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:

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

 0964–1998/20/183000

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 '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.

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
×	×		×			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).

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

The data that are analysed in the paper and the programs that were used to analyse them can be obtained from

`https://rss.onlinelibrary.wiley.com/hub/journal/1467985x/series-a-datasets`.

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).

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 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`.

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).

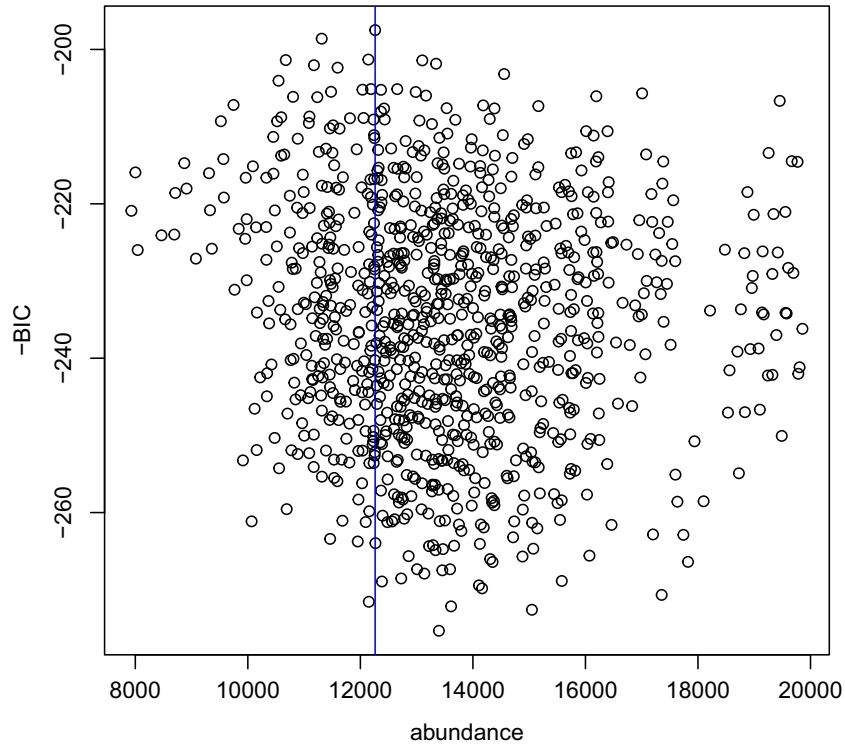

Fig. 1. Estimates of abundance plotted against the BIC, with outliers omitted, the default plot option

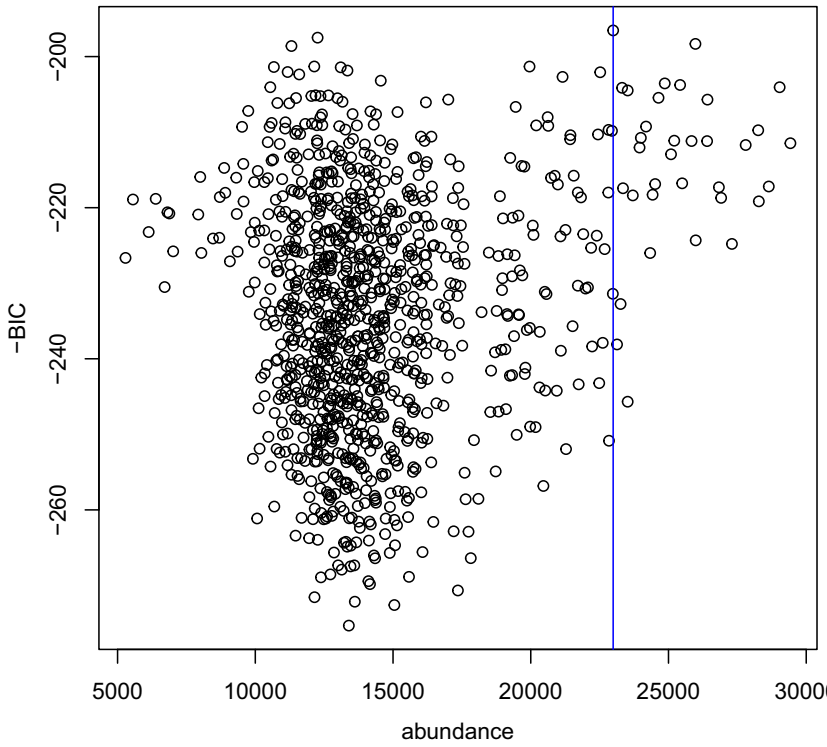

Fig. 2. Estimates of abundance plotted against the BIC, with outliers included

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 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.

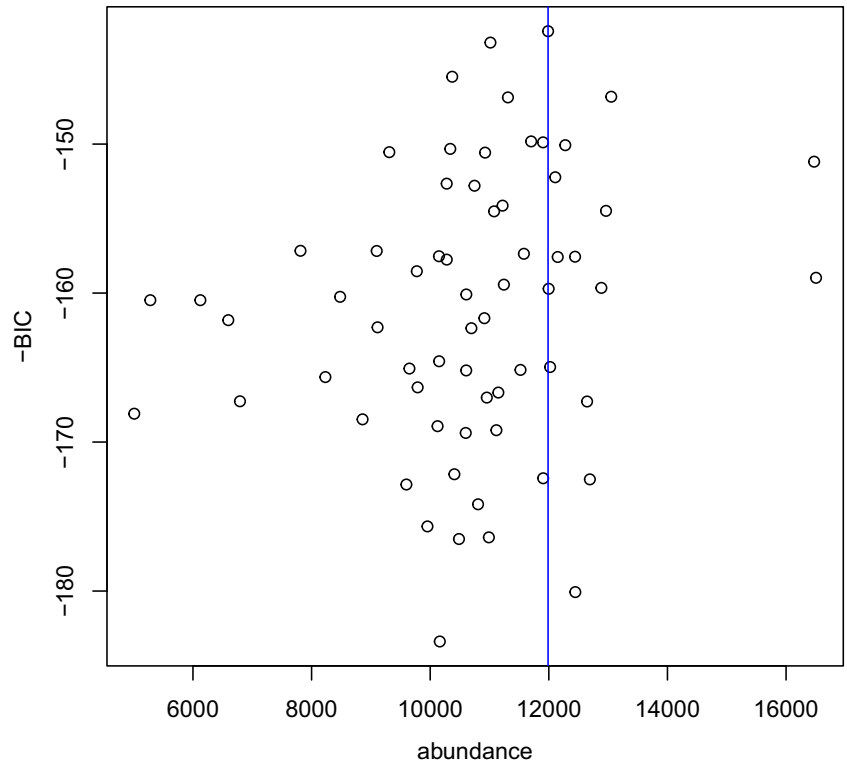

Fig. 3. Estimates of abundance plotted against the BIC, with the GP list excluded

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.

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 double 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.

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.

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.* (2019).

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.* (2019). 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.* (2019) 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.* (2019). 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.

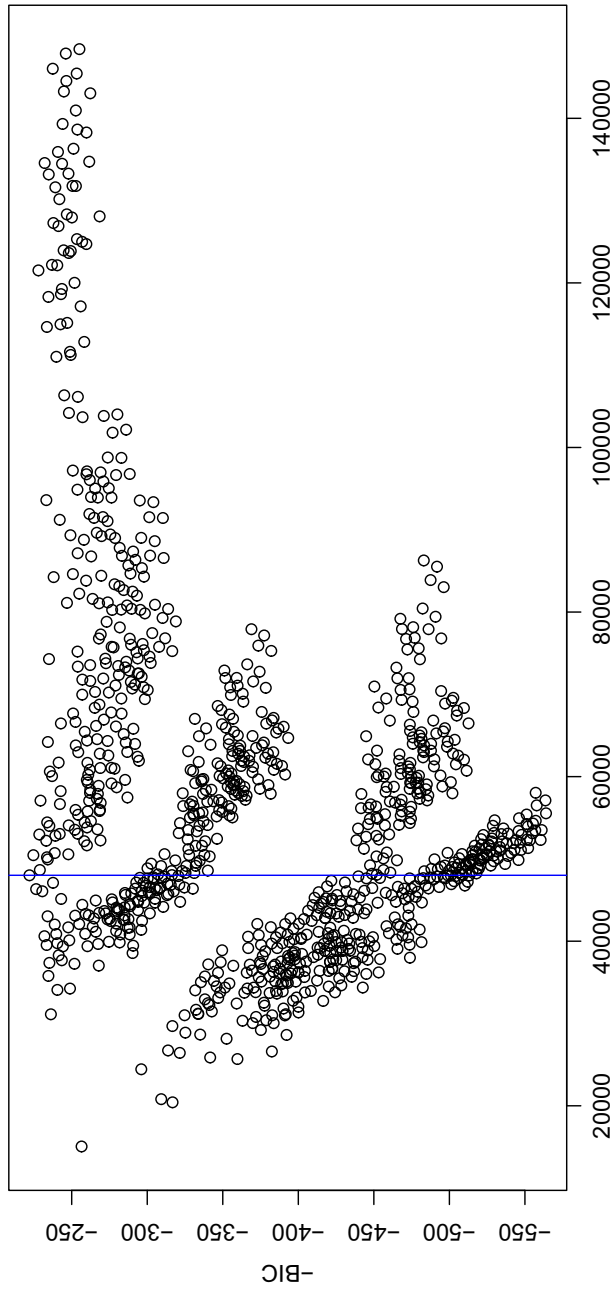

Fig. 4. Estimates of abundance plotted against the BIC, for the Netherlands data consolidated into five lists

Chan *et al.* (2019) 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 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.

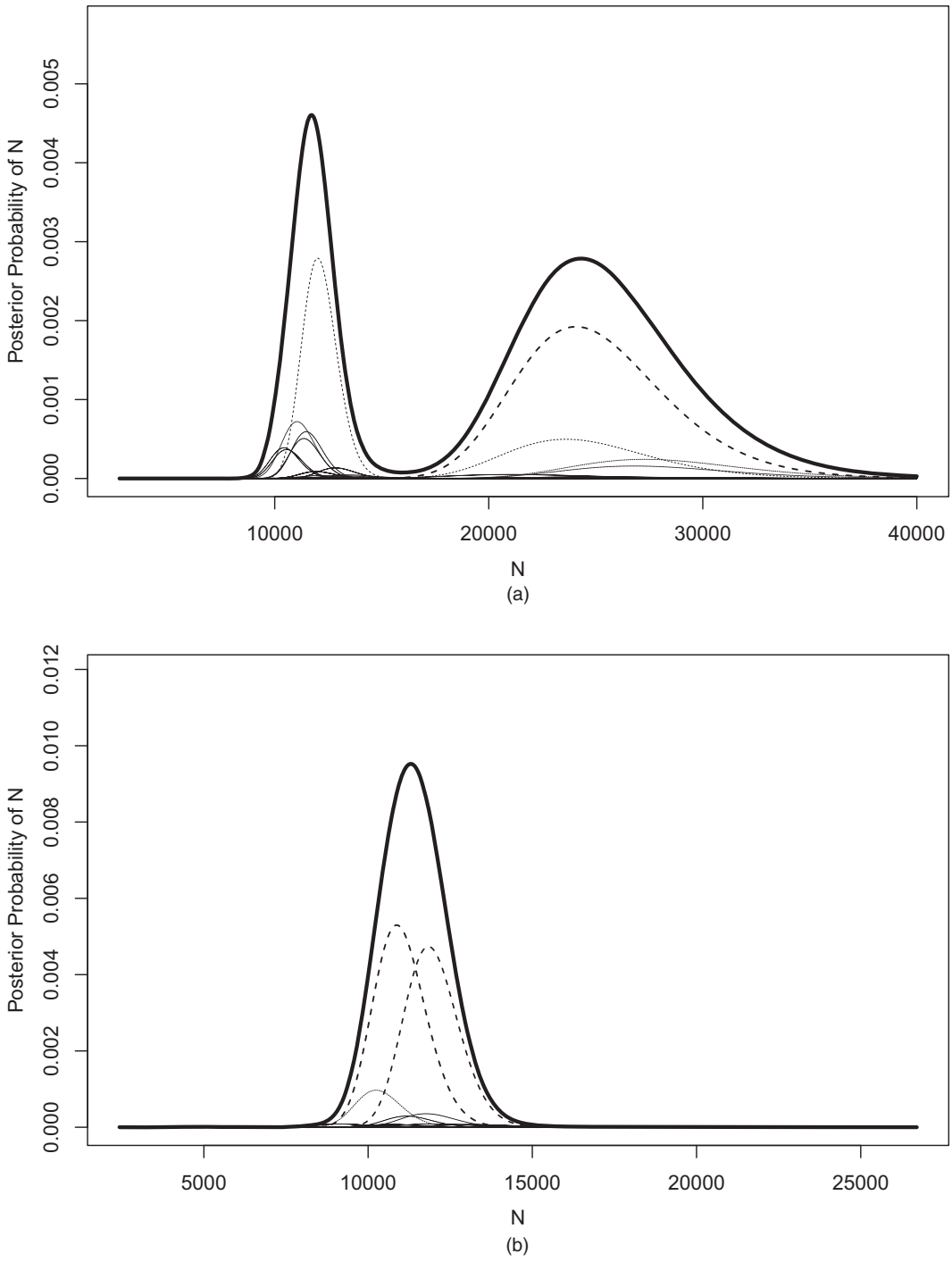

Fig. 5. Posterior distribution of the total population size for the UK data, for (a) the five-list and (b) four-list data, using the method of Madigan and York (1997): ———, averaged posterior probability; – – –, posterior probability by the model

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

5.2. Dirichlet process mixtures

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

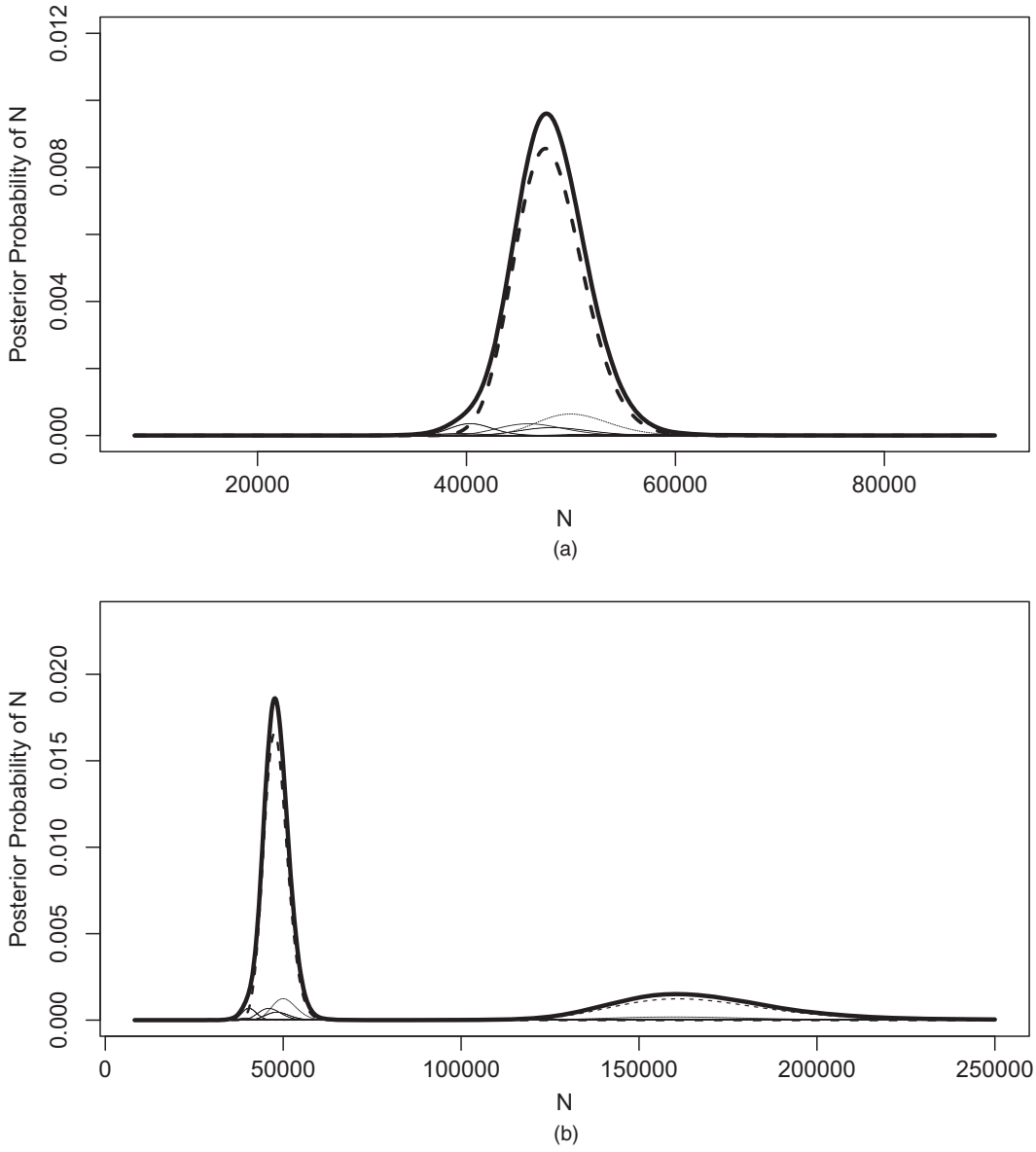

Fig. 6. Posterior distribution of the total population size for the Netherlands data, using the method of Madigan and York (———, averaged posterior probability; – – –, posterior probability by the model): (a) standard range of possible population size; (b) range extended to 250000

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.* (2019) for further discussion.

In the general MSE 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

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.

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.

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 is 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.* (2019). 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 Governments (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 (2019).

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**, 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.
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. (2019) Multiple systems estimation for sparse capture data: inferential challenges when there are non-overlapping lists. *Preprint. arXiv:1902.05156*.
Cockayne, J. (2015) Unshackling development: why we need a global partnership to end modern slavery. Freedom Fund.
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.
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. (2019) Measuring modern slavery: law, human rights and new forms of data. To be published.
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**, 22.
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*. (Available from `https://github.com/bernardsilverman/modslavmse`.)
Silverman, B. W. (2018b) Demonstrating risks is not the same as estimating prevalence. In *Proc.* Shibuya: *Delta 8.7 Modelling the Risk of Modern Slavery Symp.* United Nations University. (Available from `https://delta87.org/2018/12/demonstrating-risk-not-same-estimating-prevalence/`.)