Model fitting in Multiple Systems Analysis for the quantification of Modern Slavery: Classical and Bayesian approaches

Bernard W. Silverman
Rights Lab
University of Nottingham

July 16, 2018

Abstract

Multiple Systems Estimation is a key estimation approach for hidden populations such as the number of victims of Modern Slavery. The UK Government estimate of 10,000 to 13,000 victims was obtained by a multiple systems estimate based on six lists. A stepwise method was used to choose the terms in the model. Further investigation shows that a small proportion of models give rather different answers, and that other model fitting approaches may choose one of these. Three data sets collected in the Modern Slavery context, 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 approaches. 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 data sets in the Modern Slavery context. 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 datasets are freely and publicly available to facilitate wider implementation and further research.

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 strategic assessment of 2013. See Silverman (2014) and Bales et al. (2015). The method used for that estimation was Multiple Systems Estimation; cases that come to light are categorized onto a number of different lists, and the numbers 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. Multiple Systems Estimation is a development of the classical capture-recapture approach and has been used in many different contexts, such as counting casualties in armed conflicts (Manrique-Vallier et al., 2013) and numbers of injecting drug users (King et al., 2013).

Quantifying Modern Slavery has crucial importance for policy in this field. 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 goes on in this context to cite the use of multiple systems approaches as being 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.

Crucial to the Multiple Systems approach is the choice of the right model to fit to the observed data, in particular deciding which interactions or correlations to allow between the various lists. Some methods choose a particular model, while others seek a model averaging approach. This paper reviews a number of these and investigates in some depth their performance on a range of real data sets. Because of the particular interest in the use of the method in the Modern Slavery context, three of the data sets relate to Modern Slavery and human trafficking. The fourth is a data set also drawn from the human rights context, relating to deaths in the Kosovo conflict in 1999.

In most contexts of this kind, there is no ‘ground truth’ available to investigate the accuracy of any estimates. It is nevertheless 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 assess the performance of the various possible approaches specifically on data sets of the kinds 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 a detailed comparative study may be a contribution towards that.

Unfortunately all the existing methods display instabilities of various kinds when tested on the real data sets. In order to address this issue, we introduce a Bayesian/thresholding approach that places prior distributions on the individual terms in the standard model.

This paper is set out as follows. In Section 2 the various data sets are reviewed and tabulated. Section 3 begins with a discussion of the standard Poisson model which underlies the various possible approaches. It then goes on to examine two approaches to model selection, including the one used by Silverman (2014). In Section 4 the Bayes/thresholding method is introduced. This casts the problem in a form where a standard Markov Chain Monte Carlo package can be used to fit 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 model choice methods, but still requires care in its application. Two rather different Bayesian methods have been proposed and these are introduced and investigated in Section 5. Finally, some conclusions are drawn in Section 6.

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 described in this paper, and to reproduce its results, are given in Silverman (2018).

2 The data sets

The full data analysed by Silverman (2014), accumulated into six lists, are given in Table 1. The abbreviations of the lists are explained as follows:

LA	Local authorities
NG	Non-government organisations such as charities
PF	Police forces
GO	Government organisations, such as the Border Force and the Gangmasters and Labour Abuse Authority
GP	The general public, through various routes
NCA	The National Crime Agency

Some of the methods we shall consider do not deal with more than five lists, and so for some purposes we will combine the lists PF with the list NCA to construct the 'UK five list' data set. The NCA is not, strictly speaking, a police organisation, 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 list GP 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.

Another important analysis is that of van Dijk et al. (2017) and Cruyff et al. (2017), which considers the number of victims in the Netherlands. Table 3 of van Dijk et al. (2017) gives a table of six lists for identified victims in The Netherlands for the period 2010–2015. The data are given in Table 2 below. For a five-list version of these data, we combine the two smallest lists I and O. The total number of cases in this data set is 8234.

The third example is a data set constructed from data collected by eight agencies in New Orleans and analysed by Bales et al. (2018). These include 185 individuals who interacted with law and enforcement and service providers in the New Orleans areas during the year 2016. They are given in Table 3. Where it is necessary to reduce the number of lists, a five-list data set is constructed by combining the four smallest lists 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 killed in Kosovo in a three month period in 1999. They are available within the R package **LCMCR** (Manrique-Vallier, 2017) and are reproduced in Table 4. It can be immediately seen that 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.

3 Models and methods

In this section, we review the basic log-linear model and various approaches to model fitting and averaging, using the UK data set as an example to illustrate the discussion.

3.1 The underlying Poisson model

Suppose we have K lists altogether labelled $\{1, 2, \dots, K\}$. For each subset A of $\{1, 2, \dots, K\}$, 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. Let N_{tot} be the total number of cases including the $N_\emptyset$.

Table 1 only gives counts for 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 zero, yielding 63 observations in all. Of course, there is no observed count for the dark figure.

The basic model is to assume that each N_A has, independently, a Poisson distribution with parameter λ_A, with some structure on the λ_A. This is in itself 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. Clearly, especially if there are observed covariates to the various cases, the model will only be a jumping off point for more detailed modelling, but the simple Poisson model is at least a start. The model does not assume that the various lists are independent; the possibility of interactions between the lists is 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 too.

3.2 Log linear models, main effects and interactions

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

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

Table 1: Potential victims of trafficking in the UK, 2013. Numbers of cases on each possible combination of lists, leaving out combinations for which no cases were observed. 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 are omitted from the table, but are still taken into account in the analysis.

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

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. The lists are as follows: P = National Police; K = Border Police; I = Inspectorate SZW (Ministry of Social Affairs and Employment); R = regional coordinators; O = residential treatment centers and shelters; Z = others (for example, ambulatory care centers, organizations providing legal services, Immigration and Naturalization Service).

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

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. For reasons of confidentiality the lists are anonymised.

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

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

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

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. In some cases we also allow for higher-level interactions in addition as will be seen below.

Within this model, we have $\log \lambda_\emptyset = \mu$ because the two sums will be empty. Therefore the estimate of the dark figure is $\exp \mu$, meaning that 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 some model selection criterion to choose a particular set of parameters to fit, constraining all the others to zero. The other is to use some sort of model averaging approach, usually of a Bayesian nature.

3.3 Model selection approaches

The package **Rcapture** (Baillargeon and Rivest, 2007) can be used as the basis of various model selection approaches in this context. The simplest model selection approach is to set all interaction terms to zero, fitting main effects α_i only. Under this model, the lists themselves are independent, an assumption that may be unrealistic in many contexts. Nevertheless this model may be a good reference point and jumping-off point for more detailed analysis.

Table 5: Estimates and confidence intervals for the UK data, for the main effects model and for the stepwise AIC approach. The figures are for the numbers of thousands of victims, rounded to the nearest one hundred. The bold-face row corresponds to the analysis carried out by Silverman (2014)

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

3.3.1 Adding parameters stepwise

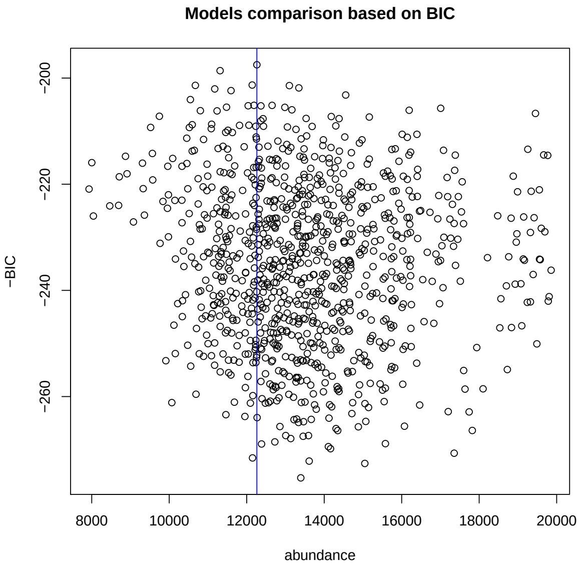
In their original work on the UK data, Silverman (2014) and Bales et al. (2015) used a stepwise approach, starting with main effects only and then adding two-list interactions β_{ij} stepwise. At each step, the interaction that best improves the AIC criterion is chosen. The process of adding interactions 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, 2018), and makes use of the package **Rcapture**.

Table 5 shows estimates and confidence limits using main effects only, and the stepwise procedure with two different p-value thresholds, for the UK data summarised into six, five and four lists as set out in Section 2. The original work used the stepwise method with $p = 5\%$ and it can be seen that both the six and five list data give a 95% confidence interval, conditional on the model choice, of 10K to 13K 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, on the other hand, give smaller estimates, but none of these effects is dramatic.

3.3.2 Choosing from a large class of models using an information criterion

The stepwise process is not the only possibility for choosing a model. Another possible approach is to fit all possible models, considering every possible sub-

Figure 1: Estimates of abundance plotted against BIC, with outliers omitted, the default plot option



set of the interactions to fit, and to choose between these using some criterion. This can be done using the routine `closedpMS.t` within the package **Rcapture**. However, 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, and so the method cannot be applied. To make the method computationally feasible in practice, the approach is only applied to the five-list data, allowing for two-list interactions only, leaving 2^{10} models to be considered. Some of the models yield a warning criterion when fitted but this will be ignored for the moment.

There are several methods for displaying the results, all based on using the BIC rather than AIC as the primary method of model choice. The BIC and AIC differ in the amount they correct for parsimony of models, with BIC having a heavier preference for more parsimonious models.

The plot shown in Figure 1 is the default plot of the results, with the vertical line on the plot showing the estimate with the highest BIC (-197.5), and estimate of 12262 for the population size, and the various estimates clustering approximately around the estimate in the original paper. However if we set the argument `omitOutliers = F` the figure becomes that shown in Figure 2. It can be seen that there is a subsidiary cloud of results corresponding to a much larger estimate for the population size, and indeed that the estimate for the highest BIC is actually within that cloud. Closer examination of the top ten models chosen by each of BIC and AIC is instructive. There are no models in the top

Figure 2: Estimates of abundance plotted against BIC, with outliers included

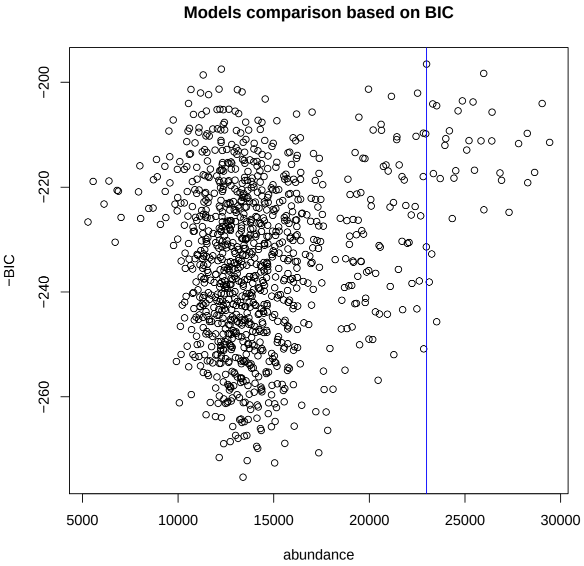


Figure 3: Estimates of abundance plotted against BIC, with list GP excluded

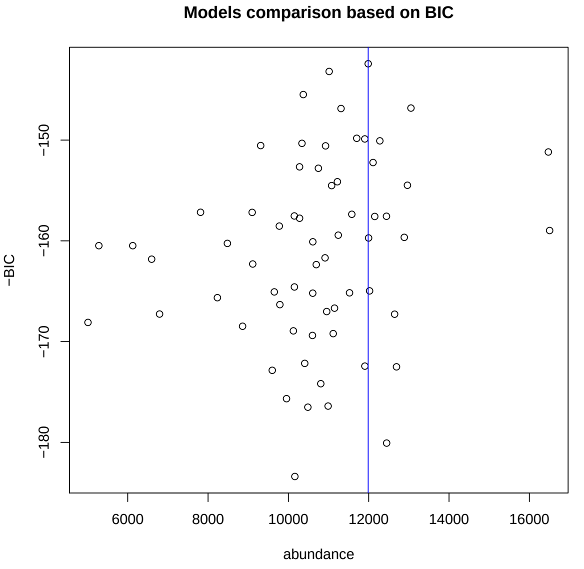


Table 6: Estimates and confidence intervals for the Netherlands, New Orleans and Kosovo data. The figures are for the numbers of thousands of victims, rounded to the nearest one hundred.

	Estimates and confidence limits				
Data	2.5%	10%	point est	90%	97.5%
Main effects only					
Netherlands	48.5	50.0	52.8	55.9	57.6
Netherlands five lists	48.6	50.0	52.9	56	57.8
New Orleans	0.7	0.7	1.0	1.4	1.7
New Orleans five 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 five lists	121.4	130.0	148.7	171.0	184.6
New Orleans	0.7	0.7	1.0	1.4	1.7
New Orleans five 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 five lists	121.4	130.0	148.7	171.0	184.6
New Orleans	1.4	1.8	3.4	7.2	∞
New Orleans five lists	0.7	0.8	1.0	1.4	1.8
Kosovo	12.5	13.1	14.3	15.7	16.5

ten AIC list which yield estimates over 17,000, and only one which yields an estimate much outside the ball park suggested in the original analysis. On the other hand, BIC chooses models yielding a much wider range of estimates.

The original analysis explained that the GP list (list 5 in this case) was unreliable when compared to the other lists, in that cases on it may not always have the kind of detail making it possible to match then to cases on the other lists. With this in mind the analysis was repeated for the four-list data. Figure 3 shows all models and demonstrates that the cloud of points corresponding to the much larger estimate disappears altogether if GP is omitted.

This discussion raises an issue which we will elaborate further in the remainder of the paper. 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.

3.3.3 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; com-

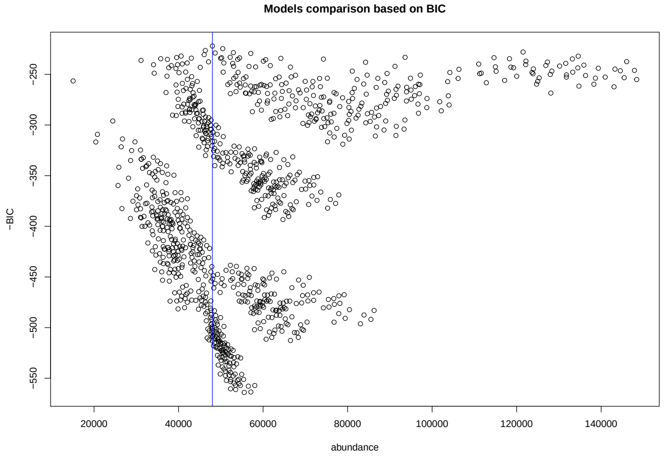


Figure 4: Estimates of abundance plotted against BIC, for the Netherlands data consolidated into five lists

bining the two smallest lists into one yields more than doubles the estimates. There is no such instability if main effects only are fitted. Some further intuition may be gained from Figure 4. There is a long tail of models with very large estimates. Although the globally optimal model according to BIC is not in this group, the stepwise method is choosing one of these.

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 the main effects only estimates, which are virtually unaffected by reducing to five lists.

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

4 A Bayesian approach to model choice

4.1 Defining the prior and thresholding the results

In this section, we set out a Bayesian 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 zero. In the second step, those interactions for which there is no strong evidence that they are not zero are dropped from the model, and the analysis repeated. The steps of the model are as follows.

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

    In every case the R package **MCMCpack** and in particular the function `MCMCpoisson` enable MCMC to be used to simulate from the posterior distribution. The improper uniform prior is the default for parameters within `MCMCpoisson`.

2. Constrain to zero 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} 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 to 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 only be included if there is very strong evidence that they are not zero. For this approach we use a threshold of 5, admittedly chosen rather arbitrarily.

4.2 Implementation issues

There are two implementation issues which are taken into account in the package **modslavmse** (Silverman, 2018). 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 while 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.

The second issue arises when using the default uniform improper prior ($\lambda = 0$) for the interaction parameters. It can be illustrated by considering the UK data as in Table 1. Examination of that table shows that no cases fall in both lists LA and GP, whether or not in combination with other lists. If the improper uniform prior is used for the corresponding interaction parameter $\beta_{LA,GP}$, then we show that the posterior distribution of $\beta_{LA,GP}$ is concentrated at $-\infty$ and set out the way that the other parameters can be estimated by MCMC.

In the general MSE model, suppose there is a pair of lists which contain no case in common. Renumbering the lists if necessary, suppose there are no cases in common between lists 1 and 2, so that $N_A = 0$ 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

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

It follows that

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

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

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

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 $\exp(-Ce^{\beta})$. Although this appears at first sight to be an improper distribution, this function has the properties that, for all y,

$$\int_{-\infty}^{y} \exp(-Ce^{\beta}) d\beta = \infty \text{ and } \int_{y}^{\infty} \exp(-Ce^{\beta}) d\beta < \infty$$

so that $P(\beta_{12} > y)/P(\beta_{12} \leq y) = 0)$. This corresponds to the distribution where $\beta_{12} = -\infty$ with probability 1. Since this is true conditional 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, while 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 conditional 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 6 lists and hence 63 observed observable combinations B, we omit the 16 N_B for which B included both lists LA and GP, 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 $\beta_{LA,NCA}$ is also estimated to be $-\infty$ with probability one. Removing (from the 47 remaining combinations) all combinations of lists containing both LA and NCA 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.

4.3 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 7, 8 and 9. The first line in each table shows the result of fitting the main effects only, with no β_{ij} considered. Once interactions are 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. If list GP is left out, there is a substantial effect if main effects only are considered or if a large threshold is used, but hardly any effect at moderate thresholding levels.

Computationally, the uniform prior (which is the default in `MCMCpack`) 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 12.3K for the overall prevalence, with an 80% credible interval (rounding to the nearest 500) of about 11.5K to 13.5K and a 95% credible interval, in round terms, of 11K to 14K. This is about a thousand more than the confidence interval obtained from the fixed model, but this is possibly because of the model averaging taking some note of the second group of models exemplified by the model chosen by the BIC criterion. 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 10. 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 5 of the 15 two-factor interactions are thresholded out. As a check, the method was run on the five-list version of the

Table 7: Quantiles of posterior distribution of the total population size, including both the observed data and the dark figure, UK data with six lists

		Quantiles of posterior				
Prior	Threshold	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 8: Quantiles of posterior distribution of the total population size, including both the observed data and the dark figure, UK data with five lists

		Quantiles of posterior				
Prior	Threshold	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 9: Quantiles of 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)

		Quantiles of posterior				
Prior	Threshold	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

data, with the two smallest lists consolidated. In most cases this gave similar results to the full data but for threshold 2 and variance 1 the difference was greater.

The New Orleans data are a smaller set of observations, and also consist of 8 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 was done in the original analysis (Bales et al., 2018). 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 8 list data are used, the MCMC algorithm requires both a long burn-in period and then a long run, and possibly other adjustments to the control parameters, to give reasonable mixing in the posterior realisations. For simplicity, therefore, we analyse the five list version of the data, and the results are given in Table 11. The variance 1 threshold 2 model (and indeed some of the other models) give results 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 interactions are thresholded out, but the model is fitted not just based on the main effects only but with three of the interactions included and estimated to $-\infty$. If the original 8 list data are considered, then the effect is much stronger, with 18 of

Table 10: Quantiles of posterior distribution of the total population size, including both the observed data and the dark figure, Netherlands data. The results for the five list data were, in general, slightly lower for thresholds 0 and 2 and slightly 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.

		Quantiles of posterior				
Prior	Threshold	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 11: Quantiles of posterior distribution of the total population size, including both the observed data and the dark figure, New Orleans data consolidated into five lists

		Quantiles of posterior				
Prior	Threshold	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

the 28 possible interaction parameters estimated as $-\infty$.

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

		Quantiles of posterior				
Prior	Threshold	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

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

4.4 Choosing the threshold for interactions

The Bayesian approach avoids the necessity of choosing a particular model, but it still contains tuneable parameters that have to be chosen in order to specify the prior. 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 13 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 GP. For the Netherlands data, 10 of the possible 15 interactions survive a threshold of 2, and the results 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, the advantage of the more parsimonious approach is that it focuses in on those pairs where there is a very clear interaction, giving pointers as to where to particularly look to gain a greater understanding of what is going on.

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. On the other hand, 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 inter-list correlations, and it is very reassuring that the high threshold adapts well to data of this kind.

Overall, consideration of these examples suggests that, with these particular parameters, the Bayesian model adapts reasonably well to the characteristics of different data sets, although it is advisable not to apply the method completely blindly.

Table 13: Interactions included in the variance 1, threshold 2 model. For threshold 5, only the effects shown in bold survive the thresholding step. For the four-list UK data, the effect LA:NG does survive up to thresholds of about 3.5.

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**

5 Other Bayesian approaches

5.1 Graphical models

A different Bayesian approach is via graphical models, developed by Madigan and York (1997) and implemented in the package **dga** (Johndrow et al., 2015). This uses every decomposable graph model of dependencies between the various lists, and obtains the joint posterior probabilities of the models and the total population size (the dark figure plus the number of cases actually observed). 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.

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

Table 14: Quantiles of posterior distribution using the method of Madigan and York. 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.

Data	Maximum population	Quantiles of posterior 2.5%	10%	50%	90%	97.5%
UK five lists	40	10.4	11.3	23.0	29.6	33.3
UK four lists (GP omitted)		9.5	10.1	11.4	12.7	13.5
Netherlands five lists		39.9	43.3	47.9	52.8	55.8
Netherlands five lists	250	44.7	50.3	166.7	204.3	228.6
New Orleans five lists		0.5	0.6	0.9	1.3	1.6
Kosovo		9.8	10.8	12.6	14.9	16.5

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 Figure 5. The dotted curves show the joint posterior probabilities of individual models and particular values for the total population size . Thus there are 822 dotted curves in the upper figure in Figure 5, and 61 curves in the lower figure; most of the models have posterior probability very close to zero for all values of the total population. A variety of line styles are used for the various models. The solid curve is the sum of the dotted curves, in other words the marginal posterior distribution of the total population size. The documentation of the package **dga** gives further details. The quantiles of the posterior distribution are given in Table 14, though of course in the case of the full five list data these are not an adequate description of the bimodal distribution.

Now turn to the Netherlands data, where the number of observed cases is 8234. The top panel of Figure 6 shows the posterior when calculated on the range of up to ten 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 second substantial mode appears. The quantiles for the two estimates are given in Table 14.

Results are also given in Table 14 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 considered, these two datasets 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, while 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.

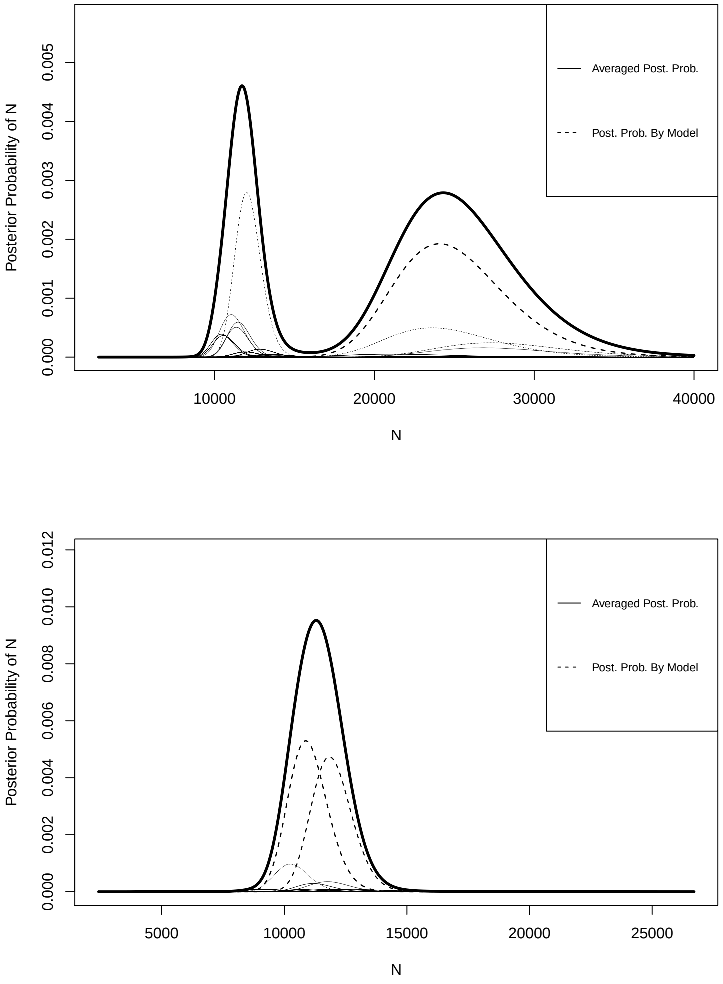


Figure 5: The posterior distribution of the total population size for the UK data, for the five list and four list data, using the method of Madigan and York

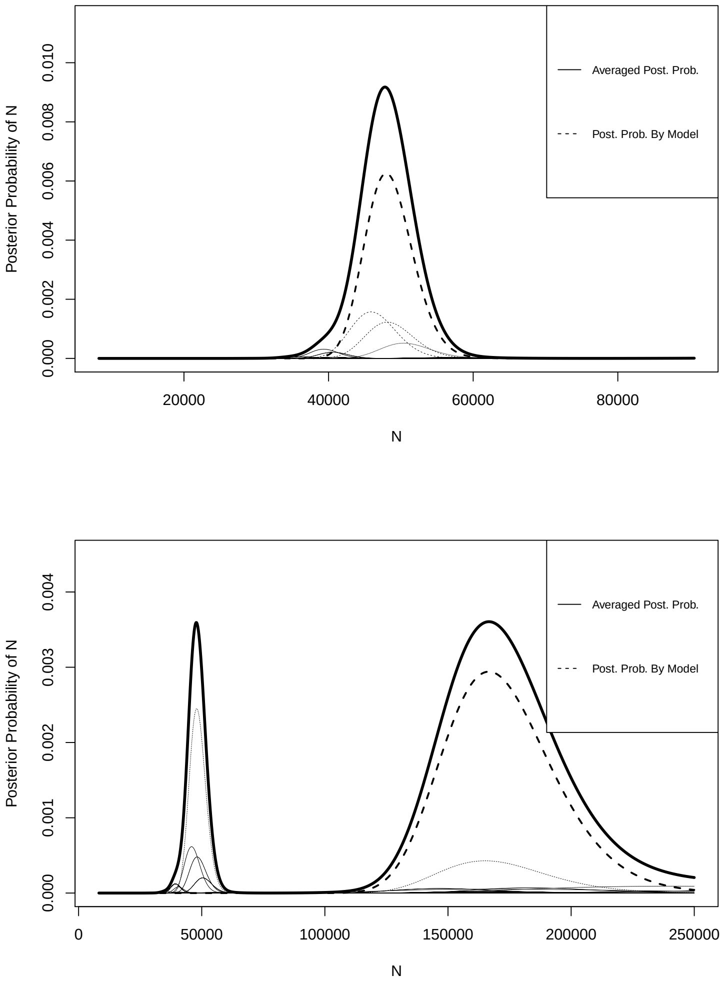


Figure 6: The posterior distribution of the total population size for the Netherlands data, using the method of Madigan and York. Upper: standard range of possible population size; lower: range extended to 250K.

Table 15: Quantiles of posterior distribution 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	43.2	45.0	49.7	55.2	58.7
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 a Markov Chain Monte Carlo estimate of the population size. In contrast with the method described in Section 5.1 above, there is no restriction on the number of lists. The results for the various data sets are shown in Table 15

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. In order 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 `burnin` value was set to 100K.

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 found by the Madigan-York method and displayed in the upper panel of Figure 5. Once GP is 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 case where the two smallest lists are combined.

6 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 it is important that there should be a standard agreed open-source methodology for the necessary modelling to produce point and interval estimates. The most stable method would, of course, be to ignore the possibility of interactions and simply fit main effects, but the

Kosovo example shows that this would be clearly inadequate in some practical cases.

The Bayesian approach of this paper, with a threshold of 5 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, it automatically rules out implausible secondary estimates which are almost certainly spurious, and it gives parsimonious explanations in cases where there may be a few interactions of particular interest. It adapts well between data such as the Kosovo data, with strong dependencies between lists, and those situations where few, if any, interactions are clearly present in the data.

The availability of real data in the public domain has been an important contribution to the study carried out in this paper, because it is likely that data in particular areas will have specific characteristics which need to be taken into account. For example, it is noteworthy that the Kosovo data, drawn from a different context, showed much more correlation between lists than did the other three data sets.

There are several avenues for future research on this methodology. 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, give us a deeper understanding of the problem itself?

7 Acknowledgements

The author gratefully acknowledges correspondence and other help from Kevin Bales, Patrick Ball, Peter van der Heijden, Ella Kaye and Daniel Manrique-Vallier.

References

Baillargeon, S. and L.-P. Rivest (2007). Rcapture: Loglinear models for capture-recapture in R. *Journal of Statistical Software, Articles 19*(5), 1–31.

Bales, K., L. Murphy, and B. W. Silverman (2018). How many trafficked and enslaved people are there in New Orleans? Unpublished report.

Bales, K. B., O. Hesketh, and B. W. Silverman (2015). Modern Slavery in the UK: How many victims? *Significance 12*(3), 16–21.

Ball, P., W. Betts, F. Scheuren, J. Dudukovich, and J. Asher (2002). *Killings and Refugee Flow in Kosovo March–June 1999.* American Association for the Advancement of Science. A Report to the International Criminal Tribunal for the Former Yugoslavia.

Cockayne, J. (2015). *Unshackling Development: Why we need a global partnership to end modern slavery.* Freedom Fund.

Cruyff, M., J. van Dijk, and P. G. M. van der Heijden (2017). The challenge of counting victims of human trafficking: Not on the record: A multiple systems estimation of the numbers of human trafficking victims in the Netherlands in 20102015 by year, age, gender, and type of exploitation. *Chance 30*(3), 41–49.

Johndrow, J., K. Lum, and P. Ball (2015). *dga: Capture-Recapture Estimation using Bayesian Model Averaging.* R package version 1.2.

Johnstone, I. M. and B. W. Silverman (2004, 08). Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Annals of Statistics 32*(4), 1594–1649.

King, R., S. M. Bird, A. M. Overstall, G. Hay, and S. J. Hutchinson (2013). Injecting drug users in Scotland, 2006: Number, demography, and opiate-related death-rates. *Addiction Research and Theory 21*, 235–246.

Madigan, D. and J. C. York (1997). Bayesian methods for estimation of the size of a closed population. *Biometrika 84*(1), 19–31.

Manrique-Vallier, D. (2016). Bayesian population size estimation using Dirichlet process mixtures. *Biometrics 72*(4), 1246–1254.

Manrique-Vallier, D. (2017). *LCMCR: Bayesian Non-Parametric Latent-Class Capture-Recapture.* R package version 0.4.3.

Manrique-Vallier, D., M. E. Price, and A. Gohdes (2013). Multiple systems estimation techniques for estimating casualties in armed conflicts. In T. Seybolt, B. Fischhoff, and J. Aronson (Eds.), *Counting Civilian Casualties. An Introduction to Recording and Estimating Nonmilitary Deaths in Conflict*, pp. 77–93. New York: Oxford University Press.

Silverman, B. W. (2014). *Modern Slavery: an application of Multiple Systems Estimation.* Home Office, UK. Published in conjunction with the UK Government Modern Slavery Strategy, available at `https://tinyurl.com/ydegfjaw`.

Silverman, B. W. (2018). *modslavmse: Multiple Systems Estimates for estimating the prevalence of Modern Slavery.* R package, available from `https://github.com/bernardsilverman/modslavmse`.

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