumptions:

(a) two-list interaction terms may sometimes be 0;
(b) three-list interaction terms are always 0.

To determine which two-list interactions to include in a model, forward selection is used. Only interactions for which there is significant evidence are included. Such tests have very little power. Indeed, if the two lists in question have no overlap, then there is no power at all, although in that circumstance the corresponding interaction term is set to 0.

This was the first feature of the method described tonight to cause me concern. It reminded me of how crossover trials, in which patients first receive one treatment and then the other in random order, used to be analysed. First, one tested for carry-over, by which the first treatment administered affected the response following the second. If there was no significant evidence of carry-over, then it was excluded from the model, and all patient responses were used to compare the treatments. Otherwise, only data from the first period were used in a much weaker analysis. However, Freeman (1989) clearly demonstrated how this two-stage procedure inflated type I error rates and led to apparent 95% confidence intervals achieving coverage as low as 56%. Following Freeman (1989), the two-stage analysis procedure disappeared from practice and from courses.

Whitehead *et al.* (2019) investigated the five-list UK modern slavery data and reproduced the results in italics in Silverman's Table 5: estimated total (in thousands) 11.3 with 95% confidence interval (9.9, 13.1). Fitting all two-list interactions gave an estimate (in thousands) of 11.2 with confidence interval (6.2, 20.8). Perhaps the accuracy suggested by the former interval is illusory. We then simulated 10000 replicate data sets from a fitted model including all two-list and no three-list interactions (one 0-count was replaced by $\frac{1}{2}$). The true Poisson mean for the 'dark figure' was 8.4. Fitting all two-list interactions to each data set gave

an average estimate of 8.8, and 94.4% of confidence intervals included 8.4. Applying Silverman's forward selection procedure gave an average estimate of 12.2, and only 39.1% of confidence intervals included the true value. The former result demonstrates the soundness of the basic fitting procedure, whereas the latter shows that pretesting is as inappropriate in this context as it is in crossover trials.

Assumption (b) is neither mentioned nor justified in Silverman's paper. I have not been able to express this assumption in an intuitive way that might be understood and therefore assessed by investigators. We simulated 10000 data sets from a fitted model including all two-list and three-list interactions (some 0-counts were replaced by 0.1). The true dark figure (in thousands) was 0.6. Fitting all two-list and no three-list interactions to each data set gave an average estimate of 9.7, and 2.0% of confidence intervals included 0.6. Applying Silverman's forward selection procedure gave an average estimate of 12.8, and none of the confidence intervals included the true value. Simulations of models incorporating three-list interactions were totally unsatisfactory.

The Bayesian–threshold method described in Section 6 of the paper uses priors incorporating assumptions (a) and (b). Do they reflect investigators' beliefs? Are they even understood by investigators? If not, the posterior inferences are not relevant to investigators.

The multiple-systems estimation approach as described tonight appears to be unreliable. Indeed, I suggest that seeking the dark figure may be futile. Changes in populations over time and place could perhaps be better tracked by using the number of observed individuals. Such an approach has shortcomings, but these are at least transparent.

The problems that are addressed in the paper are of the utmost importance. The methods proposed to address them make ingenious use of linear modelling, although I feel that more work needs to be done to assess their robustness and acceptability to investigators. It gives me great pleasure to second the vote of thanks.

The vote of thanks was passed by acclamation.

Ian Diamond (*Office for National Statistics, London*)

It gives me great pleasure to comment on this excellent paper. It is precisely the sort of work that is needed to give government the evidence on which to base policy in areas where collecting full information is incredibly difficult; and I believe we should be looking at other areas to apply this methodology. As an example let us consider the census. Although censuses collect information on the whole population it can be difficult to enumerate some subpopulations. If we are to be inclusive in our data collection then the use of a multiple-systems approach could be very helpful, e.g. to gain an estimate of the number who are homeless.

Turning to the analysis one of the key questions is, given the various models, which one should be chosen? Sir Bernard makes a compelling argument for his approach after leaving out general practitioner (GP) lists. However, it may be useful to reflect on why the GP lists bring instability in the estimates. It is well known that there are situations in which GP lists can overcount the population and, if this were true here, then it could lead to unstable estimates. Although this is speculation it does bring me to the important point that it is assumed that people on a list exist. If there is overcount then unstable estimates and, I would suggest, overestimates will occur. It is essential therefore that everything is done to clean any lists used.

Thank you again, Sir Bernard, for an excellent paper.

Sheila M. Bird (*Medical Research Council Biostatistics Unit, Cambridge*)

Tonight's paper offers new software, different from `conting` (Overstall and King, 2014), for use when applying Bayesian multiple-systems estimation (MSE) (Bird and King, 2018) to quantify modern slavery. Professor Silverman highlights smallest list sizes and combinations thereof; and the need to identify and deal with empty (or unobserved) overlaps between lists when attempting to fit first-order and higher order interactions. In the context of modern slavery Silverman's prior belief is for few interactions (and hence his choice of prior variance); he also proposes a threshold for credible interactions (2 or 5; I might interpose 3). A by-product of thresholding is its seeming ability to deal with one of the main drawbacks of MSE estimation: the widths of posterior credible intervals which, for example, make trends difficult to discern. Professor Whitehead has noted that the coverage of thresholded intervals needs to be assessed.

Attractive as some of the above features are, I note some practical reservations. First, even readily identifiable covariates, which may determine list propensities, are absent from most data sets on modern slavery: adult or child; gender; type of slavery (domestic; sex work; other labour). Secondly, clusters of victims may be rescued together and MSE needs to evolve methods to address this type of clustering

(Bird, 2019). Third, list combination should consider attributes other than list size; for example, some non-governmental organizations are set up to help child victims; others, adult females. Fourth, when assessing the UK's extent of modern slavery, it matters whether the analysis concerns 'potential victims' who have passed the UK's initial 'reasonable grounds' test or 'conclusive decision victims'. As only 55% of the UK's potential victims receive a positive decision (eventually), the UK's MSE analyses to date have related to 'potential victims'. Fifth, Professor Silverman alluded to the difficulty of matching across lists. In consultation with victims, and sensitively, we should consider whether there is a role for DNA matching to reunite trafficked victims from the same family.

Finally, a public health perspective on the care for, and follow-up of, rescued victims (Bird, 2019) should address their (and others') morbidity and mortality—*en route* to the UK; during enslavement; and for at least 2 years after rescue—so that we learn how better to care.

Paul A. Smith (*University of Southampton*)

I found this paper a fascinating insight into the workings of model fitting in multiple-systems estimation. I want to make two points. First, in Fig. 5(a), I am interested in the models which give rise to the model curves with the highest likelihoods. So I examined the five models with the largest maxima in their likelihoods for the five-source case, all visible in Fig. 5(a). Four of these models have their maxima in the lower mode, and one in the upper mode (Table 18).

There is a noticeable separation in that all the models on the left-hand side of Table 18 include multiple interactions with the general public (GP) ($\equiv 5$) source, whereas the model on the right-hand side has only a single modifying interaction for this source. This suggests that relying on the main effects for GP leads to these large estimates, which fits with Professor Silverman's view that GP is an 'outlier' source. Perhaps this kind of approach can lead to an evaluation of the quality of the different sources. I can imagine less clear-cut cases where the decision on whether or not to include a source would benefit from some evidence of internal coherence, though an alternative explanation might be that a discrepant source is better because it provides new information.

This leads to my second point, which is about the quality of the input data, and more specifically the linking (distinguishing *links* derived by using a statistical process from *matches* which are the unobservable true homologies). Let me say at once that I understand the limitations of the sources and applaud the efforts to derive sensible estimates from sparse data. But, while taking great care of the model uncertainty, I find only a little comment in the paper on the uncertainty in the underlying data. Multiple-system estimation is sensitive to the linking, and, if there are missed or false links because the original sources contain rather little identifying information, this is likely to have a substantial effect on the resulting estimates. Is there any linkage metadata, or assessment of linkage quality, to go with the data sets that have helpfully been made available? Use of such information has been extended to multiple systems by Zult *et al.* (2019). I suspect (without doing the work) that adding or removing one link may have quite an effect on the chosen model with such sparse data. Perhaps the apparent discrepancy in GP as a source would be reduced in one of these scenarios, and, if indeed a coherence statistic can be constructed, we could examine what changes in linkage would most increase the coherence of the sources.

Table 18. Five models with the highest maxima in their likelihoods, in descending order of likelihood†

Lower mode	*Maximum*	*Upper mode*	*Maximum*
[125][135][45]	0.00279	[123][34][25]	0.00192
[125][135][245]	0.00072		
[12][25][35][45]	0.00059		
[125][35][45]	0.00051		

†1, LA; 2, NG; 3, PFNCA; 4, GO; 5, GP. Terms grouped together in square brackets indicate that the interaction between all the terms and all lower order interactions involving them are included in the model.

Helen Ross (*Office for National Statistics Centre for Crime and Justice, Newport*)
The Office for National Statistics Centre for Crime and Justice is responsible for publishing official statistics on crime in England and Wales. This includes quarterly statistics as well as compendium publications and analytical articles exploring specific themes and responding to emerging issues.

To contribute towards achieving targets 5.2 ('Eliminate all forms of violence against all women and girls in the public and private spheres, including trafficking and sexual and other types of exploitation'), 8.7 ('Take immediate and effective measures to eradicate forced labour, end modern slavery and human trafficking and secure the prohibition and elimination of the worst forms of child labour, including recruitment and use of child soldiers, and by 2025 end child labour in all its forms') and 16.2 ('End abuse, exploitation, trafficking and all forms of violence against and torture of children') of the United Nations's sustainable development goals, the Centre for Crime and Justice is currently exploring ways of collating and producing data that are relevant to the measurement of modern slavery in the UK.

The hidden nature of this relatively small-scale crime in the UK makes producing an accurate prevalence measure difficult. The Office for National Statistics does not intend to use methods such as multiple-systems estimation or survey sampling to estimate prevalence directly as neither will reflect the actual extent of modern slavery in the UK. Instead, the emphasis will be on measuring or quantifying indicators and factors that are known to be linked to modern slavery.

These indicators will include sources of data from a wide variety of organizations, such as government bodies, service providers and charities. For example, data relating to criminal proceedings of modern slavery cases in the UK will give an indication of how the criminal justice system is responding to modern slavery victims and perpetrators. Additionally, sources of data showing public awareness and willingness to report modern slavery will also be important.

Monitoring known factors will help the UK to measure its progress towards combating modern slavery. This approach may also be particularly relevant in other countries where the prevalence of modern slavery is low.

The Centre for Crime and Justice will publish a report in the spring of 2020 discussing the approach in detail and bringing together indicators to quantify modern slavery.

Sarah Henry (*Office for National Statistics, Titchfield*)
I thank Bernard for his important and interesting paper, which offers opportunities for further research not only into modern slavery but also other marginal groups in our society. The work has specific implications for the Office for National Statistics and the census which we shall take forward.

I am specifically interested in the data that sit behind the analysis and would like to see more information about it provided in the paper. For example, what time period is covered in the research and is there more information about the 'sector' in which the person reported is a victim? The latter is sparked by my instinct that the stable population size arrived at with the methods proposed still underestimates the scale of the issue. I wonder whether further research into the 'sectors' may offer an independent estimate to compare the results with. A quick Google search (which I accept will not be a valid method going forward!) shows that there are around 20000 car washes and 2000 nail bars in the UK. If we consider many of the other sectors in the economy that are prone to harbouring slavery, it is likely that the 'market' is sizable and it is possible that the number of slaves is far greater than 11000–13000, especially if this figure relates to a stock. My recommendation is that more research is done on independent estimates and I look forward to following developments in the field.

Jessica L. Decker Sparks (*University of Nottingham Rights Laboratory*)
Critiques of using multiple-systems estimation (MSE) to quantify modern slavery are seemingly divided into three categories: data collection, case linkage and the statistical techniques. It is plausible that data collection and case linkage will organically improve. Indeed, modern slavery scholarship is a comparatively new field, and a field that has historically relied on proxy indicators (which have their own distinct limitations). But with a growing field of scholars, and increasingly interdisciplinary scholars, a natural diversity of quantitative and qualitative data collection methods are emerging. As a result, moving forward, scholars will be better equipped

(a) to collect the primary data that are needed for MSE and/or
(b) to collaborate with key front-line partners to design more comprehensive data collection systems, that are still ethically responsive, that interface with case identification systems—improving the reliability of list construction and the validity of case linkages.

Remaining then are the critiques of the statistical techniques, which are well articulated in response to this paper. However, the response to these critiques should not be to abandon the approach in favour of using observable counts only. Scholars, practitioners, law enforcement and even government officials agree that observable counts are not reflective of prevalence. The objective of using MSE in many social science fields is to 'illuminate the hard to reach'. Relying only on observable counts resorts back to concealing the hard to reach and probably will perpetuate the field's continued understanding of limited demographics. Also, modern slavery is often just not observable. Although domestic servitude victims or survivors are purportedly the most hidden population, fishers experiencing conditions of modern slavery on vessels at sea are not even on land; thus there is not an opportunity to 'observe' these cases. Instead, there is a robust body of literature from a range of disciplines such as ecology, public health or epidemiology that all offer potential approaches for addressing violated assumptions. For example, the spatial distribution violations of schooling fish that have been corrected for in capture–recapture techniques may be applicable to concerns of clustering in lists of cases of modern slavery. The gravity of the preliminary identification of cases of modern slavery occurring in the UK has rightfully catalysed novel approaches to quantifying the problem through MSE—an approach that provides a useful starting point that can be further tailored as uncertainty is better understood and techniques are improved.

M. S. Ridout (*University of Kent, Canterbury*)
I have enjoyed reading this interesting and important paper. It focuses on application of the methods to real data, where the true answer is unknown, and instead considers robustness and stability of estimators when lists are omitted or merged. I wondered whether it might also be useful to condition on the subpopulation of individuals who appear on a particular list. If the methods are applied to the remaining lists by using this subset of data, the estimated 'dark number' can be compared directly with the known number of individuals who appear only on that list. Although this does not help with estimating the real dark number, it may be a useful indicator of the appropriateness of the methods.

Thomas A. Louis (*Johns Hopkins Bloomberg School of Public Health, Baltimore*)
Many thanks go to Sir Bernard for debriefing on his estimation of the extent of modern slavery. The paper highlights advantages conferred by statistical models in a context wherein machine learning approaches are unlikely to be successful. Silverman's work, and that of many others on this topic, adds to the ever expanding roles of statistics and statisticians in enhancing the public good.

The following points merit consideration.

The data model: the log-linear model on the seventh page is in the standard main effects + interaction format. It is broadly applicable, with Fienberg (1972) an early example; however, when combining evidence from more than two lists, a 'logic' parameterization is worth considering (see Ruczinski *et al.* (2003) and Schwender and Ickstadt (2008)). For the logic model, let $j = 0, 1, \ldots, N$ index potential victims, and consider five lists ($K = 5$). With $a_{ji} = 1$ or $a_{ji} = 0$ according to whether potential victim j is or is not on list i, let $\mathbf{a}_j = (a_{j1}, \ldots, a_{j5})$. There are 2^K candidate indicator regressors, $I\{\mathbf{a}_j\}$, and the 'dark figure' is $\mathbf{a}_0 = (00000)$. It is possible that omitting some of the $\mathbf{a}$-vectors with two 1s and, though the prevalence of three-list appearance is low, including a few with three 1s, all with thresholding or shrinkage, will increase stability and improve performance relative to the standard parameterization.

Thresholding: thresholding stabilizes and is attractive when simple structure is the primary goal. However, stabilization by standard shrinkage is probably more effective when making predictions, e.g. the value of $\lambda_\emptyset$ (or the slope on $\mathbf{a}_0$). Related, the use of a 'slice-and-slab' prior (Ročková and George, 2018) is an alternative to Silverman's z-score thresholding.

Fuzzy matching: fuzzy matching is worth considering, with Steorts *et al.* (2016) reporting on matching administrative records. As shown by Louis *et al.* (2011) in the genomics context, there are considerable benefits to retaining the fuzziness (operate with fractions) rather than selecting the highest probability match.

Survey application: in well-curated, multiple-system contexts, such as post-enumeration surveys, there are challenges, but they are of an order of magnitude less than in Silverman's application. For post-enumeration examples, visit `https://www.census.gov/coverage_measurement/post-enumeration_surveys`.

Design: it is pleasing that Sir Bernard closes by discussing designs with the potential to improve performance. Including covariates or use of an informative prior distribution may produce identifiability and preserve validity.

Kyle Vincent (*Stittsville*)
My sincerest thanks go to Sir Bernard Silverman for writing such a significant piece on the topic of quantification using multiple-systems estimation (MSE). This paper draws the academic community's attention to the need for more rigorous estimation procedures when studying populations comprised of modern slavery victims. Silverman's use and comparison of frequentist and Bayes approaches highlights the benefits of exploring both approaches when inferring on hidden populations based on administrative lists.

Mark–recapture methods are currently being adopted and tailored for MSE. Within the context of MSE applications for studying hidden populations, there are likely to be more prevalent or new challenges that will manifest. For example, since there is a need to preserve the confidentiality of individuals who are captured on administrative lists for MSE, sharing of detailed information across administrative sources is likely to be limited. Hence, it could be that missing or erroneously entered information exists for linking individuals across lists and/or records of their covariate information. I am hoping that Sir Bernard can make a few comments regarding these challenges, in particular where future work on these topics should be directed.

A Bayesian framework can facilitate goodness-of-fit approaches to assist with choosing a model and its quality of fit based on the posterior predictive distribution. For this reason, does Sir Bernard have any thoughts on using the posterior predictive distribution to aid in choosing the most suitable thresholding parameters? If so, how does the unknown population size affect this approach?

Finally, I am keen on hearing his thoughts on how allowing for monotonically changing values for the λ- and/or τ-parameters, over the iterative steps in the analysis (as presented on the 19th page), can possibly benefit future work of this sort in MSE applications.

Christine P. Chai (*Microsoft Corporation, Redmond*)
I am pleased to see that advanced statistics has been used to address the problem of modern slavery, and this paper is an excellent example of data science for the public good. I also like that Professor Silverman has published the code and data sets, so that readers have a starting point for the implementation.

In the case of counting casualties in armed conflicts, Professor Silverman pointed out the 'dark figure' as an important part because these victims do not appear on any list. The total population estimate should be the sum of the total number of people observed and the dark figure. The Poisson distribution is an obvious choice for count data, but Professor Silverman also includes the interaction terms between the data lists in estimating the dark figure. The correlation between lists is unsurprising because some victims are easier to be found than others. The independence assumption is quite easy to make in statistical models, so I am impressed that Professor Silverman investigated the interaction effects and compared various thresholds for the model.

Since the total number of victims observed is the other part of the casualties, I would like to add some resources for record linkage, where multiple records can belong to the same person. Sadosky *et al.* (2015) applied several blocking methods to partition the casualty data set from the Syrian conflict, to reduce the amount of all-to-all comparisons needed. After the data set has been divided into mutually exclusive and jointly exhaustive blocks, only the records within each block are assumed to have a potential link. A probabilistic approach is locality-sensitive hashing, which assigns similar items to the same block with high probability, and assigns non-similar items to the same block with low probability (Steorts *et al.*, 2014). The locality-sensitive hashing method allows more flexibility than traditional blocking rules in record linkage, so we are likely to discover more potential links between the records.

Last, but not least, quantification of modern slavery requires collaboration across international organizations to obtain more accurate results. Therefore, I would suggest Professor Silverman reaches out to the Human Rights Data Analysis Group (`https://hrdag.org`) and other non-profit organizations in a similar field.

(The opinions and views expressed here are those of the author and do not necessarity state or reflect those of Microsoft.)

Malcolm Faddy (*Queensland University of Technology, Brisbane*)
I have some comments on the assumption of Poisson-distributed counts. Forward selection of interactions (after inclusion of all main effects) by using classical likelihood methods and Poisson log-linear modelling leads to the same six two-way interactions given in Table 15 for the UK six-lists data in Table 1 plus one other, NG.GO, all being significant (p-values less than 0.011). However, using alternative overdispersed and underdispersed extended Poisson process models described in Smith and Faddy (2016) indicates some underdispersion relative to Poisson (twice log-likelihood ratio 3.39 on 1 degree of freedom). This effect is much smaller if the NG.GO interaction is dropped (twice log-likelihood ratio 0.39), and if the least

significant remaining (GO.GP) interaction is also dropped then some overdispersion is apparent (twice log-likelihood ratio 1.27). So the fitted model with six interactions (Table 15) would best conform to an assumption of Poisson residual variation. How crucial then is this Poisson assumption, for the estimate of the 'dark' count is lower from the model closest to Poisson residual variation (8642 with 95% confidence interval 7501–9958) compared with that from the model with the extra (significant) NG.GO interaction (9587 with 95% confidence interval 8193–11217)?

Kuldeep Kumar (*Bond University, Gold Coast*)
I join the chorus that there is a strong need for more research in estimating the victims of heinous crimes such as human trafficking, especially for laying down the foundation for policy implementation. The present study reviews various methods and investigates their performance on a range of real data sets. My suggestion to Professor Silverman would be to utilize the rich data set that he has for various other research avenues using graph analysis techniques which include link prediction analysis, clustering and visualization. These techniques can be used to find suspicious individuals, existing network relationships, unusual changes and anomalous network structure as well as geospatial dispersions. Besides, the visualization of dynamic and complex data in the form of graphs leads users to have a detailed overview of the data, to filter, select and look into networks details. I would be personally keen to have a look at the data and to use the expertise that is at my disposal to do the same and to take advantage of the hard work and commendable efforts put in by Professor Silverman in the collection and compilation of the data from several sources.

The following contributions were received in writing after the meeting.

Serge Aleshin-Guendel, Mauricio Sadinle and Jon Wakefield (*University of Washington, Seattle*)
We thank Professor Silverman for this paper, which will undoubtedly prompt methodological discussions around an important application of multiple-systems estimation (MSE). We point out issues with the approach proposed and connections with existing literature.

Uncertainty from model selection: the proposed 'Bayesian' thresholding approach leads to an understatement of uncertainty. Inferences on the 'dark figure' are still conditioned on the model selected, which is subject to sample variability, so the approach will almost surely have poor operating characteristics (Regal and Hook, 1991; Whitehead *et al.*, 2019).

Connection with spike-and-slab priors: the thresholding approach in the paper is justified as

> 'an approximation to a prior … which is a mixture of an atom of probability at zero and some other distribution'.

Such priors are well known as *spike-and-slab* priors (see for example Rockova *et al.* (2012)) and have already been employed in MSE, where they are presented as a model averaging approach that is equivalent to using spike-and-slab priors (King and Brooks, 2001; Overstall and King, 2014). Although intuitively appealing, the spike-and-slab approach is not feasible in general, as it does not scale to a large number of lists, especially when considering higher order interactions. As an alternative, we are currently working on continuous shrinkage priors (see for example Bhadra *et al.* (2019)) for exploring the space of log-linear models.

Robustness and stability: Professor Silverman relies on notions of 'robustness' and 'stability' to evaluate MSE approaches. These concepts refer to procedures that lead to similar estimates, regardless of the model selected. Although he acknowledges that following these guidelines suggests fitting only main effects models, his proposed approach still focuses on searching over a narrow set of models. At play here is a bias–variance trade-off: the author is favouring a potentially very biased approach in favour of lower variance. It is preferable to have a procedure that provides us with honest assessments of uncertainty and thereby avoids misleading and overconfident results. For illustration, the dark figure is essentially known for the Kosovo data (it is around 6001; see Manrique-Vallier (2016)). None of the 95% intervals obtained with the thresholding approach include this value, but fitting a log-linear model with all three-way interactions gives a 95% interval of [4922, 31584], which includes the known value. This indicates that a method that searches over a larger model space, yet still encourages parsimony, could be beneficial.

Matching: recent advances in MSE with probabilistically linked data (fuzzy matching) include Tancredi and Liseo (2011) and Sadinle (2018). Incidentally, we can only wonder whether undermatching might be partially responsible for the sparse tables that are presented.

Olivier Binette and Rebecca C. Steorts (*Duke University, Durham*)

Silverman (2014) and Bales *et al.* (2015) estimated 10000–13000 potential victims of modern slavery in the UK by using multiple-systems estimation (MSE) via Cormack (1989, 1992) and using the `Rcapture` package of Baillargeon and Rivest (2007), which implements a Poisson log-linear model. Professor Silverman now compares MSE methods in an application to modern slavery and human rights data, proposing a 'Bayesian threshold' approach as a solution to observed instabilities. His approach raises the following questions. Is it reasonable to assume a one model fits all approach? Why does the accuracy of the estimates receive little attention given the fact that many MSE methods were considered?

One model does not fit all

Professor Silverman alludes to the intrinsic difficulties with modern slavery, trafficking and human rights data by stating that

> 'our comparative study using real data sets ... will demonstrate that, unfortunately, all the existing methods display instabilities of various kinds ...'.

However, we are unsure of the cause of these instabilities, and we specifically question his black box application of MSE models. For instance, with the New Orleans data, *only* five lists considered sex trafficking victims, *only* one considered victims of labour trafficking and *only* one considered simultaneous victims of both crimes (Bales *et al.*, 2019). Since this is not accounted for in statistical analyses, it is not immediately clear that the assumptions underlying the use of MSE are satisfied (van der Heijden *et al.*, 2012; Lum *et al.*, 2013). Could the instabilities that are showcased by Professor Silveman be consequences of ignoring such features of the data? Should we expect one model to fit all data sets?

The accuracy of multiple-systems estimation

Professor Silverman states that 'no true prevalence or "ground truth" is available to investigate the accuracy of any estimates'; however, why are simulation studies not utilized to study model robustness and sensitivity? Recently, using a parametric bootstrap, Whitehead *et al.* (2019) studied the accuracy of MSE methods and estimated the coverage of the confidence intervals that were used by Silverman (2014) and Bales *et al.* (2015) to be around 40%. This low coverage occurs as the model selection uncertainty is ignored when the confidence intervals are reported. Does the Bayesian threshold approach proposed by Professor Silverman suffer from a similar issue? Given the thresholding step, do the posterior quantiles reported in Tables 9–14 adequately capture the uncertainty in the estimates?

Pietro Coretto (*University of Salerno*)

I congratulate Professor Silverman for this interesting and inspiring paper. He combines deep statistical insights with methodological craftsmanship to analyse data about social phenomena often ignored. The major challenge here is that a 'ground truth' is not available to investigate the performance of the methods under study. Therefore, Professor Silverman argues that the evaluation of the methods' performance should be based on a careful evaluation of their statistical properties. Here stability is considered the key property. One thing to learn from this paper is certainly the quest for meaningful tuning of the methodology in a difficult environment where classical model validation techniques cannot be used for the aforementioned reasons. Although the trend in our field is to provide users with fully data-driven unsupervised procedures, this paper is an example of why conscientious supervised use of the methods is still a valuable attitude for modern statisticians. Let us introduce two more specific points.

(a) In this paper, stability is understood as a 'not-too-strong sensitivity' to changes in the data (aggregating small lists), or changes in tuning parameters (setting prior distributions). Although I think that it is not possible to formalize a universal notion of statistical stability, is it at least possible to formulate a standard definition that works for this particular problem so that some numerical performance evaluation can be done? This would go into the direction of having an 'agreed standard approach' advocated in Section 1.

(b) The thresholding is the key step of the method proposed, where the pair (λ, τ) is its crucial ingredient. Based on a thorough analysis of the resulting estimates it is argued why prior variance 1 and threshold 2 produced the best results across the data sets. However, these numbers cannot be taken as generally appropriate for any data sets at hand. If the structure of the underlying $\{\beta_{ij}\}$ is sufficiently sparse with a reasonable gap between zero and non-zero coefficients, varying λ and τ we should observe a clear transition towards a stable subset of thresholded β-coefficients. In contrast, in a situation like the Kosovo data set, we expect a smooth pattern and that this sharp transition

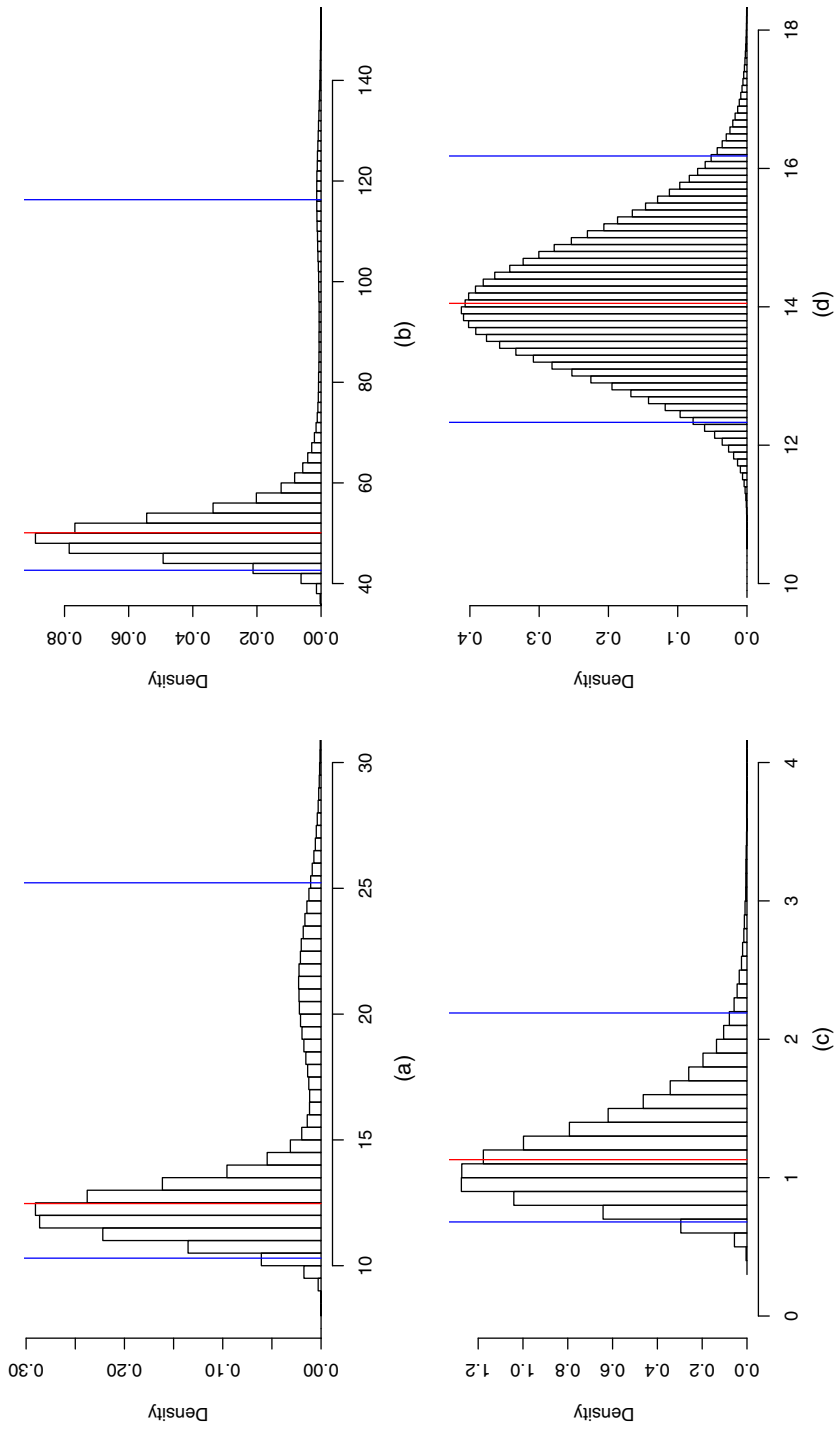

Fig. 9. Estimated posterior distributions of population sample size N in thousands from MCMC runs in a full Bayesian analysis for (a) the UKdat.5, (b) Ned5, (c) NewOrl and (d) Kosovo data sets: the posterior median estimates are 12.47×10^3, 50.09×10^3, 1.13×10^3 and 14.05×10^3 for the respective data sets, as indicated with a vertical line, with 2.5% and 97.5% posterior quantiles on either side

does not happen. It would be interesting whether, based on this, one could build some graphical or numerical tool to be used as a general guideline for setting the thresholding prior parameters.

Ian Dryden, James Goulding and Simon Preston (*University of Nottingham*) **and Lax Chan** (*The Open University, Milton Keynes*)

Professor Silverman's paper gives a fascinating overview of the challenges of estimating population sizes from sets of sparse overlapping lists of individuals. This is a difficult practical problem and requires careful modelling and close collaboration with stakeholders.

An alternative to thresholding is to consider a full Bayesian approach where a mixture prior is given for the interaction parameters in the model. We introduce latent variables $w_i \in \{0, 1\}, i = 1, \ldots, q$, for each of the q interaction parameters, with independent Gaussian mixture priors of two components 0 and 1 with distributions $N(0, \sigma_0^2)$ and $N(0, \sigma_1^2)$ respectively. The latent variable indicates component membership, with mixing proportions p_{mix} for component 0 and $1 - p_{\mathrm{mix}}$ for component 1. It is straightforward and fast to simulate from the posterior by using `MCMCpack` (Martin *et al.*, 2011) for a given set of $w_i, i = 1, \ldots, q$, and then we use an outer loop of Metropolis–Hastings steps for switching w_i labels between 0 and 1 changing one at a time chosen at random.

In Fig. 9 we present some analysis where component 0 has a very small 'spike' variance $\sigma_0^2 = 1/1000$ to indicate that the parameter is effectively 0 with prior probability $p_{\mathrm{mix}} = 0.9$ and $\sigma_1^2 = 1$. For these results the algorithm involves 1000 updates of the latent class variables and in between 1000 updates of all other parameters by using the existing Markov chain Monte Carlo (MCMC) algorithm with the appropriate prior precision matrix (which depends on the latent variables). We have thrown away the first 10% of each chain as burn-in and we have pooled 100 independent runs of the MCMC algorithm. The chains can remain in different modes where interaction parameters are not strongly identifiable, which leads to high or low estimates of the population size as seen in data sets `UKdat_5` and `Ned5`. Such bimodality is also seen in Figs 2, 4, 5 and 6 in the paper. There is little in the data to distinguish between these parameters that lead to very different population estimates. It is difficult to see how this problem could be resolved without either more data or more overlap in the lists.

A final point is about the model assumption that individuals are homogeneous, which is clearly unrealistic. There are various models that take into account specific forms of heterogeneity (Baillargeon and Rivest, 2007). In simulation studies with heterogeneous individuals, we have found the pairwise interaction models of the paper quite effective, whereas the main-effects-only models clearly fail.

Peter G. M. van der Heijden and Maarten Cruyff (*University of Utrecht*)

We are not happy with the presentation of the Dutch data in the paper.

Professor Silverman reduces the number of lists from 6 to 5, taking together the I (Inspectorate) and O (residential treatment centres and shelters) as they have the lowest number of observations. However, such decisions should be driven by substantive grounds, e.g. that the organizations have similar objectives in making the lists and that the relationships between lists and background characteristics are similar. For the data constructed by Professor Silverman, it is unclear what they stand for. Below, we eliminate list K (Border Police), because many individuals on that list never entered the Netherlands because they were denied entrance at the border (see van Dijk *et al.* (2017)).

Contrary to what Professor Silverman states, covariates measuring background characteristics are usually available. For the United Nations Office on Drugs and Crime and the Walk Free Foundation we have produced multiple-system estimates for the Netherlands, the Republic of Ireland, Romania, Slovenia, Serbia and Slovakia, and for all these countries the data included covariates and multiple years.

Making use of covariates is essential. The instabilities that Professor Silverman observes are due to the presence or absence of certain two-factor interactions between lists. Inclusion of covariates, however, may replace list–list interactions by list–covariate interactions. This is because the covariates 'cause' list–list interactions when marginalizing over them. The International Working Group for Disease Monitoring and Forecasting (1995) have coined this *apparent dependence*.

To illustrate this, we first fitted all two-factor interaction models to the data with our choice of the five lists. Fig. 10 shows the relationship between the abundance estimates and the Akaike information criterion (AIC) and Bayesian information criterion (BIC) values of these models. Here the estimates lie in the range of 20000–70000: much smaller than the range in Fig. 4 of the paper.

All models in Fig. 10 with estimates in excess of 60000 have the interactions IZ, OP, OZ, PR and PZ in common. The models (IZ, OP, OZ, PR, PZ, RZ) and (IZ, OP, OZ, PR, PZ), as depicted by the black dots in Fig. 10, respectively have the lowest AIC and BIC values. Subsequently we have fitted these models

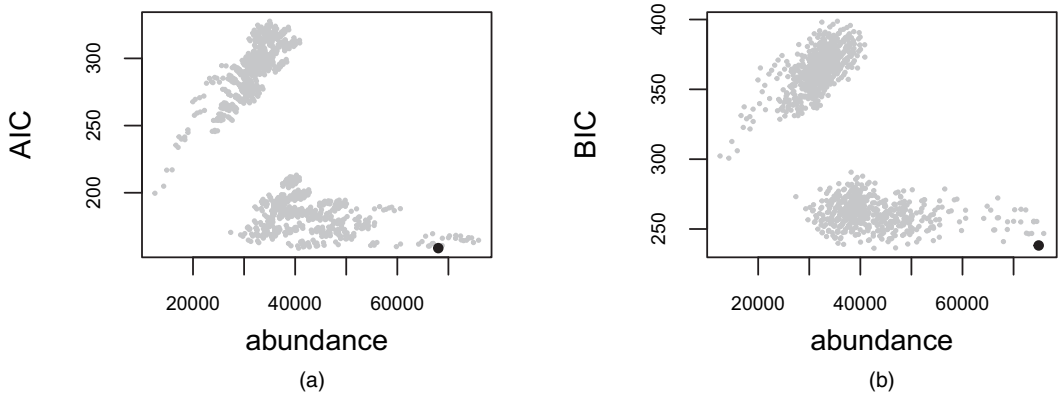

Fig. 10. Abundance estimates of all two-factor interaction models for the five-list Netherlands data based on (a) the AIC and (b) the BIC

with the main effects of the covariates sex, age, nationality and type of exploitation included, and then performed a stepwise model search. The AIC model yields an abundance estimate of 44000, with OP no longer in the model, and the BIC model yields an estimate of 42000, with IZ, OP, PR and PZ no longer in the model. The estimate produced in van Dijk *et al*. (2017) is also 42000.

James Jackson (*Lancaster University*)
The modern slavery data sets that are suitable for multiple-systems estimation (MSE) analysis are sparse. From Tables 1–3, the number of observable cells with 0-counts are, for the UK, 38/63 (60%), for the Netherlands, 39/63 (62%), and, for New Orleans, 236/255 (93%). Professor Silverman expects this to be the norm and has diligently developed methods to account for sparseness (see Chan *et al.* (2020)). Expecting sparseness is equivalent to expecting negative dependences between lists. Understanding why this is so is intrinsic for current methods to develop.

In the MSE model, for any list combination A, every individual in the population has the same probability of being in A. This may be unrealistic in a modern slavery environment, where several factors affect the visibility of a case; some poor victims may even have a probability of detection that is close to 0. The level of heterogeneity may also vary between lists, which would give rise to negative correlation. Agencies may have different abilities of reaching specific groups of the population. For example, a list deriving from the general public may be effective at identifying victims from car washes, but less effective at observing workers growing cannabis. Or an agency may be effective at identifying victims in one geographical area, but less so in another.

Moreover, undermatching would create a false impression of negative list dependences. Anonymized data make it difficult—or even impossible—to match individuals across lists, leading to one-list counts that are too high and multiple-list counts that are too low.

Neither the UK four-list data nor the Kosovo data (Table 4) contain any 0-counts, and estimates resulting from these data display impressive stability. Since MSE enables lists to be combined, the problem of sparse data could be relieved by combining lists in such a way as to minimize the number of 0-count cells. But, to begin to address the issue of sparseness, information about the list sources is essential. In the New Orleans data, even the names of the list sources are suppressed. To end with the 'fish in the lake' analogy, it needs to be checked that each list is catching from the same lake!

Antony Overstall (*University of Southampton*) **and Ruth King** (*University of Edinburgh*)
Multiple-systems estimation has a long history but has only recently been applied to modern slavery. Discussion of the multiple-systems estimation assumptions and their suitability to modern slavery would be useful, particularly as violations can have a significant effect on the population estimates (e.g. Overstall *et al.* (2014)).

A Bayesian threshold approach is used (applied to the ratio of the posterior mean to standard deviation) considering only two-way interactions. We raise two specific issues.

(a) Estimation of the total population size is dependent on a single model, and model estimates can vary substantially.

Table 19. Quantiles of the posterior distribution of the total population size and Bayesian p-value under the χ^2-discrepancy (small or large values indicate a lack of fit)

Data set	*Results ($\times 10^3$) for the following quantiles of the posterior:*					*Predictive p-value*
	2.5%	*10%*	*50%*	*90%*	*97.5%*	
UK—4 lists	9.6	10.2	11.5	12.9	13.8	0.14
UK—5 lists	10.2	10.9	12.5	25.5	30.3	0.28
UK—6 lists	10.3	10.9	12.2	14.2	21.4	0.52
Netherlands—5 lists	106.2	113.3	133.6	152.3	164.4	0.12
Netherlands—6 lists	105.8	113.7	129.4	149.0	161.8	0.04
New Orleans—5 lists	0.6	0.7	1.0	1.5	1.9	0.17
Kosovo	8.8	9.4	10.8	14.4	18.5	0.37

(b) Step 2 of the algorithm (removing interaction terms) essentially assumes that posterior parameter distributions are independent of the fitted model.

Model-averaged approaches have been proposed to obtain population estimates (see Section 5.1 for decomposable graphical models) and simultaneously provide informative posterior probabilities of the presence or absence of interactions. King and Brooks (2001) applied a model-averaged approach considering the set of all log-linear models, with a hierarchical prior structure on the interaction terms, and applied via a transdimensional Markov chain Monte Carlo algorithm. The R package `conting` (Overstall and King, 2014) implements this approach and incorporates other features including checks of model adequacy and inclusion of covariates. Further, `conting` has no upper bound on the total population size nor limits the number of lists (as for `dga`).

For the data considered in Section 6, the corresponding model-averaged estimates obtained by `conting` are presented in Table 19, in addition to the posterior predictive p-value using a χ^2-discrepancy function. Only two-way interactions are considered, excluding any such terms leading to estimation issues due to parameter redundancy. The exception is for Kosovo, where additional three-way interactions are considered because of apparent model inadequacy by using only two-way interactions; the Netherlands (six-lists) data also suggested some lack of fit but it was difficult to fit more complex models because of the limited overlap of lists.

Similar results are obtained to those of Table 7 (for model-averaged decomposable models) except for the Netherlands. The results are generally slightly larger than for the threshold approach representing the additional model uncertainty. For the Netherlands data, no individuals are observed by both lists I (Inspectorate) and K (Border Police), or K and R (regional co-ordinators)—thus these terms should typically be omitted from the analysis (see Section 6.2 and Sharifi Far *et al.* (2020)). The inclusion of these interactions terms in the analyses in the paper and exclusion in our analyses appear to lead to the discrepancy observed.

Louis-Paul Rivest and Sophie Baillargeon (*Université Laval, Québec*)

We thank and congratulate Professor Silverman for bringing a new type of capture–recapture data, concerned with human slavery, to the attention of the statistical community and for his methodological contributions to analysing this type of data. Our comments will mostly focus on Sections 3 and 4 of the paper.

Bayesian information criterion or Akaike information criterion: in a stepwise search for a model (see Tables 5 and 6), there are no differences between the Bayesian information criterion (BIC) and Akaike information criterion as they select the same model at each step. As a global criterion they, of course, lead to different estimates as highlighted in Section 4. Which criterion to use might depend on whether there is heterogeneity in the data (Rivest and Baillargeon, 2014). Such a heterogeneity is likely in human right data sets such as the Kosovo data (Lum *et al.*, 2013). It typically translates into fitted log-linear models with many positive interactions (Rivest, 2011) and with population size estimates $\hat{N}$ that increase with the number parameters; Fig. 11. The more stringent BIC might then be a better criterion as it offers

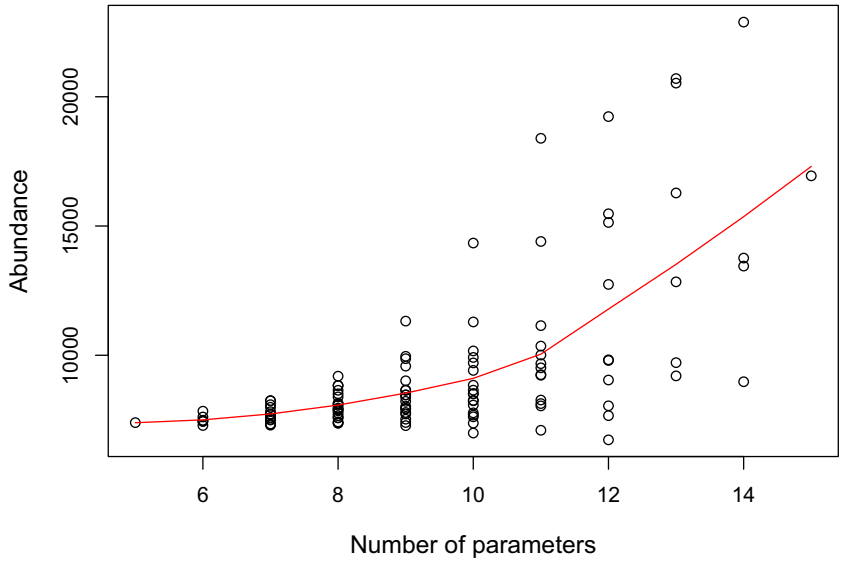

Fig. 11. Estimate $\hat{N}$ as a function of the number of log-linear parameters for the Kosovo data

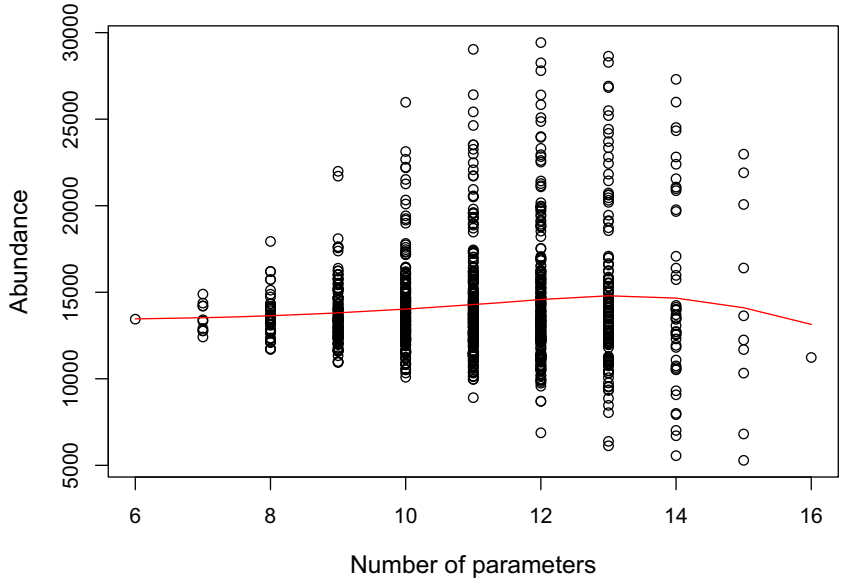

Fig. 12. Estimate $\hat{N}$ as a function of the number of log-linear parameters for 2^{10} models fitted to the UK five-list data

protection against large $\hat{N}$ associated with an overfitted model. Fig. 12, constructed with 2^{10} log-linear models containing bivariate interactions fitted to the UK five-list data, does not show such an upward trend. The case for the BIC is not as strong for this data set as demonstrated in Section 4.2.

Sparsity: for the UK data, less than 10% of the recorded units belong to more than one list. There is a need to investigate whether the methodology used to produce Tables 5 and 6 is reliable. In `Rcapture`, we address this problem by associating warnings to estimates $\hat{N}$. The `glm` warnings for the Poisson regression fit, such as *fitted rates numerically 0 occurred* or *the algorithm did not converge*, are flagged in the `infoFit` column in the printed output of any `closedp` function. Also the asymptotic bias of each log-linear estimate $\hat{N}$ is

evaluated as proposed in Rivest and Lévesque (2001) and is available in the `bias` output object. Instances where this absolute bias is larger than 10% of $\hat{N}$ are flagged as being indicative of a problematic model. Clearly, more work is needed to ascertain the validity of likelihood-based inference for sparse capture–recapture data and we commend Professor Silverman for highlighting the importance of these issues.

Serveh Sharifi Far (*University of Edinburgh*) **and Michail Papathomas** (*University of St Andrews*)
This interesting paper highlights the important issue of estimation and uncertainty quantification for the size of hidden populations. The modelling approaches presented consider only the inclusion of first-order interactions, which is a restriction that can have a considerable effect on inferences. Although we recognize that allowing for higher order interactions will increase the probability of non-estimability of parameters, further discussion on the balance between non-estimability and allowing for a less smooth joint distribution between contingency table factors would be useful.

Our main interest in the paper lies in model identifiability within the classical framework, which we approach as an 'extrinsic parameter redundancy' problem. The approach utilized in this paper is based on checking the existence of the maximum likelihood estimator using linear programming, presented by Fienberg and Rinaldo (2012a, b). For a non-identifiable model, this method provides a subset of the initial parameters of the model as the estimable parameters. In contrast, the parameter redundancy approach (as described in Sharifi Far *et al.* (2019)) provides a set of estimable parameters that may include linear combinations of the initial parameters. Thus, when considering the identifiability of the model adopted, the parameter redundancy approach may lead to a different reduced log-linear model, in terms of the model parameters. We believe this is important as the estimates of the model main effects and interactions are of interest here.

As an example, consider analysing the given Netherlands data in the paper. The main effects model, after a stepwise Akaike information criterion approach, gives a point estimate of 52.8×10^3 with 95% confidence interval (CI) $(48.3 \times 10^3, 57.3 \times 10^3)$. Adding first-order interactions to this model results in having two non-estimable parameters (I:K and K:R interactions: note that the model in Table 15 includes these two parameters that are non-estimable within the classical framework). After reducing the model and using a forward and backward step Akaike information criterion method, the population estimate is increased to 95.3×10^3, with 95% CI $(70.7 \times 10^3, 119.8 \times 10^3)$. Adding second-order interactions to the model makes 17 parameters non-estimable and the population estimate is highly increased to 298.8×10^3 with 95% CI $(139.3 \times 10^3, 458.2 \times 10^3)$.

Finally, we believe it is worth noting that a possible identifiability problem, which could arise in other data sets, is that, because of the specific pattern of 0s in the data, the intercept of the model may not be estimable. This will prohibit the estimation of the 'dark figure'.

Joe Whittaker (*Lancaster University*)
This is an interesting paper; however, I believe that certain key results, the estimates of the missing population given in Figs 5(a) and 6(b), are spurious and misleading. The figures show a mixture of two components: whereas the lower component is informative, the component giving the upper mode is an unwitting fabrication of numerical inaccuracies due to the non-convergence of the fitting algorithm applied.

In the context of maximum likelihood the (so-called) multiple-system estimation method smooths the multiway table of observed counts by fitting a low order log-linear model to the observed tabular data. It uses the resulting approximation to extrapolate the counts to the missing cell and so gives an estimate of the missing population. The fitting routine maximizes a log-likelihood by using the iteratively reweighted least squares algorithm to find the optimum. It is well known that an optimum is characterized by equating observed and fitted margins, and the margins are determined by the minimal sufficient statistic of the log-linear model.

Numerical problems arise when this minimal statistic contains 0s. The corresponding cell in the fitted margin has a linear predictor that must take the value $-\infty$. Consequently more than one of the fitted log-linear parameters must be unbounded. What happens in numerical practice is that during the iteration certain parameters attain very large (finite) values. The iteration terminates when a numerical criterion is met (e.g. 100 iterations) and wild parameter values are reported! Similar remarks apply to the Bayesian approaches of Silverman.

The default starting values are not problematic; the reported fitted values are reasonable because they are close to observed, often the result of judicious parameter cancellation; and consequently goodness-of-fit measures are believable. However, the log-linear intercept parameter is not believable; nor is the extrapolation to the missing cell.

Consequently in the repetitions made for Figs 5 and 6 the higher estimates are often spurious, and particularly misleading because they cast aspersions on the credibility of the lower estimates.

To avoid this pitfall a precaution is to compute convergence of the iteration with a criterion that is sensitive to movement in the log-linear parameter space. Also, one can identify those log-linear models which lead to unbounded parameter estimates from an effective rank criterion, before fitting.

However, these steps do not solve the issue of population inference when the data have sparse low dimensional margins. My feeling and that of colleagues Maarten Cruyff and Peter van der Heijden is that regularization is the right approach. A fully Bayesian analysis with proper priors should also eliminate spurious numerical artefacts.

Congratulations go to the author for a stimulating paper.

The **author** replied later, in writing, as follows.

I am extremely grateful to the discussants of the paper. My hope, which has certainly been realized, was that the paper and the discussion together would provide a resource for future development of methodology in this important area, and there are many suggestions which will repay future study. It may be helpful to draw out some common themes among the points that have been raised.

The paper concentrated on data that were currently available in the public domain and that was why it took what some believe to be an inappropriately simplistic approach. My own view is that it is better to attempt to answer the question at hand, but to be clear about the assumptions that have been made. One discussant explicitly says that attempting to estimate the 'dark figure' is a hopeless quest and that we should instead concentrate on what is actually observed. This is of course a perennial issue in criminology; the problem with any hidden crime is that, if the reported numbers increase, we do not know whether the underlying phenomenon has changed or whether we have become better at spotting it.

Another calls me to task for not using covariate information but refers to data that are not in the public domain, thereby making it impossible to verify or build on the cited work. This underlines the need for as much as possible to be made available to researchers, so that we can all make progress on developing methodology. However, the lack of detail in published data is often for serious reasons of privacy and confidentiality. Obviously we would like, if possible, to know more about the various lists and the individuals on them, but data in this area are often collected under conditions where assuring complete anonymity is morally, and often legally, essential. In New Orleans, some of the agencies were not allowed to be identified and others were happy to disclose data only for collation by a trusted person on condition of anonymity. In the case of the UK data set, only a single individual, who spent several weeks assembling the data, was allowed to see any more detail than is presented in Table 1. We might wish to have more information, but we also have to do the best with what is actually available. We cannot insist, as one discussant suggests, that 'information about the list sources is essential'.

Some of the robust discussion of the paper was only possible because I put my own data and software into the public domain, and I am grateful to the discussant who thanked me for this. Both of the criticisms discussed above raise a serious dilemma about the perfect being the enemy of the good, or even the quite good being the enemy of the not quite so good. In this, I side with George Box, who famously wrote (e.g. Box (1979)) 'All models are wrong, but some are useful'. But an interesting topic for future research might be to design some sort of anonymization, perhaps involving encryption, that would allow more information about data to be disclosed without compromising necessary confidentiality. The prize would be that researchers would be able to develop methodology on much more granular data than are currently available.

Clearly, where there is information, it should be used. The general public list GP within the UK data, referred to by more than one discussant, is a case in point. Performing the analysis with and without that list was an important robustness check, and a similar point can now be made about list K (Border Police) within the Netherlands data, though including that list may mean that we are simply estimating the size of a wider population, those that could possibly come to the authorities' attention, whether or not they actually enter the country.

I agree completely with those respondents who expressed the desirability of including covariate information. As already stated above, there is a paucity of publicly available data sets which contain such information in this area, but extending Bayesian approaches to deal with covariates, and of course with the issue of fuzzy matching and uncertain linkages, is an obvious topic for future research. It is clear that the population is not homogeneous and that different segments of the population will behave in different ways. Whatever method is being used, it can be conceptually elaborated to handle data with known covariates, but the details will need to be worked out. A word of warning though: within the multiple-systems

framework, if we were to break down our observed data into subpopulations, then the problem of sparse tables could become more acute. If covariates are continuous rather than discrete this may not explicitly be so much of an issue, but it will still need to be handled. The bigger question remains: if we do not have any covariate information, should we still do what we can, while of course being explicit about our assumptions?

What is the underlying justification for the Bayesian model? This is very much tied up with issues around parsimony. To quote George Box again (Box, 1976):

> 'Since all models are wrong the scientist cannot obtain a "correct" one by excessive elaboration. On the contrary following William of Occam he should seek an economical description of natural phenomena. Just as the ability to devise simple but evocative models is the signature of the great scientist so overelaboration and overparameterization is often the mark of mediocrity'.

The most parsimonious multiple-systems model would be a model which is usually unrealistic in our context: to assume that the lists are independent of each other. Most of the approaches to model selection in multiple-systems estimation seek, in one way or another, to formalize the notion of seeking a parsimonious model, whether through information criteria which penalize for the inclusion of additional parameters, stepwise approaches which add parameters successively or Bayesian spike-and-slab approaches, such as the approach the paper aims to approximate, which give absence of a parameter a high prior probability. One discussant claims that a spike-and-slab approach is not feasible for larger numbers of lists, whereas another actually implements it! The Bayesian–thresholding method does focus on a particular model, and the full implementation is indeed preferable if it is feasible. In some circumstances there may be scope for a hybrid approach, where some sort of thresholding is used to narrow down the set of effects whose presence or absence from the model is still uncertain on the basis of the data observed.

What causes sparsity in tables? Some discussants suggest that it is the result of undermatching. I genuinely do not think that this is the main factor in the case of modern slavery. Clearly a negative correlation between capture on two lists will make it more likely that they will not show any cases in common, but that possibility is built into the modelling. The main reason, though, is that any list is likely to contain only a relatively small proportion of the actual population, and so overlaps, in particular between three or more lists, are increasingly likely to be empty just by chance. This does seem to be a qualitative difference between the modern slavery context and some others, but, even if there is scepticism about multiple-systems estimation for modern slavery, the issues raised by sparsity are clearly deserving of careful attention whatever the context. As an aside, the extension of the methodology of the paper to incorporate three-list effects is straightforward; it was not considered formally because of the very low levels of higher order overlaps.

Some important technical issues around sparsity are touched on in Section 4.4 of the paper and dealt with in much more detail by Chan *et al.* (2020). Several discussants made points where these matters are relevant. One cannot simply regard certain interactions as being 'inestimable' and omit them from the analysis. Rather, if one is to use a maximum likelihood approach, as the paper explains, one needs to move to an extended maximum likelihood framework which allows $-\infty$ as a parameter value but still checks whether estimates within this extended context exist. There is no 'non-convergence of the fitting algorithm' if this issue is approached carefully; nor are `glm` warnings generated. Several discussants use methods based on information criteria even though, as pointed out in the paper, the underlying theory does not hold in the extended scenario; I am glad that a discussant agrees with me that a more detailed dispassionate investigation of whether this really matters in practice would be an interesting topic for future research.

How should we account for model selection when assessing the accuracy of an estimate? Within the frequentist paradigm, an obvious approach is to use the bootstrap. Whatever method is being used to select and fit the model, this can be applied to repeated bootstrap samples, and thereby a confidence interval can be obtained that is not conditional on the particular model fitted. Chan *et al.* (2020) set this approach out in detail in the context of a stepwise method of choosing and fitting the model, but in principle the same approach could be used whatever model selection method is used. The only *caveat* is, of course, that the method should be sufficiently economical computationally that it can be used on the 1000 or more bootstrap replications that are drawn. The BC_a method of DiCiccio and Efron (1996) is particularly appropriate because it is second-order accurate and is equivariant under whatever transform is being used in the population domain. It requires a jackknife step, where individual data points are left out and the estimate recalculated, but Chan *et al.* (2020) describe an algorithmic approach (in principle applicable to any model selection method) that can shorten this step dramatically. Fuller details of results by using the bootstrap approach are given there, but for the UK data the method, using a stepwise method for fitting

the model, yields a 95% confidence interval of $(9.3 \times 10^3, 14.3 \times 10^3)$ for the number of victims. Combining the PF (Police Forces) and NCA (National Crime Agency) lists, and/or omitting the GP list, does not substantially affect this result. It is not surprising that this interval is somewhat wider than the interval $(10 \times 10^3, 13 \times 10^3)$ conditional on the chosen model. A limited simulation study, from a somewhat more parsimonious model than the model allowing all possible two-list effects, confirms the coverage probability of the BC_a confidence intervals.

To what extent should we rely on simulation studies? It is clearly important to devise simulation studies that are relevant to the problem in hand. One simulation model proposed in the discussion is to fit all two- and three-list parameters to the UK data set and then to simulate from that fitted model. Such a model will be by any account overparameterized and indeed gives a total population which is only slightly larger than the number of cases actually observed. This model overfits when considered statistically, but more to the point is clearly not based on a good estimate. Whatever is the true figure of UK victims, it is indisputable (not least from figures in subsequent years, reflecting increasing consciousness of modern slavery, as well as the new legislative framework) that in 2013 the dark figure could not have been only in the hundreds. Perhaps this is an extreme case, but it illustrates a wider issue about whether non-parsimonious models are a good basis for simulation studies in this context. It does, however, suggest a possible future project. Given the number of different methods for multiple-systems estimation that have now been proposed, is there any possibility of agreeing a 'simulation competition' on a range of simulation models, assessing both longer-standing and more recently proposed methods? This would be an interesting complement to the comparison of methods on real data sets. In this context, one interesting proposal in the discussion is that of starting with a real data set and then conditioning on the subpopulation of those that fall on a particular list.

It was very pleasing that the discussion referred to the use of multiple-systems estimation in the census, both in the UK and the USA. The need for approaches to the quantification of hard-to-count populations is, if anything, likely to increase over time, for three reasons. Firstly, because of long-term social and economic developments, some groups may become increasingly hard to reach by conventional methods. Secondly, the level of resource available, e.g. in terms of actual enumerators who visit addresses in person, may decrease. And thirdly the expectation that the census will accurately capture such populations, and the importance of doing so, is increasing. However, individuals will show up on a range of administrative sources as well as on post-enumeration surveys. Multiple-systems approaches, no doubt combined with other methods, may well be relevant.

My colleague Kevin Bales has drawn the comparison between our present understanding of modern slavery and the state of knowledge some decades ago about climate change. Public and political consciousness about modern slavery has been raised in recent years, partly because of prevalence estimates. As in the case of climate change, it will be a long process to gain a detailed understanding of the nature of modern slavery and to learn what works best in combating it. Several discussants have rightly pointed out the need to know more about individual sectors, about the effect within demographic and geographic groups, about the more precise nature of what happens to victims both before they have been identified and after they have been freed, and so on. It is of great importance to understand societal, economic and individual risk factors, especially those that can be influenced or mitigated, to address the problem at source. Prevalence estimation has a role to play in all these aspects, and it is my hope that the paper and discussion will inform and prompt further research into this enormously important and disturbing area of human rights and public policy.

References in the discussion

Baillargeon, S. and Rivest, L.-P. (2007) Rcapture: loglinear models for capture-recapture in R. *J. Statist. Softwr.*, **19**, no. 5, 1–31.

Bales, K., Hesketh, O. and Silverman, B. W. (2015) Modern slavery in the UK: how many victims? *Significance*, **12**, no. 3, 16–21.

Bales, K., Murphy, L. T. and Silverman, B. W. (2019) How many trafficked people are there in Greater New Orleans?: Lessons in measurement. *J. Hum. Traffckng*, to be published, doi https://doi.org/10.1080/23322705.2019.1634936.

Bhadra, A., Datta, J., Polson, N. G. and Willard, B. (2019) Lasso meets horseshoe: a survey. *Statist. Sci.*, **34**, 405–427.

Bird, S. M. (2019) Public health perspective on UK-identified victims of modern slavery. Submitted to *Crime Delinq*.

Bird, S. M. and King, R. (2018) Multiple systems estimation (or capture-recapture estimation) to inform public policy. *A. Rev. Statist. Appl.*, **5**, 95–118.

Box, G. E. P. (1976) Science and statistics. *J. Am. Statist. Ass.*, **71**, 791–799.
Box, G. E. P. (1979) Robustness in the strategy of scientific model building. In *Robustness in Statistics* (eds R. L. Launer and G. N. Wilkinson), pp. 201–236. Cambridge: Academic Press.
Chan, L., Silverman, B. W. and Vincent, K. (2020) Multiple systems estimation for sparse capture data: inferential challenges when there are non-overlapping lists. *J. Am. Statist. Ass.*, to be published, doi 10.1080.01 621459.2019.1708748.
Cormack, R. M. (1989) Log-linear models for capture-recapture. *Biometrics*, **45**, 395–413.
Cormack, R. M. (1992) Interval estimation for mark-recapture studies of closed populations. *Biometrics*, **48**, 567–576.
DiCiccio, T. J. and Efron, B. (1996) Bootstrap confidence intervals. *Statist. Sci.*, **11**, 189–228.
van Dijk, J. J., Cruyff, M., van der Heijden, P. G. M. and Kragten-Heerdink, S. L. J. (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, Vienna.
Fienberg, S. E. (1972) The multiple recapture census for closed populations and incomplete 2k contingency tables. *Biometrika*, **59**, 591–603.
Fienberg, S. E. and Rinaldo, A. (2012a) Maximum likelihood estimation in log-linear models. *Ann. Statist.*, **40**, 996–1023.
Fienberg, S. E. and Rinaldo, A. (2012b) Maximum likelihood estimation in long-linear models: supplementary materials. *Technical Report*.
Freeman, P. (1989) The performance of the two-stage analysis of two-treatment, two-period cross-over trials. *Statist. Med.*, **8**, 1421–1432.
van der Heijden, P. G. M., Whittaker, J., Cruyff, M., 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. *Ann. Appl. Statist.*, **6**, 831–852.
Hook, E. B. and Regal, R. R. (1995) Capture-recapture methods in epidemiology: methods and limitations. *Epidem. Rev.*, **17**, 243–264.
International Working Group for Disease Monitoring and Forecasting (1995) Capture–recapture and multiple record systems estimation: Part 1, history and theoretical development. *J. Am. Epidem.*, **142**, 1059–1068.
King, R. and Brooks, S. (2001) On the Bayesian analysis of population size. *Biometrika*, **88**, 317–336.
Louis, T., Carvalho, B., Fallin, M., Irizarry, R., Li, Q. and Ruczinski, I. (2011) Association tests that accommodate genotyping errors. In *Bayesian Statistics 9*, pp. 393–420. Oxford: Oxford University Press.
Lum, K., Price, M. E. and Banks, D. (2013) Applications of multiple systems estimation in human rights research. *Am. Statistn*, **67**, 191–200.
Manrique-Vallier, D. (2016) Bayesian population size estimation using Dirichlet process mixtures. *Biometrics*, **72**, 1246–1254.
Martin, A. D., Quinn, K. M. and Park, J. H. (2011) MCMCpack: Markov Chain Monte Carlo in R. *J. Statist. Softwr.*, **42**, no. 9, 1–21.
Overstall, A. M. and King, R. (2014) conting: an R package for Bayesian analysis of complete and incomplete contingency tables. *J. Statist. Softwr.*, **58**, no. 7, 1–27.
Overstall, A. M., King, R., Bird, S. M., Hutchinson, S. J. and Hay, G. (2014) Incomplete contingency tables with censored cells with application to estimating the number of people who inject drugs in Scotland. *Statist. Med.*, **33**, 1564–1569.
Regal, R. R. and Hook, E. B. (1991) The effects of model selection on confidence intervals for the size of a closed population. *Statist. Med.*, **10**, 717–721.
Rivest, L.-P. (2011) A lower bound model for multiple record systems estimation with heterogeneous catchability. *Int. J. Biostatist.*, **7**, 1–21.
Rivest, L.-P. and Baillargeon, S. (2104) Capture-recapture methods for estimating the size of a population: dealing with variable capture probabilities. In *Statistics in Action: a Canadian Outlook* (ed. J. F. Lawless), pp. 289–304. Boca Raton: CRC Press.
Rivest, L.-P. and Lévesque, T. (2001) Improved log-linear model estimators of abundance in capture-recapture experiments. *Can. J. Statist.*, **29**, 555–572.
Ročková, V. and George, E. I. (2018) The spike-and-slap LASSO. *J. Am. Statist. Ass.*, **113**, 431–444.
Rockova, V., Lesaffre, E., Luime, J. and Löwenberg, B. (2012) Hierarchical Bayesian formulations for selecting variables in regression models. *Statist. Med.*, **31**, 1221–1237.
Ruczinski, I., Kooperberg, C. and Leblanc, M. (2003) Logic regression. *J. Computnl Graph. Statist.*, **12**, 475–511.
Sadinle, M. (2018) Bayesian propagation of record linkage uncertainty into population size estimation of human rights violations. *Ann. Appl. Statist.*, **12**, 1013–1038.
Sadosky, P., Shrivastava, A., Price, M. and Steorts, R. C. (2015) Blocking methods applied to casualty records from the Syrian conflict. *Preprint arXiv:1510.07714*.
Schwender, H. and Ickstadt, K. (2008) Identification of SNP interactions using logic regression. *Biostatistics*, **9**, 187–198.
Sharifi Far, S., Papathomas, M. and King, R. (2019) Parameter redundancy and the existence of MLE in log-linear models. *Statist. Sin.*, to be published, doi 10.5705/ss.202018.0100.

Silverman, B. W. (2014) Modern slavery: an application of multiple systems estimation. Home Office, London. (Available from `https://www.gov.uk/government/publications/modern-slavery-an-application-of-multiple-systems-estimation`.)

Smith, D. M. and Faddy, M. J. (2016) Mean and variance modeling of under-dispersed and over-dispersed count data. *J. Statist. Softwr.*, **69**, no. 6, 1–23.

Steorts, R. C., Hall, R. and Fienberg, S. E. (2016) A Bayesian approach to graphical record linkage and deduplication. *J. Am. Statist. Ass.*, **111**, 1660-1672.

Steorts, R. C., Ventura, S. L., Sadinle, M. and Fienberg, S. E. (2014) A comparison of blocking methods for record linkage. In *Proc. Int. Conf. Privacy in Statistical Databases*, pp. 253–268. New York: Springer.

Tancredi, A. and Liseo, B. (2011) A hierarchical Bayesian approach to record linkage and size population problems. *Ann. Appl. Statist.*, **5**, 1553–1585.

Whitehead, J., Jackson, J., Balch, A. and Francis, B. (2019) On the unreliability of multiple systems estimation for estimating the number of potential victims of modern slavery in the UK. *J. Hum. Traffckng*, to be published, doi 10.1080/23322705.2019.1660952.

Zult, D., de Wolf, P.-P., Bakker, B. and van der Heijden, P. G. M. (2019) A general framework for multiple-recapture estimation that incorporates linkage error correction. *Discussion Paper 2019-07*. Centraal Bureau voor de Statistiek, The Hague. (Available from `https://www.cbs.nl/-/media/_pdf/2019/19/wmr-model-def-2019dp07.pdf`.)