This is an Accepted Manuscript of an article published by Taylor & Francis in the Journal of Human Trafficking on 6th September, 2019, available online: https://doi.org/10.1080/23322705.2019.1660952

On the unreliability of Multiple Systems Estimation for estimating the number of potential victims of modern slavery in the UK

John Whitehead[a+], James Jackson[a], Alex Balch[b] and Brian Francis[a]*

Abstract

Accurate records of victims of modern slavery identified by various agencies allow investigators to compare different jurisdictions, track fluctuations in prevalence over time and evaluate preventative interventions. As well as enumerating those victims known to agencies, it would be desirable to know how many are working undetected under conditions of modern slavery and thus deduce the total number involved.

To estimate the number of undetected potential victims of modern slavery in the UK, Bales, Hesketh and Silverman (2015) applied the method of Multiple Systems Estimation. Their approach involves fitting a statistical model to data listing victims detected by different agencies. In doing so, (a) they assume that various terms in the model are equal to 1, and (b) they only include terms not assumed to be 1 if they achieve statistical significance. In this paper, simulated datasets with known properties are used to show that if (a) is valid then (b) leads to substantial overstatement of the reliability of the estimates computed, and that if (a) is not valid then the estimation procedure is totally unsound. We conclude that Multiple Systems Estimation is not a suitable method for estimating numbers of potential victims of modern slavery.

Keywords: *estimating numbers of potential victims of modern slavery; multiple systems estimation; potential victims of modern slavery*

[a]Department of Mathematics and Statistics, Lancaster University, Lancaster UK.
[b]Department of Politics, University of Liverpool, Liverpool, UK.
** Correspondence to John Whitehead, 6 Alderman Road, Lancaster, LA1 5FW, UK.*
[+]E-mail: j.whitehead@lancaster.ac.uk

On the unreliability of Multiple Systems Estimation for estimating the number of potential victims of modern slavery in the UK

Introduction

"The most robust estimate to date of the scale of modern slavery in the UK was produced by the Home Office in 2014. The estimate suggested that there were between 10 000 and 13 000 potential victims of modern slavery in the UK in 2013" (p. 4). This statement is made in the 2018 UK Annual Report on Modern Slavery, published in October of that year (Home Office, 2018). In the report, "potential victims of modern slavery" are taken to be people who have been identified and referred as possibly having worked under conditions of modern slavery, and this includes both those whose victim status is eventually confirmed and those for whom it is not. For brevity, such people will be referred to as "pvoms". In this paper we will show that the figures quoted above are unreliable, and that the method used to derive them is inappropriate.

The range (10 000, 13 000) quoted in the Home Office report is drawn from a series of analyses presented by Bales, Hesketh and Silverman (2015) (henceforth "BHS") who deduced these figures using a method known as Multiple Systems Estimation (MSE). MSE is a generalisation of the method of capture-recapture, used (for example) to estimate fish populations over a century ago (Goudie & Goudie, 2014; Manrique-Vallier, Price & Gohdes, 2013). In such an application, a sample of fish from a stable population would be caught, counted, marked, and thrown back into the water. Later, a second sample would be caught, and the numbers of marked and unmarked fish would be counted. These counts, plus certain modelling assumptions, would then allow the estimation of the total population of fish in the waters sampled. For application to the prevalence of modern slavery, multiple lists of pvoms detected by different agencies are required. These lists take the place of the two catches of fish in the example above, and using more than two lists allows the modelling assumptions to be relaxed in a way that will be discussed below (see Details of the MSE approach used by Bales, Hesketh and Silverman). BHS reported six different analyses, all based on MSE but using different sets of lists. Throughout this paper we will focus on just one of their analyses, based on what they refer to as "Model B". The analysis makes use of data from five lists of pvoms from 2013, compiled by government organisations (GO), the

general public (GP), local authorities (LA), non-governmental organisations (NG), and police forces and the National Crime Agency (PF). From this analysis, BHS derive an estimate of 11 313 and a 95% confidence interval of (9 889, 13 063) for the total number of pvoms in the UK in 2013. The other five analyses presented by BHS provide similar results: all estimates lie between 10 900 and 11 500 and all confidence intervals between 9 500 and 13 200. Model B has been chosen for study here because it is the case that BHS feature in Table 3 of their paper, and because that table provides estimates and standard errors which we have used to validate our own computer programs. The data used by BHS to derive their estimate based on Model B are shown in Table 1 below.

_____

Insert Table 1 here

_____

The BHS estimate has not only become a central reference point for the UK government's strategic thinking on modern slavery (Home Office, 2018), it is also widely used by scholars working in the area: see for example Craig (2017), and is seen as a potentially valuable source of data, even by scholars otherwise demanding an improvement in the quality of research in the field (Patterson & Zhuo, 2018). Following publication of the BHS paper, Durgana and Zador (2017) claimed that MSE "is generally considered to be the most reliable [method] and ideally suited for developed countries" (p. 51). Van Dijk and van der Heijden (2016) applied MSE to Dutch data from 2014, obtaining an estimate of 17 812 pvoms with 95% confidence interval (14 026, 23 874). Cruyff, van Dijk and van der Heijden (2017) extended the MSE modelling to fit additional factors concerning age, gender, type of exploitation and year of exploitation to Dutch data from the period 2010 to 2015. From their Figure 1, the new estimate for 2014 is 7 840 with 95% confidence interval (7 040, 8 450). It is of some concern that the two confidence intervals for 2014 given in these papers do not overlap, and indeed are quite distant from one another. At least one of them must be seriously misleading, possibly both.

In this paper, the assumptions inherent when applying the MSE method to modern slavery data will be critically examined. BHS use a two-stage approach, first using significance testing to select terms for inclusion in their model, and then fitting that model to compute estimates and confidence intervals for the total number of pvoms. The effect of the model selection stage on the accuracy of confidence intervals derived from the model

fitting stage will be explored. BHS also assume that certain model terms are equal to 1, without explanation of what these terms mean or justification of why they should be set to 1. The effects on the BHS approach should these terms not be 1 are explored, and efforts are made to include them in the model. Our findings are rather disturbing. Both the validity of confidence intervals and the unbiasedness of estimates are called into question. As a result, we are unable to recommend the use of the MSE approach in this context. There do not appear to be any quick fixes to the problems that we have discovered, and indeed it may prove infeasible to derive reliable estimates of or confidence intervals for the total number of pvoms from available data.

Below, the MSE method, as implemented by BHS, will be presented. This is followed by discussion of other applications of two-stage statistical procedures and the presentation of results from simulation studies of the accuracy of the procedure in the context of modern slavery. Next, certain simplifying assumptions made by BHS are examined, and the results of a sensitivity analysis in which the assumptions are weakened are presented, followed by examination of whether recently suggested alternative methods of analysis based on a Bayesian approach could resolve the difficulties uncovered. Finally, we present conclusions for future attempts to quantify the prevalence of modern slavery.

This paper is intended for people involved in the study, control and prevention of modern slavery. However, in order to appreciate the flaws in the MSE approach as applied by BHS, it is necessary to understand some of the key statistical issues, and these are presented in some detail. Practitioners may not be equipped to construct and fit statistical models themselves, but they do need to consider what assumptions underlie proposed statistical methods and how their accuracy has been investigated. In the present case, lack of such consideration has contributed to the adoption of an inappropriate method.

Details of the MSE Approach Used by Bales, Hesketh and Silverman

From the five lists of pvoms mentioned above, 31 counts are derived. There are the counts of pvoms who appear *only* on the GO list, *only* on the GP list, and so on – giving 5 counts. Then there are those on both the GO and GP list and no others, those on both the GO and LA list and no others, and so on for every pair of lists – giving 10 counts. There are those on all lists *except for* GO and GP, those on all lists *except for* GO and LA, and so on – giving 10 counts; those on all but one list – giving 5 counts; and those on every list – giving 1

count. The count of the number of pvoms on none of the five lists, melodramatically referred to by BHS as "the dark number", is the quantity to be estimated from the data. This can then be added to the total of the other 31 counts, which is 2,744 for the BHS data under Model B, to provide an estimate of the total number of pvoms. The counts used by BHS and in this paper to derive estimates based on Model B are those in Table 1.

In the MSE approach, the 31 observed counts are modelled as independent Poisson random variables. A Poisson random variable is a quantity taking one of the values 0, 1, 2, ... , governed by a specific probability distribution function. Details can be found in Hilbe (2014) or online, but two important characteristics should be noted. First, the variance of the distribution is the same as the mean, and indeed all properties of the distribution can be derived from the value of the mean. Second, the sum of two independent Poisson random variables also follows a Poisson distribution. If all of the pvoms living in the UK fell into the 32 classifications mentioned above (including those on no lists) independently of one another, then the counts would follow a Poisson distribution. If any of the classifications were merged, then some counts would be added together, but they would still satisfy the Poisson model. The question of independence will be returned to in the conclusion.

In this paragraph, Model B of BHS will be defined. The mean of the count of pvoms who appear *only* on the GO list can be denoted by $\kappa \times \lambda_{GO}$, the mean of the count of pvoms who appear *only* on the GP list can be denoted by $\kappa \times \lambda_{GP}$, and so on. This introduces one "constant" parameter κ (kappa), and 5 "main effect" parameters denoted by λ (lambda) with a subscript referring to one of the lists. The mean of the count of those on both the GO and GP list and no others can be denoted by $\kappa \times \lambda_{GO} \times \lambda_{GP} \times \mu_{GO,GP}$, and for those on both the GO and LA list and no others by $\kappa \times \lambda_{GO} \times \lambda_{LA} \times \mu_{GO,LA}$, and so on. This introduces 10 "two-way interaction" parameters denoted by μ (mu) with a subscript referring to two of the lists. No further terms are included: the means for all other counts are made up of κ and the λs and μs already introduced. So, for example, the mean of the count of pvoms on all lists *except for* GO and GP is given by

$$\kappa \times \lambda_{LA} \times \lambda_{NG} \times \lambda_{PF} \times \mu_{LA,NG} \times \mu_{LA,PF} \times \mu_{NG,PF} ,$$

and the mean of the count of pvoms on all lists *except for* GO is given by

$$\kappa \times \lambda_{GP} \times \lambda_{LA} \times \lambda_{NG} \times \lambda_{PF} \times \mu_{GP,LA} \times \mu_{GP,NG} \times \mu_{GP,PF} \times \mu_{LA,NG} \times \mu_{LA,PF} \times \mu_{NG,PF} .$$

The unobserved dark number is assumed to be a Poisson random variable with mean κ. In fact, the analyses estimate κ and set 95% confidence limits for κ. According to the model, κ is the mean of the dark number. This justifies presenting an estimate of κ as an estimate of the dark number itself. The 95% confidence interval does not take account of the variability of the dark number about its mean κ, but in this application that additional variability is small relative to the width of the interval, and it is reasonable to ignore it (see also the last paragraph of Section 3.1 of Silverman (2019)). This formulation is known as a log-linear model, as it can be written on the log transformed scale as a linear sum of terms. It can be fitted using any statistical analysis package that caters for such an approach, for example PROC GENMOD of SAS or the glm function of R.

The 31 counts that feature in Model B are characterised using 16 parameters. This implies that some assumptions have been made concerning the data. First consider an assumption that has *not* been made. Including the two-way interaction parameters (the μs) means that appearances on any two lists are not assumed to be independent. For example, if the two-way interaction between the GO and GP lists were ignored, and $\mu_{GO,GP}$ set to the value 1 in the model, it would imply that the probability that someone was on the GO list, given that they were on the GP list would be equal to the probability that they were on the GO list, given that they were not on the GP list. Symbolically, that assumption is

$$P(\text{on GO} \mid \text{on GP}) = P(\text{on GO} \mid \text{off GP}). \qquad (1)$$

There are many reasons why assumption (1) might not be true for any pair of lists. For example careless traffickers and their victims might be detected by multiple agencies, while competent operators might be observed by none. Assumption (1) is inherent in simple capture-recapture estimation based on two samples (for example of fish), but fortunately in MSE with more than two lists, it can be avoided.

The assumption that has been made is to neglect three-way and higher interaction parameters. If three-way interaction parameters were to be included, then the mean of the count of pvoms on all lists *except for* GO and GP would be given by

$$\kappa \times \lambda_{LA} \times \lambda_{NG} \times \lambda_{PF} \times \mu_{LA,NG} \times \mu_{LA,PF} \times \mu_{NG,PF} \times \nu_{LA,NG,PF}\,, \qquad (2)$$

and so on. Ten ν (nu) parameters would be introduced, raising the total number of parameters to 26. The simplifying assumption that all of the ν parameters are equal to 1 cannot be reliably tested from the data as there is insufficient power to do so. The assumption is also difficult to understand or to justify intuitively. For example, the three-way interaction between the GO, GP and LA lists can be ignored if

$$\begin{aligned} &P(\text{on GO} \mid \text{off GP, off LA}) \times P(\text{on GO} \mid \text{on GP, on LA}) \\ &\quad = P(\text{on GO} \mid \text{on GP, off LA}) \times P(\text{on GO} \mid \text{off GP, on LA}). \end{aligned} \qquad (3)$$

Here, the first term is the probability that a pvoms appears on the GO list, given that they are not on the GP list and that they are not on the LA list either. The meanings of the other terms are similar. The truth of (3) is assumed by BHS without justification or even discussion.

In principle, 5 four-way interaction parameters and 1 five-way interaction parameter could also be introduced. That would give a total of 32 parameters, which could not be estimated from 31 counts. Some form of assumption is necessary for any estimation to proceed.

BHS do not include three-way or higher interaction parameters in any of their models: in effect they set these terms equal to 1. Neither do they automatically include all two-way interactions. Instead, they use a method of forward stepwise selection in order to determine which μ parameters to include. This leads to a simpler model, and a tighter confidence interval around the estimate of the number of pvoms. First, they fit a model including κ and all five λ parameters. Then they add each μ parameter to the model in turn to find out which would make the biggest improvement in the fit (judging by Akaike's Information Criterion). If the estimate of that parameter is significantly different from 1

(according to a Wald test) then that parameter is added to the model. The process is then repeated, with each remaining μ parameter being considered for addition to the model, until it is found that the one which would make the biggest improvement to the fit is not significantly different from 1. The process is then stopped, with the final non-significant parameter not being added to the model. Silverman (2019) clarifies that the significance tests were conducted at the 5% significance level against a two-sided alternative. Having selected which parameters to include in the model, BHS then estimate the value of κ using maximum likelihood and provide a 95% likelihood-based confidence interval. Results for the total number of pvoms follow by addition of the 2744 observed victims. For Model B, all two-way interaction parameters *except for* $\mu_{GO,LA}$, $\mu_{GO,PF}$, $\mu_{GP,LA}$, and $\mu_{NG,PF}$ were selected for inclusion in the model. Thus, the relationship illustrated in (1) is assumed to be true for these four pairs of lists. In particular, it is assumed for GP and LA even though no pvoms were on both of these lists, an observation that might indicate a negative correlation between them rather than independence.

For the purposes of this paper, the fitting method described above was implemented using the R function **MSEfit** from the package modslavmse (Silverman, 2018), which in turn uses the function **closedpCI.t** from the package **Rcapture** (Baillargeon & Rivest, 2007). The output provides an estimate of 11 313 for the total number of pvoms, with a standard error of 803, based on a Poisson profile likelihood, and an estimate of 11 304 and a 95% confidence interval of (9 889, 13 063) based on a multinomial profile likelihood. The Poisson and multinomial methods are approximately equivalent, and the resulting estimates are similar. BHS quote the Poisson estimate and the multinomial confidence interval: they do not explain why they made these choices.

Pre-testing Invalidates Confidence Intervals

The method of BHS incorporates pre-testing. That is, significance tests are used to determine which two-way interactions to include before the final model is fitted and a 95% confidence interval is deduced. The calculation of the interval takes no account of the pre-testing that has taken place. This invalidates its defining feature: that in repeated applications the interval would contain the true value in 95% of cases.

Pre-testing used to be a common feature of statistical methods, and was widely taught and recommended in statistical courses. For example, it was often applied when

comparing two samples assumed to have been independently drawn from normal distributions. First, one tested whether the standard deviations of the two samples were the same, using an F-test. If so, a t-test was then used to compare the two means; if not, a more robust testing procedure was adopted. Moser and Stevens (1992) outline the reasons why this approach is invalid and no longer recommended, showing simulation studies that indicate that the true rate of wrongly declaring the two sample means to be different, when in fact they are the same, exceeds the intended 5%. Another well-known instance of pre-testing occurred in the analysis of cross-over clinical trials. In such studies, half of the patients receive Treatment A during one time period, and then later receive Treatment B for another time period. The remaining patients receive the two treatments in the opposite order. All of the data can be used to compare the effects of the two treatments, provided that the effects of that used in the first period do not *carry-over* and affect patient responses during the second period. In the past, statisticians would use the data collected to test whether there was significant evidence of a carry-over effect. If not, they would use all of the data to compare the treatments. If there was evidence of carry over, they would use only the results from the first treatment period to perform a less powerful analysis. Freeman (1989) reported simulation results that demonstrated that the two-stage procedure led to rates of false conclusions of treatment differences, when in truth there were none, well in excess of the intended 5%.

The consequence of studies such as those of Moser and Stevens and of Freeman is that pre-testing is rarely used these days in the analysis of clinical trial data, nor in many other application areas. Instead, the assumptions that previously would have been pre-tested are assessed for their credibility separately from the study data. Sometimes the study design is altered to enhance adherence to the assumptions – in cross-over trials a long wash-out period without treatment might be included between the treatment periods. If pre-testing is used, its influence is allowed for when tests are conducted and confidence intervals are computed.

The problem of computing confidence intervals in MSE analyses that incorporate model selection is discussed by the International Working Group for Disease Monitoring and Forecasting (IWGDMF) (1995). They point out that without allowance for model selection, intervals will be too short, implying that they will have a coverage rate that is less than 95%. They provide a discussion of the use of bootstrap methodology to derive valid confidence

intervals, but no examples are given and some serious difficulties with application of the bootstrap method are identified. BHS employ pre-testing, but they do not appear to have assessed its influence through simulation nor allowed for it in calculating confidence intervals.

To appreciate the potential magnitude of the pre-testing effect, an MSE model was fitted to the data of Model B in which all 10 two-way interactions were included without question, although no three-way or higher interactions were included. To implement this, the function **closedpCI.t** from the package **Rcapture** was used directly. The resulting estimate of the total number of pvoms is 11 233 and the 95% confidence interval is (6 155, 20 761). Following BHS, the Poisson version of the estimate is quoted, together with the multinomial confidence interval. The estimate is very similar to that of BHS, but the confidence interval is much wider. Exclusion of four of the two-way interaction terms by BHS has led to greater apparent precision: but is that greater precision spurious?

Before investigating this question, a technical feature of the model which includes all two-way interactions must be discussed, as it includes the parameter $\mu_{GP,LA}$, even though no pvoms were found to be on both the GP and the LA lists. The number of pvoms on the GP and LA lists is the sum of 8 of the basic 31 counts being modelled. That is, those on GP and LA only, plus those on GP and LA and one of the three others, plus those on GP and LA and all but one of the three others, and those on all five lists. Each of these counts is observed to be zero, and in the MSE model, each has a mean that includes the factors κ, λ_{GP}, λ_{LA} and $\mu_{GP,LA}$. The first three of these parameters appear in the expressions for the means of other, non-zero counts, but $\mu_{GP,LA}$ does not appear anywhere else in the model. Consequently, $\mu_{GP,LA}$ is estimated to be 0. The fitting method does not give a standard error for this estimate, as the approach used in all other cases breaks down when the estimate takes the value zero. Nevertheless, 0 is a valid estimate for $\mu_{GP,LA}$, and the estimate and confidence interval for the dark number, and those for the total number, of pvoms remain valid. One way to be reassured on this point is to replace the 0 count of pvoms on LA and GP only by a count of ½. This is, of course, not possible in practice. However, the fitting routine still works, and it leads to an estimate of 11 176 and 95% confidence interval of (6 133, 20 634): very close to the values given when the true count of 0 is used but with no zero parameter estimates. The three analyses reported so far are summarised in Table 2.

_____

Insert Table 2 here

_____

Some Simulation Results

Now, the results of a simulation study will be presented. Morris et al. (2018) discuss the principles of using simulation to evaluate the accuracy of approximate theoretical statistical results in practice. Our study was based on the parameter estimates taken from the model just described: that is one deduced from the data modified to include a count of ½ for pvoms on the GP and LA lists only. These parameter values were used to compute the "true means" of the 31 observable counts. A data set was then generated by simulation, in which each of the 31 counts was independently simulated as a Poisson random variable with the appropriate true mean. Thus, the simulated data set is an example of what another modern slavery study might yield if the model was correct and those true means were indeed governing the data. It is not the same as the data observed, but it resembles those data. No value for the dark figure was simulated, as this is (obviously) not used in the analysis procedures. This process was independently repeated, always using the same true means, until 10,000 replicate data sets had been collected. Each of these data sets will be different (unless by very small chance two are exactly the same!). The adjusted data set with the added count of ½ was used to set the truth, because if the model fitted to original data set had been used, no individual would ever be on both the GP and the LA list in any of the 10,000 replicate data sets. That adjusted data set led to an estimate of 11 176 for the total number of pvoms, and subtracting 2,744 which was the number of pvoms actually observed leads to an estimate of 8 432 for the dark number. In the simulations, the true value of κ is set to be 8 432.

The MSE model including all two-way interaction parameters, but no three-way or higher interactions, was fitted to each of the 10,000 simulated data sets in exactly the same way as it was to the real data above (and without replacing zero counts by ½). The average of the 10,000 estimates for the dark number was found to be 8 807. Of the 10,000 confidence intervals for the dark number, 94.4% contained the true value 8 432. Thus the procedure appears to overestimate the dark number by a moderate amount, but to set a confidence interval that contains the truth in close to the intended 95% of cases. The method is based on several approximations that are exact in samples of infinite size, but are

here being applied to just 31 observations: the accuracy found is quite reassuring. For the total number of pvoms, the observed number would be added to the estimate of the dark number and to the lower and upper confidence limits. In each of the 10,000 replicate runs, that observed number would be different. However, as the true total would be the true dark number plus the replicate-specific observed number, the same conclusions about bias and confidence interval accuracy apply.

Next, the MSE model was fitted to each of the 10,000 data sets using the forward selection method adopted by BHS and implemented as described above (refer to "Details of the MSE Approach Used by Bales, Hesketh and Silverman"). The average of the 10,000 estimates for the dark number was found to be 12 193. As the true value underlying the datasets was 8 432, this demonstrates considerable bias. Of the 10,000 confidence intervals for the dark number, only 39.1% contained that true value. This procedure overestimates considerably more than when all two-way interactions are fitted automatically, and the 95% confidence intervals are too narrow and fail to fulfil their defining function. The latter exhibit only 39.1% confidence. Results of the two simulation studies are presented in Table 3.

_____

Insert Table 3 here

_____

Inclusion of Three-way Interactions

It is feasible to include three-way interactions in the model. These are the parameters denoted by ν in expressions such as equation (2) above. The resulting estimates tend to have larger standard errors than those corresponding to two-way interactions, and the R function **closedpCI.t** for the calculation of likelihood-based confidence intervals often fails. Because of this tendency to fail, an alternative form of confidence interval that is more robust in these circumstances will be used throughout this section. This is based on the mean and standard error of $\exp(\kappa)$, and is known as a Wald confidence interval. An attempt to fit all 10 two-way interactions and all 10 three-way interactions to the data of Table 1 leads to an estimate of the dark number equal to 0, with an associated 95% Wald confidence interval that is essentially $(0, \infty)$. Replacing the zero counts in all but the last six rows in Table 1 by 0.1 does lead to interpretable results: the point estimate of the dark

number of 648 and the 95% confidence interval is (7, 61 629). That is a very wide confidence interval. Adding the 2 744 observed pvoms to each number gives a point estimate of the total of 3 392 with 95% confidence interval (2 751, 64 373). The value 0.1 is inserted, rather than ½, so that the total of the introduced counts remains small: there is no need to replace the zero counts amongst the last six rows of Table 1 as no four- or five-way interactions are being fitted.

If forward selection is used to choose interaction terms to include in the model, with both the μ and ν parameters under consideration, then no three-way interactions are selected. The final model is that fitted by BHS. An alternative procedure is known as backward elimination (Draper & Smith, 1998, pp. 339-341). In this case, we start with the model in which all two-way and three-way interactions are included. Then each of the 20 interaction terms are excluded in turn. The one that leads to the least reduction in goodness-of-fit, as judged by the Akaike Information Criterion (AIC), is identified. If dropping this term improves the AIC, then it is dropped. The AIC is constructed to play off the goodness-of-fit of a model against its complexity in terms of the number of parameters included. So we drop a parameter if the complexity resulting from retaining it is not justified by the improvement of the fit that it brings. An exception is that a two-way interaction cannot be dropped if a corresponding three-way interaction is in the model: for example $\mu_{GO,LA}$, $\mu_{GO,NG}$ and $\mu_{LA,NG}$ cannot be dropped if $\nu_{GO,LA,NG}$ remains in the model. The procedure continues until no interaction parameters can be dropped according to this criterion. Backward elimination was conducted using the R function **stepAIC** from the **MASS** package (Ripley et al., 2013). For the data of Model B, the resulting model includes all of the two-way interaction parameters except for $\mu_{LA,PF}$ and the three-way interaction terms $\nu_{GO,LA,NG}$ and $\nu_{GO,PF,NG}$. The estimated total number of pvoms is 5 552 with 95% confidence interval (4 407, 7 485). The latter does not overlap with that presented for Model B by BHS. As backward elimination is as well established as forward selection for choosing model parameters, there is no real basis for preferring one of these analyses over the other. Hence, it is of concern that they lead to confidence intervals that are mutually exclusive. The analyses reported above are summarised in Table 4.

Insert Table 4 here

Further Simulation Results

A similar simulation exercise to that reported above for models without three-way interactions was conducted. This time, the "truth" was based on the fitted parameters arising from fitting all two-way and all three-way interactions in the MSE, replacing zero counts by 0.1 as described above. Each of the 10,000 samples generated was then analysed (a) by fitting all 20 two-way and three-way interactions (using 0.1 counts where necessary), (b) by fitting all 10 two-way interactions and no three-way terms, (c) using the BHS forward selection approach in which only the two-way interaction parameters were considered for inclusion and (d) using the backward elimination approach starting with all 20 interaction terms. The results from Method (a) demonstrate that this is an unreliable approach, even when certain zero counts are replaced by 0.1. Many estimates for the dark number are essentially infinite, and many confidence intervals are essentially $(0, \infty)$. The latter are uninformative, but they do contain the true value and thus enhance the coverage probability. Estimates of the dark number for Methods (b) and (c) were 9 664 and 12 782 respectively and that for Method (d) was essentially infinite. Estimated coverage probabilities for the 95% confidence intervals produced by Methods (b), (c) and (d) were 0.0201, 0.0000 and 0.4375. The true value of κ in this investigation was 648, as reported in the first paragraph of this section, so the bias in estimation is enormous. The results from Methods (b) and (c) show that, if three-way interactions are indeed present but are ignored in the analysis, then the analysis is invalid. The results from Methods (a) and (d) show that attempts to allow for three-way interactions fail, whether one fits them all or looks for those that are most important. The second set of simulation results is summarised in Table 5.

_____

Insert Table 5 here

_____

Bayesian Approaches

Silverman (2019) describes some Bayesian analyses of these and similar data. The Bayesian approach is fundamentally different from the methodology explored so far in this paper (which is referred to as the "frequentist approach" when distinguishing it from the Bayesian method). It requires investigators to express their beliefs about the values of

study parameters before any data are collected: these are called “prior opinions”. Once the data have been collected, they are combined with the prior opinions to form the new impressions about parameter values, known as “posterior opinions”. Sometimes, prior opinions are carefully elicited from investigators during pre-study meetings (see for example, Johnson et al., 2010), but Silverman (2019) takes another approach in which the prior opinions are specified for the investigators by the study statisticians, or else are to be selected from a narrow range of possibilities. It should be borne in mind that, unless policy makers shared the prior opinions used in the analysis before learning of the study results, there is no reason why they should automatically accept the derived posterior opinions. In particular, if a prior opinion specifies that a certain parameter is definitely equal to 1, then no amount of evidence to the contrary will reverse that judgment, and the posterior opinion will also insist that the parameter takes the value 1.

Despite the deep differences between them, Bayesian and frequentist methods often lead to similar numerical results. This can be seen by comparing Table 2 above with Silverman’s Table 8. In Silverman’s table, the middle values in each row (labelled 50%) provide alternative estimates of the number of pvoms. The first and last values in each row (labelled 2.5% and 97.5%) form corresponding 95% *credibility* intervals for the total number of pvoms. The Bayesian estimate and credibility interval are defined in a different way from frequentist estimates and confidence intervals, but they are used in a similar way by practical investigators to assess the likely value of a parameter and how accurate that estimate is. The top row of Silverman’s Table 8 (with threshold = 0 and labelled “uniform”) provides values that are similar to those in the second row of Table 2 here: his results are estimate = 11 100, interval = (5 900, 22 200). The top row of the second block of results (with threshold = 2 and labelled “uniform”) provides values that are similar to those in the first row of Table 2 here: Silverman’s results are estimate = 12 200, interval = (10 700, 13 900).

Essentially, setting a threshold of 0 amounts to fitting all two-way interactions, while putting the threshold equal to 2 requires the data to exhibit evidence that an interaction parameter is needed before including it in the model. The uniform case corresponds to prior opinion expressing very little knowledge about the magnitude of two-way interactions, whereas the other cases include more informed prior opinion. In the Bayesian method, the 95% credibility interval is not an interval that will include the true parameter value in 95% of

replicate data sets, and thus its accuracy cannot be assessed through simulation as performed here.

The Bayesian analyses adopt similar ploys to the frequentist approach of BHS, such as pre-assessing parameters before including them in the model by setting a threshold equal to 2 or 5, and ignoring three-way interactions. The Bayesian approach described by Silverman (2019) does not involve eliciting from investigators what they believe about the likely magnitude of two- or three-way interactions: instead it allows (for two-way) or insists (for three-way) that these magnitudes are precisely 1. The former amounts to believing that it is *possible* that two lists are *totally* independent of one another, as in equation (1), and the latter to believing that three-way relationships such as equation (3) are *certain* to be true. As mentioned above, the latter judgment cannot be overturned by the data.

If the argument of this paper is accepted, and the frequentist justification of estimating the number of pvoms as around 11 500 pvoms with a 95% interval of (10 000, 13 000) is discarded, then some very strong justification is needed before accepting (roughly) those same figures as the valid outcome of a Bayesian analysis.

Conclusions

In this paper, two key difficulties inherent in estimating the number of pvoms using MSE have been exposed. First, even if there truly are no three-way interactions, then using a two-stage procedure in which selection procedures are used to identify which two-way interactions to include and then basing the analysis on that chosen model, leads to confidence intervals that are far too narrow and whose coverage probability can be closer to 0.40 than to the intended 0.95. That difficulty could be overcome by fitting all two-way interactions without any pre-analysis. However, there is no a priori reason to suppose that three-way interactions do not exist. If they do, then they have the potential to invalidate totally analysis methods that rely on their absence. Furthermore, the data are insufficient to either allow for them without question, or to search for and allow for those which can be detected as important.

There are more potential problems with MSE besides the two mentioned above. First, the assumption that the 31 observed counts follow the Poisson distribution is questionable. Such a model arises naturally when counting events that occur completely independently of one another. However, some trafficked people travel in family or

friendship groups, and if one is discovered then it is likely that all will be. Further simulations could be conducted in which the counts were generated from such a mechanism, and the inaccuracies introduced to analyses that ignored these dependencies could then be quantified.

There are also practical problems concerning compilation of the lists, in particular whether all refer to precisely the same time frame. Were MSE to be routinely and repeatedly applied, agencies would learn that their lists were inconsistent with those of other agencies. It would become natural to collaborate, but in doing so the basis of the method would be undermined. Problems can also arise from inconsistency in the interpretation of what constitutes modern slavery (previously referred to in the UK as human trafficking) among the relevant authorities, as this could impact upon practices of identification and support (ATMG, 2010). In addition, victims do not always wish to be recorded as such. Research into the UK's National Referral Mechanism, the system of support offered to all potential victims that have been identified, has noted that many choose not to be registered - referral is optional and they may have reason not to trust the authorities (Beddoe et al., 2015).

Important mathematical aspects of fitting log-linear models to count data such as those in Table 1 are descibed by Fienberg and Rinaldo (2012). They point out that if certain patterns of zero counts occur within the data, then estimation methods may not only be unreliable, but might break down altogether. In particular, maximum likelihood estimates may not exist and Bayesian estimates may not make appropriate use of the data. These issues are relevant to this paper because maximum likelihood estimates are employed within the MSE method as discussed above, and Bayesian methods have also been considered. A discussion of the general impact of such difficulties in the context of MSE models and software to check for the existence of maximum likelihood estimates has been provided by Chan, Silverman and Vincent (2019a and 2019b). Results presented In Section 3.3.4 of Silverman (2019), and our own investigations using this software have shown that the non-existence problem does not affect any of our investigations of models without three-way interactions, nor any in which zero counts are replaced by non-integer values such as ½ or 0.1. The software cannot be used to check the models that have been fitted using backward elimination, but it should be noted that the implications would be even

harsher than those presented here: for certain datasets the estimate of the dark number allowing for three-way interactions would not even exist.

As the work of Cruyff, van Dijk and van der Heijden (2017) and Silverman (2019) shows, ever more elaborate approaches can be used to address some of the problems alluded to here. However, the sparseness of the data – there are only 31 observations in the analyses explored here – is the ultimate barrier to reliable interpretation. Sadly, it is the conclusion of this paper that the difficulties in analysis arise from the very nature of the data themselves, and that it is unlikely that any statistical approach is capable of reliably uncovering the true number of pvoms from them.

Government agencies, NGOs and academics throughout the world are informing their policies and shaping their research in the light of estimates of the numbers of pvoms emanating from various sources. If lists of observed records of pvoms cannot be used to uncover the "dark number", as suggested here, what should such researchers do? One possibility is to use estimates derived from other sources. The Walk-free Foundation (2018), in their Global Slavery Index, publishes worldwide estimates of the prevalence of modern slavery, and their 2018 estimate pertaining to the United Kingdom is 136 000. This value is far larger than the estimates presented by BHS. The definition of a modern slave used by the Walk-free Foundation is far wider than that used in the BHS calculation, including for example women in forced marriages, and the methodology is based on data from international surveys combined with various techniques for extrapolation. While this method has been developed over time, various iterations of the Walk-free approach have been criticised for their methodology and claims to accuracy (Guth et al., 2014; Gallagher, 2017). We are not sufficiently confident of the Walk-free methodology to recommend its UK figure as an acceptable alternative estimate.

Estimation of the number of people not seen is intrinsically difficult, and perhaps the attempt to do so should be abandoned. Much of the use of estimates of the numbers of people in conditions of modern slavery is to make comparisons through time and across countries of the world or regions within countries. For these purposes, the "light number" could be used: that is the total number of pvoms actually observed (2 744 in Table 1). If this increases over time or varies geographically, then it is likely to be a reflection of a variation in the total number of pvoms. Variation in detection rates, especially over time, might be known about and allowed for, especially when they result from deliberate changes in policy.

Although actual values may never be known, changes might still be detected. Using MSE amounts to taking the light number, passing it through a rather unreliable "inflating machine", and then using the result. We are not convinced that this is more reliable in detecting changes than using the light number directly. The limitations of the latter are clear for users to appreciate, whereas the MSE method runs the risk of imparting a false sense of security to estimates produced.

In order to propose the adoption of a new methodology it is necessary to demonstrate its validity over a wide range of parameter values. In this paper, we demonstrate that the MSE method as implemented by BHS is unsound, and for this purpose a single counterexample showing that it fails is sufficient. We have not sought some pathological dataset for this demonstration: indeed we have chosen the parameter values most consistent with the observed data. The findings presented here may have implications for other applications of the MSE approach. IWGDMF (1995) show that some users of MSE appreciate the inference problems due to model selection. On the other hand, assuming that high order interactions are absent is essential for fitting MSE models. We believe that this assumption is far more dangerous in MSE models than in other regression models, and that it would certainly be of interest for researchers applying MSE to any form of data to check.

References

ATMG (2010). *Wrong kind of victim? One year on: an analysis of UK measures to protect trafficked persons*, London: The Anti-Trafficking Monitoring Group.

Baillargeon, S., & Rivest, L.-P. (2007). Rcapture: Loglinear models for capture-recapture in R. *Journal of Statistical Software*, 19, 1–31. Available at: https://www.jstatsoft.org/article/view/v019i05.

Bales, K., Hesketh, O., & Silverman, B. (2015). Modern slavery in the UK: How many victims? *Significance, 12,* 16-21.

Beddoe, C., Bundock, L., & Jardan, T. (2015). *Life Beyond the Safe House for Survivors of Modern Slavery*, London: Human Trafficking Foundation.

Chan, L., Silverman, B. W., & Vincent, K. (2019a). Multiple Systems Estimation for sparse capture data: Inferential challenges when there are non-overlapping lists. Available at: https://arxiv.org/pdf/1902.05156.pdf (accessed on 4 July 2019).

Chan, L., Silverman, B. W., &Vincent, K. (2019b). SparseMSE: Multiple systems estimation for sparse capture data. R package. https://cran.r-project.org/web/packages/SparseMSE/index.html (accessed on 11 July 2019).

Craig, G. (2017). The UK's modern slavery legislation: An early assessment of progress, *Social Inclusion*, *5,* 16–27

Cruyff, M., van Dijk, J., & 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.

Draper, N. R., &Smith, H. (1998). *Applied Regression Analysis*, (Third Edition). New York: Wiley.

Durgana, D. P., & Zador, P. L. (2017). Fighting slavery through statistics: A discussion of five promising methods to estimate prevalence in the United States. *Chance, 30*, 50-53.

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

Freeman, P. (1989). The performance of the two-stage analysis of two-treatment, two-period cross-over trials. *Statistics in* Medicine, *8,* 1421-1432.

Gallagher, A. (2017). 'What's wrong with the Global Slavery Index?' *Anti-Trafficking Review*, *8*, 90-112.

Guth, A., Anderson, R., Kinnard, K., & Tran, H. (2014). Proper methodology and methods of collecting and analyzing slavery data: An examination of the Global Slavery Index. *Social Inclusion, 2,* 14-22.

Goudie, I. B. J., & Goudie, M. (2007). Who captures the marks for the Petersen estimator? *Journal of the Royal Statistical Society: Series A (Statistics in Society*), 170, 825-839.

Hilbe, J. M. (2014). *Modeling Count Data*. Cambridge: Cambridge University Press.

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

International Working Group for Disease Monitoring and Forecasting (IWGDMF). (1995). Capture-Recapture and Multiple-Record Systems Estimation I: History and Theoretical Development. *American Journal of Epidemiology*, *142*, 1047-1058.

Johnson, S. R., Tomlinson, G. A., Hawker, G. A., Granton, J. T., Haddas, A, Grosbeing, B. M., & Feldman, B. M. (2010). A valid and reliable belief elicitation method for Bayesian priors. *Journal of Clinical Epidemiology*, 63, 370-383.

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

Morris, T. P., White, I. R., & Crowther, M. J. (2018). Using simulation studies to evaluate statistical methods. arXiv:1712.03198v2 [stat.ME]. https://arxiv.org/pdf/1712.03198.pdf

Moser, B. K., & Stevens, G. R. (1992). Homogeneity of variance in the two-sample means test. *American Statistician,* 46, 19-21.

Patterson, O., & Zhuo, X. (2018) Modern Trafficking, Slavery, and Other Forms of Servitude. *Annual Review of Sociology, 44*, 407–39.

Ripley, B., Venables, B., Bates, D. M., Hornik, K., Gebhardt, A., & Firth, D. (2013). Package `MASS'. Available at: https://cran.r-project.org/web/packages/MASS/MASS.pdf.

Silverman, B. W. (2018). modslavmse: Multiple Systems Estimates for estimating the prevalence of Modern Slavery. Available at:

https://github.com/bernardsilverman/modslavmse.

Silverman, B. W. (2019). Model fitting in Multiple Systems Analysis for the quantification of modern slavery: Classical and Bayesian approaches. Available at: https://arxiv.org/abs/1902.06078 (accessed on 22 April 2019).

Van Dijk, J.J.M., & van der Heijden, P.G.M. (2016). Research Brief. Multiple Systems Estimation for estimating the number of victims of human trafficking across the world. Vienna: UNODC. http://bit.ly/2vk7ypg.

Walk-free Foundation (2018). *Global Slavery Index.* https://www.globalslaveryindex.org/2018/data/country-data/united-kingdom/

Table 1: Counts used in the analysis of Model B

Each count is of the number of victims appearing on the lists indicated by "" and not on any of the other lists. Thus, for example, there are 695 people appearing only on the GO list, and 8 people on both the GO and the GP lists and on no others.*

GO	GP	LA	NG	PF	Count	GO	GP	LA	NG	PF	Count
Total count					2744	*	*	*			0
*					695	*	*		*		0
	*				316	*	*			*	0
		*			54	*		*	*		1
			*		463	*		*		*	0
				*	995	*			*	*	4
*	*				8		*	*	*		0
*		*			3		*	*		*	0
*			*		19		*		*	*	0
*				*	76			*	*	*	1
	*	*			0	*	*	*	*		0
	*		*		1	*	*	*		*	0
	*			*	11	*	*		*	*	0
		*	*		15	*		*	*	*	1
		*		*	19		*	*	*	*	0
			*	*	62	*	*	*	*	*	0

Table 2: Summary of the estimates of the total number of pvoms obtained when no three-way interaction parameters are fitted

Method	Estimate	95% confidence interval
Forward selection (BHS)	11 313	(9 889, 13 063)
All 2-way interactions	11 233	(6 155, 20 761)
As above, with ½ as GP-LA count	11 176	(6 133, 20 634)

Table 3: Summary of the different simulation investigations of analyses in which no three-way interactions are fitted (10,000 replicate runs)

Method	True value of κ	Average estimate of κ	confidence interval coverage
All 2-way interactions	8 432	8 807	0.9437
Forward selection (BHS)	8 432	12 193	0.3907

Table 4: Summary of the estimates of the total number of pvoms obtained when three-way interaction parameters are included

Method	Estimate	95% confidence interval
All 2- and 3-way interactions (using 0.1 counts)	3 392	(2 751, 64 373)
Backward elimination	5 552	(4 407, 7 485)

Table 5: Summary of the different simulation investigations of analyses in which three-way interactions are fitted (10,000 replicate runs)

Method	True value of κ	Average estimate of κ	confidence interval coverage
(a) All 2- and 3-way interactions	648	10^{14}	0.9967
(b) All 2-way interactions	648	9 664	0.0201
(c) Forward selection (BHS)	648	12 782	0.0000
(d) Backward elimination	648	10^{21}	0.4375