Article

Crime & Delinquency
1–17

DOI: 10.1177/0011128720981908
journals.sagepub.com/home/cad


Multiple System Estimation of Victims of Human Trafficking: Model Assessment and Selection

Maarten Cruyff[1], Antony Overstall[2], Michail Papathomas[3], and Rachel McCrea[4]

Abstract

Recently, multiple systems estimation (MSE) has been applied to estimate the number of victims of human trafficking in different countries. The estimation procedure consists of a log-linear analysis of a contingency table of population registers and covariates. As the number of potential models increases exponentially with the number of registers and covariates, it is practically impossible to fit and compare all models. Therefore, the model search needs to be restricted to a small subset of all potential models. This paper addresses principles and criteria for model assessment and selection for MSE of human trafficking with special attention to sparsity which is typical to human trafficking data. The concepts are illustrated on data from Slovakia and Romania.

Keywords

log-linear modeling, information criteria, modern slavery, BIC, AIC

[1]Utrecht University, Utrecht, Netherlands
[2]University of Southampton, Southampton, Hampshire, UK
[3]University of St. Andrews, St. Andrews, UK
[4]University of Kent, Canterbury, Kent, UK

Corresponding Author:
Maarten Cruyff, Faculty of Social Sciences, Utrecht University, Padualaan 14, Utrecht, 3584 CH, Netherlands.
Email: m.cruyff@uu.nl

Introduction

In 2016 the United Nations adopted the Elimination of Human Trafficking/Forced Labor as Target 16.2 of its 2030 Agenda for Sustainable Development. In this context, the UN Statistical Commission's Interagency and Expert Group on SDG Indicators (IAEG-SDGs) recommended to monitor the number of victims of human trafficking per 100,000 population, by sex, age, and form of exploitation. The institution responsible for collecting data on this indicator is the United Nations Office on Drugs and Crime (UNODC), and according to the UNODC, the indicator is composed of two parts: detected and undetected victims. The detected victims can be counted on a national level from activities of criminal justice systems, NGO's and other service providing institutions, while the number of undetected victims has to be estimated. According to UNODC, the methodology for this should allow for estimating the victims' sex, age, and form of exploitation.

The fact that these data are collected by a variety of institutions makes multiple systems estimation (MSE; e.g., Silverman, 2020) ideally suited for the estimation of the undetected number of victims. In its most simple form, the data for MSE consists of a cross-classification of two incomplete population registers A and B. This results in a two-by-two contingency table with the cells n_{10} representing the number of victims that have been observed in A but not in B, n_{01} representing the number of victims observed in B but not in A, and n_{11} representing victims observed in both A and B. The cell n_{00} representing the number of victims not observed in A nor in B, is to be estimated. The following assumptions are a necessary, but not sufficient, condition for unbiased estimation; (i) the inclusion probability in one register is independent of the inclusion probability in the other register, and (ii) for at least one register the inclusion probabilities are homogeneous in the population, that is, all members of the population have the same probability to be observed in that register. If these assumptions hold $\hat{n}_{00} = n_{10}n_{01} / n_{11}$. As an example, consider the table below. Given the observed frequencies, the estimate of the unobserved cell 00 is $\hat{n}_{00} = 100 \times 50 / 10 = 500$.

	$B = 0$	$B = 1$
$A = 0$	??	50
$A = 1$	100	10

Alternatively, the estimation can be performed with the log-linear model (A, B)

$$\log(\mu_{ij}) = \lambda_0 + \lambda_i^A + \lambda_j^B,$$

for $i, j \in \{0,1\}$, where μ_{ij} is the expectation of n_{ij}. The frequencies in the table above correspond to the following sets of parameters:

	$B = 0$	$B = 1$
$A = 0$	e^{λ_0}	$e^{\lambda_0+\lambda_1^B}$
$A = 1$	$e^{\lambda_0+\lambda_1^A}$	$e^{\lambda_0+\lambda_1^A+\lambda_1^B}$

Given that in our example $\hat{n}_{00} = 500$, the estimate $\hat{\lambda}_0 = \log(500) = 6.125$. The estimate $\exp(\hat{\lambda}_1^A) = -1.609 = \log(10/50)$ denotes the odds of inclusion in A given inclusion in B. Note that these odds are identical to the odds $100/500 = 1/5$ of inclusion in A and exclusion from B.

If the inclusion odds of one register do not depend on the status on the other register, the registers are said to be mutually independent. The log-linear model that allows for dependence has an additional parameter λ_{ij}^{AB}. This parameter is called a "two-way" interaction, as it corresponds to two lists. However, since the cell n_{00} is not observed, this parameter is not estimable. Consequently, with two registers, the log-linear model necessarily assumes mutual independence. The advantage of using log-linear models for population size estimation is that they are easily extended to data that include more than two lists and covariates.

The mutual independence assumption of the lists can be relaxed when there are more than two lists. With a third list C it becomes possible to estimate pairwise dependencies between the lists. This becomes possible because the number of persons not observed in two of the lists is observed for the persons observed in the third list. The notation for the log-linear model with three lists and all two-way interactions is (AB, AC, BC), also known as the homogeneous association model. This model assumes that pairwise dependencies between two lists do not depend on the level of the third list. The model that allows for heterogeneous associations contains three-way interactions and is denoted by (ABC). But this model cannot be estimated since the cell n_{000} denoting the numbers of persons not in A, B, and C is not observed.

The assumption of homogeneous inclusion probabilities can also be relaxed by extending the model to include covariates. Covariates may play either a passive or an active role in the model (Van der Heijden et al., 2012). For example, suppose we have the two lists A and B, and that males have a different probability to be included in list A than females. The correct model would than be (AS, B), where S denote the variable Sex. In this model S plays a passive role, since it affects the estimated composition of the population, but it does not affect the population size estimate. If S also interacts with list B, the correct model would be (AS, BS). Now S plays an active role, since it affects both the composition of the population and the population size estimate. Therefore, if covariates are available it is important to consider them in the analysis.

Each unique combination of interactions is referred to as a model. Different models can provide very different estimates of the number of victims. For this reason, model selection is critical. This is essentially a balancing act between model fit (i.e., suitability to observed cell counts) and model complexity (measured by the number of interactions). As a model becomes more complex, the fit to observed counts improves, reducing bias in the population size estimate. However, at the same time, the chance of including spurious interactions increases which can lead to high variability in the estimates, observing large changes in the parameter estimates for small changes in the data set due to random fluctuations. Furthermore, including a large number of interactions in the model increases the likelihood of fitting models with non-estimable parameters, with very wide associated confidence intervals. This typically manifests itself in unrealistically large (exploding) estimates. Conversely, a too simple model does not fit the observed counts, providing estimates with low variability but with potentially high bias. The aim of model selection is to find parsimonious models that are neither too simple nor too complex, and thus trade-off between bias and variance (e.g., Hastie et al., 2009, Chapter 7).

Typically when performing model selection, associations are assessed using some measure of statistical significance, with only significant associations included in the model. One commonly used approach is to assess significance with an information criterion, such as the AIC and BIC, that penalize model complexity in order to prevent overfitting of the data (these criteria are discussed in more detail in section 2.2). Since different information criteria may result in different models (and hence different population estimates), the choice of a criterion is of crucial importance. Once the criterion is chosen, a strategy for model selection has to be defined. When the number of variables in the model is small, this procedure may simply consist of fitting all possible models, and choosing the one which performs

best with respect to the chosen information criterion. For example, with two lists there is only one possible model, and with three lists there are seven possible models, and adding a single covariate increases this number to 28. However, when the number of variables further increases, it rapidly becomes impossible to fit all potential models. There are a large number of statistical methods for model selection including hypothesis testing; information criteria; and Bayesian methods, each aiming to balance model fit and parsimony. However estimating human-trafficking victims brings unique challenges to model selection for log-linear models (see Chan et al., 2019; Larsen & Durgana, 2017). Most notable of these challenges is data sparsity. The human trafficking data of the countries that have taken part in the UNODC monitoring project, the Netherlands (Cruyff et al., 2017; UNODC, 2017), Ireland (UNODC, 2018a), Serbia (UNODC, 2018b), Romania (UNODC, 2018c), and Slovakia (in press), are typically collected over a period of two or more years and include three or more lists, and three or more covariates. The numbers of observed victims are typically small in relation to the dimension of the contingency table, so the majority of the cells are sparsely filled or empty. In the case of the Netherlands, for example, the data were collected over a period of 6 years and included six lists and five dichotomous covariates (age, gender, form of exploitation, nationality, and year), yielding a contingency table with $2^{11} \times 6 = 12,288$ cells, of which $2^5 \times 6 = 192$ cells are not observed and are therefore structural zeros. The total number of observed cells is therefore $12,096$, and with a total of $8,324$ observations the average number of observations per cell is less than 1. This sparseness of data seriously limits the potential complexity of log-linear models for human trafficking data. Difficulties with model-fitting for multiple systems estimation human trafficking data are detailed in Silverman (2020), along with a discussion on the lack of robustness in the estimation of the size of hidden populations, for different model choices.

Typically, the scope of models eligible for selection is restricted to models that only include pairwise associations between the variables. This means for example that with three lists A, B, and C and one covariate S, the model (AB, AC, BC, AS, BS, CS) with 11 parameters is the most complex model allowed in the model search, while the more complex model (ABS, ACS, BCS) with 14 parameters could in theory also be fitted. For reasons of uniformity, these criteria were applied to all countries.

The aim of this paper is to highlight the importance of model selection in estimating human-trafficking. The information criteria approach to model selection is adopted as an example, but the arguments also apply to other model selection methods. Information criteria provide a score for each model,

balancing model fit and complexity, with the chosen model giving the smallest criterion. The datasets from Slovakia and Romania are considered, and it is shown how sparsity can affect model selection.

The remainder of the paper is organized as follows. Section 2 provides a technical description of log-linear models and model selection methods. In section 3, the results of applying information criteria to the Slovakia and Romania datasets are shown. The paper ends in section 4 with a discussion of future research directions.

Log-Linear Models and Model Selection for Estimating Human-Trafficking

Contingency tables and log-linear models

We provide a brief description of contingency tables and log-linear models (e.g., Fienberg, 1972) in the context of MSE. Suppose there are L lists available, labeled $1,\ldots,L$ and C covariates. If the c th covariate has g_c levels, then an incomplete contingency table with $2^L \times g_1 \times \ldots \times g_C$ cells is constructed. Each cell count gives the number of observed individuals by each list and covariate classification. The total population size is given by the sum of all cell counts. However the $g_1 \times \ldots \times g_C$ cell counts corresponding to not being observed on any of the lists (one for each classification of covariates) are unknown.

The basic idea underlying MSE is to fit statistical models to the observed cell counts, to identify underlying patterns (associations or interactions between lists and/or covariates), and to use this information to estimate the total population size.

Let y_i denote the cell count for cell $i = 1,\ldots,n$, where n is the number of observed cell counts. Log-linear modeling assumes that, independently

$$y_i \sim \text{Poisson}\left(\mu_i\right),$$

where the expected cell count is given by

$$\log \mu_i = \mathbf{x}_i^T \beta,$$

where $\mathbf{x}_i$ is a $p \times 1$ design vector and β a $p \times 1$ vector of unknown parameters. The design vector identifies the main effects and interactions relevant to cell $i = 1,\ldots,n$.

The total population size of victims of human trafficking can be estimated as follows. First, estimates, denoted by $\hat{\beta}$, for example through the use of maximum likelihood, are found for the unknown parameters, β. From this,

the maximum likelihood estimate of an unknown cell count with design vector $\mathbf{x}_0$ is given by $\exp\left(\mathbf{x}_0^T\hat{\beta}\right)$. An estimate for the total population size can therefore be formed by summing these estimates over all unknown cell counts. Finally, a confidence interval should be produced (e.g., Silverman, 2013), providing a representation of uncertainty in the estimate.

However this approach assumes that the interactions present are known which is rarely the case in practice. Different models can provide significantly different estimates of the total population size and therefore model uncertainty should to be taken into account.

Model Selection

Initially, a set of models $\mathcal{M}$ is posited that includes all models under consideration. This is sometimes referred to as the model set. Each model corresponds to a different combination of interactions, and $\mathcal{M}$ is typically restricted. First, the highest order interaction is usually specified to be much less than the theoretical upper limit of $L+C-1$. This is either to aid interpretation (higher-order interactions are non-trivial to interpret) or due to the sparsity of the data where it may not be possible to estimate all interactions in a given model (e.g., Sharifi Far et al., 2019). Again, for interpretation, the interactions present in a model usually obey the effect hierarchy principle giving $\mathcal{M}$ as a set of hierarchical log-linear models (e.g., Dellaportas & Forster, 1999). This means that if a model has a particular interaction present, then all lower-order interactions involving the terms in that interaction must also be present. For example, a model with the three-way interaction between A, B, and C denoted ABC, must also have the two-way interactions AB, AC, and BC. There exist more restrictive sets of log-linear models, that is, graphical and decomposable (see, e.g., Dellaportas & Forster, 1999 for details).

One of the simplest approaches to model selection is hypothesis testing (e.g., Davison, 2003, section 4.5). Two nested models are compared: one with a given interaction and one without. If the inclusion of the interaction leads to a *significant* increase in likelihood then the model with the interaction is retained. The main disadvantage of hypothesis testing is that it can only compare two models. An alternative approach to model selection is that of information criteria (Burnham & Anderson, 2002). Information criteria assess models in terms of complexity (i.e., the number of parameters in the model) and how well they fit the observed cell counts. The likelihood provides a measure of the fit of a given model, with higher values indicating superior fit. Using the likelihood as the criterion for model selection, however, will result in the model including all potential interactions. This model will fit the data perfectly, but its predictions will be unreliable due to high variance; the

parameter estimates may have large standard errors, and small changes in the data may produce large changes in the parameter estimates. By penalizing the number of parameters in the model the information criteria search for the most parsimonious model, that is, a model that fits the data adequately while keeping the variance of the model in check. Each model is given a score, and the model with the smallest score is then used to estimate the total population size. An information criterion (e.g., Davison, 2003, section 4.7) for a model can be written as follows

$$\mathrm{IC} = -2l(\hat{\beta}; \mathbf{y}) + c(p),$$

where $l(\hat{\beta}; \mathbf{y})$ is the value of the maximized log-likelihood for the model in question, and $c(p)$ is a penalty function increasing with the number of parameters, p. The information criterion decreases as $l(\hat{\beta}; \mathbf{y})$ increases and the model fit improves. However this is balanced against the model becoming more complex, with p (and $c(p)$) increasing. Different penalties correspond to different information criteria. The two most commonly used are the Akaike Information Criterion (AIC; Akaike, 1974) with $c(p) = 2p$, and the Bayesian Information Criterion (BIC; Schwarz, 1978) with $c(p) = \log(n)p$. AIC aims to choose the model which is optimal in terms of prediction. Conversely, BIC approximately chooses the model that is "closest" to the unknown data-generating process.

Estimation of the total population size is based on the model with the smallest information criterion. However even identifying this model can be problematic. The gold standard is to fit all models and select the model with the smallest criterion (e.g., Davison, 2003, section 8.7.3). However, even with the restricted set of models described above, there are typically a large number of models that can be fitted to a given contingency table. For example, suppose that attention is restricted to models with only pairwise interactions and that there are $P = L + C$ lists and covariates. Then there are are $2^{\binom{P}{2}}$ hierarchical log-linear models available, for example, if there are five lists and two covariates, then there are over two million models. This can render fitting all models prohibitively expensive, especially if higher-order interactions are also considered. Instead step-wise approaches are taken, that is, forward or backward selection (e.g., Davison, 2003, section 8.7.3).

Forward selection is given by completing the following steps.

0. Start by fitting the simplest model under consideration. This is the current model and calculate its IC.
1. Construct a proposal set of models by augmenting to the current model one interaction term at a time (while obeying effect hierarchy).

2. Calculate the IC of every model in the proposal set.
3. If the IC for the current model is less than the smallest IC from the proposal set, then stop, the current model is the final chosen model. Else, set the current model to be the model with the smallest IC from the proposal set and return to step 1.

The model chosen as the starting model in step 0 is usually the model with all main effects but no interactions.

Backward selection is similar but starts in step 0 with the most complex model under consideration and step 1 is replaced by the following step.

1. Construct a proposal set of models by removing from the current model one interaction term at a time (while obeying effect hierarchy).

Beyond the computational difficulties of finding the model that minimizes the chosen information criterion, there do exist other disadvantages. The theoretical basis of both AIC and BIC are based on a number of assumptions and approximations whose validity may be questionable for sparse data. Additionally, the estimate of the population size is based on the final chosen model. However there may be many models with near identical information criteria and these models may give significantly different estimates.

Alternatively, a fully Bayesian approach can be taken (King & Brooks, 2001; Madigan & York, 1997). This uses Bayes' theorem to construct a posterior distribution over models, that is, each model is assigned a probability of being the true model. These posterior model probabilities are not available in closed form but can be approximated using trans-dimensional Markov chain Monte Carlo (MCMC) methods. The most common is the Reversible Jump MCMC approach, introduced by Green (1995). The Bayesian approach does not use one single model to estimate the total population size. Instead the estimate of the total population size is averaged over models with a model's averaging weight equal to its posterior model probability. There do exist non-Bayesian model averaging techniques (Buckland et al., 1997) but we defer discussion until section 4.

Note that if the contingency table is sparse then there may be a large number of observed zero counts and the issue of parameter redundancy can arise. As described in section 1, this is a common occurrence in multiple systems estimation of human trafficking. In Chan et al. (2020), checks for parameter estimability based on a linear programming approach are used. Model selection is then conducted via a stepwise algorithm based on a predetermined threshold p-value. The algorithm fits a main effects model and then adds the

most significant interaction terms one-by-one subject to estimability checks, until a final model is adopted. An R package titled "SparseMSE" (Chan et al., 2019) is publicly available for the implementation of their methods to MSE data. See also Bales et al. (2020) for an application of this method to data from New Orleans-Metairie, USA. Whitehead et al. (2019) demonstrate that, under certain conditions, the reliability of estimates for the parameters of MSE models suffers when a two-stage procedure is used to identify which first order interaction terms to include in the model. A novel Bayesian approach is proposed in Silverman (2020) that allows for the inclusion of first-order interactions. This approach is based on reducing model parameters to zero with the use of a threshold value, and on checking the existence of estimates using the Chan et al. (2020) linear programming technique.

Examples

In this section, we present an overview of model selection for Slovakia and Romania. To avoid confusion about the definitions of human trafficking that were used in these analyses, we start the overview for each with a brief description of the organizations that collected the data, and the definitions of human trafficking that were used by organizations. For a detailed discussion of the definitions of human trafficking, refer to Bales et al. (2020).

Slovakia

The data from Slovakia are collected by the Information Centre to Combat Trafficking in Human Beings and Crime Prevention of the Ministry of Justice, and cover the period 2016 to 2018. The data used for carrying out MSE were composed by linking the registers from the police, support organizations, the Programme for Support and Protection, and other organizations on a yearly basis. These registers are denoted by R1 to R3, respectively. Over this 3-year period a total number of 189 victims were identified. The original registers distinguish between the trafficking categories sexual exploitation, forced begging, forced labor, and forced marriage, which in turn are subdivided in the primary, secondary, and tertiary type of exploitation. For the analysis, the primary type of exploitation was recoded into sexual and non-sexual exploitation. Aside from Exploitation (E), the data also includes the covariates Sex (S), Age (A - minors vs. adults), and Year of observation (Y).

With three registers and the four covariates S, A, E, and Y, the contingency table has $2^6 \times 3 - 2^3 \times 3 = 192 - 24 = 168$ potentially observed cells. The model search was performed using forward selection under both the BIC and

Table 1. Stepwise Model Selection Based on AIC and BIC for Slovakia.

Step	Df	Deviance	Resid. Df	Resid. Dev	AIC	BIC	Nhat
	—	—	159	261.5	179.7	130.5	387.4
+ S:E	−1	87.6	158	173.8	94.1	48.1	387.4
+ R2:R3	−1	28.6	157	145.3	67.5	24.8	523.7
+ R2:A	−1	15.7	156	129.6	53.8	14.3	523.7
+ R1:E	−1	13.0	155	116.6	42.8	6.6	523.7
+ R1:R3	−1	8.6	154	108.0	36.2	3.2	374.7
+ S:A	−1	8.4	153	99.6	29.8	0.0	366.8
+ S:Y	−2	8.4	151	91.2	25.5	2.2	366.8
+ E:Y	−2	7.6	149	83.6	21.8	5.0	366.8
+ R3:E	−1	3.2	148	80.4	20.7	7.1	369.3
+ A:Y	−2	4.6	146	75.8	20.1	13.0	369.7
+ R1:Y	−2	5.5	144	70.3	18.6	18.0	373.9
+ R2:Y	−2	10.4	142	59.9	12.1	18.0	543.9
+ R3:Y	−2	12.5	140	47.4	3.6	16.0	26844358178.0
+ R1:R2	−1	3.5	139	43.9	2.2	17.8	328.5
− R2:R3	1	0.3	140	44.2	0.4	12.8	712367104.0
+ R3:A	−1	2.5	139	41.7	0.0	15.6	2316938752.7

AIC information criteria. The starting model was the main effects only model $(R1, R2, R3, A, S, E, Y)$ and $\mathcal{M}$ was restricted to pairwise interactions only. Table 1 shows the results of this model search where the BIC and AIC values are re-scaled so that a value of 0 corresponds to the best model. The first row of Table 1 shows the AIC, BIC and population size estimate for the main effects model with no interactions. The second line shows the result of the first iteration of forwards selection. Inclusion of the interaction between Sex and Exploitation reduces both AIC and BIC but this does not affect the population size estimate. Subsequently including the interaction between the R2 and R3 lists reduces both information criteria and significantly affects the population size estimate. This procedure carries on until each information criteria stops decreasing. It is clear from Table 1 that different models lead to different estimates of the population size.

The best model as determined by BIC is $(SE, R2R3, R2A, R1E, R1R3, SA)$ has $p = 14$ parameters and yields an estimate of 367, with a 95% parametric bootstrap confidence interval (269, 622). The best model as determined by AIC has $p = 28$ parameters, but the model is obviously too complex for the data; the population estimate has exploded.

Romania

The data from Romania are maintained by the National Agency against Trafficking in Persons (ANITP) of the Ministry of the Interior, and include registers form the Police/NATP plus Border Police, IOM, NGOs, foreign authorities (mainly police forces) and other (mainly diplomatic missions). These are designated as R1 to R5. The data for the MSE were collected on a yearly basis over the years 2015 and 2016, and included a total of 1636 observations. The types of exploitation used for the MSE are sexual exploitation, beggary, and forced labor. Aside from type of Exploitation, the data included the covariates Sex, Age (minors vs. adults), Destination (D - transnational vs. domestic exploitation), and Year. With five registers and the five covariates S, A, D, E, and Y the contingency table consists of $2^9 \times 3 - 2^4 \times 3 = 1,488$ potentially observed cells. The model search procedure was performed analogously to that for Slovakia. Table 2 shows the results of the model search.

The model supported by BIC has $p = 36$ parameters, and yields a population estimate of 2541 with a 95% confidence interval (2011, 4153). The model supported by AIC has 11 more parameters, and yields an estimate of 2112.

Discussion

This paper has investigated the issue of model selection for multiple systems estimation models using two commonly used information criteria. The motivation for such work is a result of different models resulting in substantially different estimates of population size, which is of course the primary parameter of interest. The paper has focused on the use of two information criteria; AIC and BIC. However there are many alternative methods of model selection which have not been considered in this paper, in addition to those discussed in the Introduction. For some applications it is possible to fit all models in a prespecified model set, however it is sometimes necessary to take a practical exploration of model space, and one approach using information criteria has been described. Step-wise strategies using alternative test statistics have been considered in many fields. A step-up approach utilizing score tests has been applied in the related field of ecological capture-recapture models (McCrea & Morgan, 2011), and has the advantage that only the simpler model has to be fit to the data for the evaluation of the test statistic, thus avoiding the need to fit models which are not supported by the data. Disadvantages of such an approach however include the issue of multiple testing and the associated decision about significance levels and reliance on statistical significance has been heavily criticized in recent statistical articles (Wasserstein & Lazar,

Table 2. Stepwise Model Selection Based on AIC and BIC for Romanian Data.

Step	Df	Deviance	Resid. Df	Resid. Dev	AIC	BIC	Nhat
	—	—	1476	3600.8	3031.5	2850.1	2330.0
+ S:E	−2	858.6	1474	2742.2	2176.8	2006.2	2330.0
+ A:D	−1	586.0	1473	2156.1	1592.8	1427.6	2330.0
+ R2:R4	−1	344.9	1472	1811.3	1249.9	1090.1	2578.3
+ E:D	−2	237.9	1470	1573.4	1016.0	867.0	2578.3
+ R4:R5	−1	190.8	1469	1382.6	827.2	683.6	2937.2
+ R3:R4	−1	179.6	1468	1203.0	649.6	511.4	3500.0
+ R4:D	−1	155.9	1467	1047.0	495.7	362.9	3500.0
+ R1:R5	−1	107.1	1466	940.0	390.6	263.2	2169.0
+ A:E	−2	87.0	1464	853.0	307.7	191.1	2169.0
+ R5:E	−2	67.1	1462	785.9	244.6	138.8	2169.0
+ R1:R3	−1	40.6	1461	745.3	206.0	105.6	1855.9
+ E:Y	−2	43.1	1459	702.2	166.8	77.2	1855.9
+ R2:R5	−1	28.1	1458	674.1	140.8	56.6	1857.2
+ R1:R4	−1	22.7	1457	651.4	120.1	41.3	2777.3
+ R5:D	−1	20.4	1456	631.0	101.7	28.3	2777.3
+ R2:E	−2	24.3	1454	606.7	81.4	18.8	2787.9
+ D:Y	−1	11.7	1453	595.0	71.7	14.5	2787.9
+ R3:Y	−1	11.7	1452	583.3	62.0	10.2	2750.2
+ R1:D	−1	10.2	1451	573.1	53.8	7.4	2952.9
+ R1:R2	−1	8.2	1450	565.0	47.6	6.6	2632.7
− R2:R5	1	0.7	1451	565.7	46.4	0.0	2540.6
+ R3:E	−2	12.5	1449	553.2	37.9	2.3	2260.8
+ R5:S	−1	7.2	1448	546.0	32.6	2.4	2260.8
+ S:A	−1	6.0	1447	540.0	28.6	3.8	2260.8
+ S:D	−1	7.0	1446	532.9	23.6	4.2	2261.0
+ R2:R3	−1	5.5	1445	527.5	20.1	6.1	2435.0
+ R2:Y	−1	7.2	1444	520.3	14.9	6.3	2437.7
+ R4:Y	−1	12.6	1443	507.7	4.4	1.2	2454.6
+ A:Y	−1	4.4	1442	503.3	1.9	4.1	2456.2
+ R5:Y	−1	3.5	1441	499.8	0.5	8.1	2443.4
+ R3:R5	−1	2.5	1440	497.3	0.0	13.0	2112.2

2019). Further, the calculation of the expected information matrix, which is required for the evaluation of the score test statistic, is not always straightforward and approximations using the observed information matrix have been found to be flawed for some applications (Morgan et al., 2007).

An exploration of the model set in a classical paradigm has been proposed by Brooks et al., (2003), implementing a trans-dimensional simulated annealing approach. This is the classical counterpart to the Bayesian reversible jump MCMC approach (King et al., 2001) described in section 2.2. Generally however such approaches need tuning, and therefore do not offer a model selection solution which will work for all applications. Therefore this paper has focused on two approaches which can be easily implemented for all MSE applications. As mentioned earlier, it is possible to provide model-averaged estimates of the population size accounting for model-uncertainty. A decision on whether to model-average based on AIC or BIC is well-debated in the literature. Fletcher (2019) suggests that AIC weights should be better than BIC weights, as model averaging is about estimation and prediction and this is the goal of AIC. In contrast, BIC is focused on identifying the true model. For a fuller discussion of this, see Yang (2005). Some alternative criteria for model-averaging have been reviewed in Dormann et al. (2018) and a systematic review of the state of the art is given in Chapter 3 of Fletcher (2019).

This paper has shown that model assessment and selection play a crucial role in MSE of victims of human trafficking. It has also alluded to the alternative strategies that can be followed to arrive at a population size estimate. With the growing number of publications on MSE of human trafficking, —aside from the above mentioned UNODC studies there have also been studies in the UK (Bales et al., 2015), the US (Bales et al., 2020), and Australia (Lyneham et al., 2019)—our knowledge and understanding of the intricacies of human trafficking data will grow. From a methodological point of view, we envisage that future research will lead to a better understanding on how sparsity affects the estimability of model parameters for this type of data, and the prospect of modeling higher order interactions and more complex dependencies between the lists and covariates. All this will help us to optimize our strategies for model assessment and selection for MSE of human trafficking victims. This in turn will help national governments to formulate more substantiated policies on resource allocation for detection and prevention of human trafficking.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Maarten Cruyff https://orcid.org/0000-0002-6808-8518

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

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

Bales, K., Murphy, L. T., & Silverman, B. W. (2020). How many trafficked people are there in Greater New Orleans? Lessons in measurement,. *Journal of Human Trafficking 6*, 375–387.

Brooks, S. H., Friel, N., & King, R. (2003). Classical model selection via simulated annealing, *Journal of the Royal Statistical Society Series B*, *65*, 503–520.

Buckland, S. T., Burnham, K. P., & Augustin, N. H. (1997). Model selection: An integral part of inference. *Biometrics*, *53*, 603–618.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information – Theoretic approach* (2nd ed.). Springer.

Chan, L., Silverman, B. W., & Vincent, K. (2019). "*SparseMSE: Multiple systems estimation for sparse capture data*," R Package Version 2.0.1.

Chan, L., Silverman, B.W., & Vincent, K. (2020). Multiple systems estimation for sparse capture data: Inferential challenges when there are nonoverlapping lists. *Journal of the American Statistical Association*. https://doi.org/10.1080/01621459.2019.1708748

Cruyff, M. J. L. F., 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.

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

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

Dormann, C. F., Calabrese, J. M., Guillera-Arroita, G., Matechou, E., Bahn, V., Barton, K., Beale, C. M., Ciuti, S., Elith, J., Gerstner, K., Guelet, J., Keil, P., Lahoz-Monfort, J. J., Pollock, L. J., Reineking, B., Roberts, D. R., Schröder, B., Thuiller, W., Warton, D. I., . . . Hartig, F. (2018). Model averaging in ecology: A review of Bayesian, information-theoretic, and tactical approaches for predictive inference. *Ecological Monographs*, *88*, 485–504.

Fienberg, S. E. (1972). The multiple recapture census for closed populations and incomplete 2^k contingency tables. *Biometrika*, *59*, 591–603.

Fletcher, D. (2019). *Model averaging*. Springer.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, *82*, 711–732.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. Springer.

King, R., & Brooks, S. P. (2001). On the Bayesian analysis of population size. *Biometrika*, *86*, 615–633.

King, R., Morgan, B., Gimenez, O., & Brooks, S. P. (2009). *Bayesian analysis for population ecology*. Chapman & Hall/CRC.

Larsen, J. J., & Durgana, D. P. (2017). Measuring vulnerability and estimating prevalence of modern slavery. *Chance*, *30*, 21–29. https://doi.org/10.1080/09332480.2017.1383109

Lyneham, S., Dowling, C., & Bricknell, S. (2019). *Estimating the dark figure of human trafficking and slavery victimisation in Australia*. Statistical Bulletin No. 16 Canberra: Australian Institute of Criminology. https://www.aic.gov.au/publications/sb/sb16

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

McCrea, R. S., & Morgan, B. J. T. (2011). Multi-state mark-recapture model selection using score tests. *Biometrics*, *67*, 234–241.

Morgan, B. J. T., Palmer, K. P., & Ridout, M. S. (2007). Negative score test statistic. *The American Statistician*, *61*, 285–288.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Sharifi Far, S., Papathomas, M., & King, R. (2019). Parameter redundancy and the existence of the MLE in Log-linear models. *Statistica Sinica*. In press. https://doi.org/10.5705/ss.202018.0100

Silverman, B. (2013). *Modern slavery: An application of multiple systems estimation*. Home Office.

Silverman, B.W. (2020). Multiple-systems analysis for the quantification of modern slavery: Classical and Bayesian approaches (with discussion). *Journal of the Royal Statistical Society, Series A*, *183*: 691–736.

UNODC. (2017). *Monitoring Target 16.2 of the United Nations Sustainable Development Goals: A multiple systems estimation of the numbers of presumed victims of trafficking in persons in the Netherlands in 2010–2015 by year, age, gender, form of exploitation and nationality*. Research Brief. Vienna: UNODC.

UNODC. (2018a). *Monitoring Target 16.2 of the United Nations Sustainable Development Goals: Multiple systems estimation of the numbers of presumed victims of trafficking in persons: Ireland*. Research Brief. Vienna: UNODC.

UNODC. (2018b). *Monitoring Target 16.2 of the United Nations Sustainable Development Goals: Multiple systems estimation of the numbers of presumed victims of trafficking in persons: Serbia*. Research Brief. Vienna: UNODC.

UNODC. (2018c). *Monitoring Target 16.2 of the United Nations Sustainable Development Goals: Multiple systems estimation of the numbers of presumed victims of trafficking in persons: Romania*. Research Brief. Vienna: UNODC.

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

Wasserstein, R. L., & Lazar, N. A. (2019). Editorial: The ASA's statement on p-values: Context, process and purpose. *The Americam Statistician*, *70*, 129–133.

Whitehead, J., Jackson, J., Balch, A., & Francis, B. (2019). On the unreliability of multiple systems estimation for estimating the number of potential victims of modern slavery in the UK. *Journal of Human Trafficking*. https://doi.org/10.1080/23322705.2019.1660952

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika*, *92*, 937–950.

Author Biographies

Maarten Cruyff is Associate Professor at Methods & Statistics at Utrecht University. His research interest are population size estimation and randomized response.

Antony Overstall is Associate Professor at Mathematical Sciences at University of Southampton. His research interest are optimal experimental design and the analysis of categorical data, particularly incomplete contingency tables.

Michail Papathomas is Senior Lecturer in Statistics at the School of Mathematics and Statistics of the University of St Andrews. His research interests are linear modeling, model comparison, parameter identifiability and Bayesian inference.

Rachel McCrea is a Senior Lecturer in Statistics at the School of Mathematics, Statistics and Actuarial Science at the University of Kent. She is also Director of the National Centre for Statistical Ecology. Her research interests include model assessment methods, in particular for integrated population models, multi-state models, capture-recapture models and removal modeling.