Biometrics ??, **??–??** DOI: 10.1111/???
XXXXX XXXX

Discussion to "The central role of the identifying assumption in population size estimation"

Ruth King[1], Rachel McCrea[2] and Antony Overstall[3]

[1] School of Mathematics, University of Edinburgh, James Clerk Maxwell Building,
The King's Buildings, Peter Guthrie Tait Road, Edinburgh, UK, EH9 3FD.

[2] Department of Mathematics & Statistics, Lancaster University, Lancaster, UK, LA1 4YF.

[3] School of Mathematical Sciences, University of Southampton, Highfield, Southampton, SO17 1BJ.

**email:* Ruth.King@ed.ac.uk

Summary: In this discussion response we consider some practical implications of the authors' consideration of the no highest order interaction model for multiple systems estimation which permits the authors to derive the explicit (albeit untestable) identifying assumption related to the unobserved (or missing) individuals. In particular, we discuss several aspects, from the standard process of model selection to potential poor predictive performance due to overfitting and the implications of data reduction. We discuss these aspects in relation to the case study presented by the authors relating to the number of civilian casualties within the Kosovo war, and conduct further preliminary simulations to investigate these issues further. The results suggest that the no highest order interaction models considered, despite having a potentially useful theoretical result in relation to the underlying identifying assumption, may perform poorly in practice.

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

We would like to thank the authors for this interesting paper that challenges the readers, and more particularly the practitioners of multiple systems estimation, to consider further the underlying identifying assumption of the models. Common modelling techniques for multiple systems estimation require an untestable identifying assumption to be made to estimate the unobserved population size, with the assumption relating to the unobserved, (and hence missing sector) of the population. However this issue is typically overlooked and no identifying assumption explicitly stated. The paper asserts that if no identifying assumption can be justified for a given dataset, then no estimate of the population size should be provided.

The paper shows that for the no highest-order interaction (NHOI) log-linear model (i.e. for a dataset with K lists, the log-linear model with all interactions present, up to and including $(K-1)$-way interactions), the associated identifying assumption can be expressed in the form of a log odds ratio equivalence. Consequently a level of sensitivity of the identifying assumption can be investigated by varying this defined log odds ratio. However, it is noted that this ratio is difficult to interpret when there are more than $K = 3$ lists; and thus when the number of lists is $K > 3$, the paper suggests reducing to the consideration of the marginal K' $(< K)$ NHOI if there is suitable domain knowledge to permit this list reduction. This approach is applied to the case study relating to the number of civilian casualties in the Kosovo war where there are $K = 4$ lists, but using domain knowledge this is reduced to $K' = 2$ (where the domain knowledge is that the 2 lists are likely to be independent).

We strongly agree with the authors that the sensitivity of the estimates should be assessed to deviations from the statistical assumptions made. Consequently, our discussion focuses on the implications of this paper to the practice of multiple systems estimation, focusing on the use of the Kosovo example presented as a practical case study. In particular in relation to the consideration of using the NHOI model we discuss further issues relating to (i) model selection and the bias-variance trade-off; (ii) the robustness of the estimate due to potential over-fitting; and (iii) data reduction.

2. Model selection

The mainstream approach in multiple systems estimation (e.g. Silverman, 2020) involves considering the set of hierarchical log-linear models. The estimate of the total population size can vary substantially dependent on the log-linear model in terms of the presence/absence of interactions within the model being fitted to the data (e.g. Hook and Regal, 1995). Model selection uses a data-driven approach adopting the principal of parsimony, balancing the trade-off between bias and variance within the estimation process. To investigate the sensitivity of the estimated population size for the Kosovo data in relation to the NHOI model, we conducted a model selection process. In particular, we considered Akaike's Information Criterion (AIC) to compare competing hierarchical log-linear models, using a step-wise search algorithm adding/removing interaction terms until no improvement in AIC statistic was found. Applying this approach, the model determined to be optimal contained all interactions except those that involved an interaction between the sources ABA and HRW. The corresponding maximum likelihood estimate for the total population size under this model is 12740, which is indeed somewhat different from the estimate of 16940 under the NHOI model. The MLEs of the estimated model parameters for the NHOI and model deemed optimal via the AIC are provided in Appendix A. In practice, we would strongly advocate that the sensitivity of the population estimate is investigated with respect to different competing (plausible) models. Further, model-averaging techniques may be applied to obtain an estimate of the total population size that takes into account both the parameter and additional model uncertainty, (e.g. Buckland et al., 1997; Hoeting et al., 1999; King and Brooks, 2001).

3. Over-fitting

In general, applying a more complex model than necessary may lead to both over over-fitting and associated poor precision for the model parameters. This, in turn, can lead to poor predictive performance. We investigate these issues for the NHOI models fitted to the Kosovo data.

Firstly, we consider the precision of the population size estimates. For the NHOI model, we obtain an MLE of the total population size of 16941 with 95% (non-parametric bootstrap) confidence interval (CI) of (8980, 36100). (For consistency throughout we calculate non-parametric confidence intervals which avoids the need for any asymptotic normality assumptions, and so these CIs differ to those of the authors). For comparison, the model deemed optimal via the AIC statistic has an MLE of the total population size of 12741 with 95% CI of (9740, 18100). The CI for the more complex NHOI model is, as expected, significantly wider with the upper limit essentially twice that of the optimal AIC model (despite only having 3 additional parameters).

Secondly, to investigate the issue of the potential over-fitting of the model to the data, we conducted a simulation study. In particular, we simulated new datasets by randomly perturbing each observed cell entry uniformly between $\pm$ 10% (and 5% and 1%) of the observed cell entry and refitted the NHOI model and model deemed optimal via the AIC statistic for the original data. For completeness, we also fit the reduced $K' = 2$ NHOI model to the perturbed data. A summary of the results obtained is presented in Table 1. We note that the sensitivity of the estimates of the unobserved population is substantially more marked for the NHOI model. For example, at the 5% perturbation level, $>$ 50% of the perturbed datasets led to an estimate of the number of unobserved individuals that differed to the associated estimate for the (unperturbed) observed data by at least 10%; whereas for both the model deemed optimal and the reduced $K' = 2$ NHOI model, $>$ 50% of the perturbed datasets led to an estimate that differed by at most 5% to the corresponding estimate obtained for the observed data. Thus these preliminary results suggest that the estimates obtained from the NHOI model may not be robust to even relatively small perturbations of sampling variability we would reasonably expect within the observed data.

[Table 1 about here.]

4. Discarding data

The model deemed optimal via the AIC statistic corresponds to all interactions (up to order 3) present except those between ABA and HRW, suggesting that the independence assumption between ABA and HRW appears valid, given the observed data. It would seem a useful exercise (and good practice) that, in general, any such assumed domain knowledge is tested with the observed data, and we discuss this further in Section 5. However, reducing the dataset to K' lists, and in the paper the authors consider $K' = 2$ lists (ABA and HRW), discards a substantial amount of information (including 2295 individuals not identified by either of these sources). There is no discussion or investigation in the paper of the impact of discarding such data on population estimates, including, for example, whether interactions present with other lists may lead to biases in the associated estimates when the lists are marginalised and reduced in this way. This would seem to be of primary importance before marginalising the data and considering only a subset of the available data based on domain knowledge. See Sharifi Far et al. (2021) for further discussion regarding investigating the sensitivity of multiple systems estimation via list inclusion/exclusion. We discuss this particular issue next via a simulation study.

To investigate the issue of reducing the data to K' lists, we simulated 50,000 data sets from the fitted NHOI model to the Kosovo data (for the associated parameter values see Table 1) and applied the reduced $K' = 2$ model, considering only the data observed by ABA

and HRW. In all cases the estimated MLE of the $K' = 2$ NHOI model was lower than the simulated true value. Further, in all cases, the 95% (non-parametric bootstrap) CI for the total population size did not contain the true value from the model, and in particular the MLE was 64-70% below the true value 95% of the time. This suggests that, for at least data similar to the observed Kosovo data, the estimated population size is not robust to reducing the data to considering the absence/presence of individuals on only the 2 data lists, (associated with ABA and HRW) when considering the NHOI model. We repeated the process using the fitted model deemed optimal via the AIC statistic, where the lists ABA and HRW were independent in the model, leading to 89% of the corresponding 95% CIs containing the true value for the total population size. The associated difference of the MLE to the true value of the total population size was (-13%, 20%) in 95% of cases.

To investigate the performance of the reduced K' model, we considered a further simulation study, considering $K = 4$ lists and the model with all two-way interactions present, except between two lists, which we label S_3 and S_4. We then reduced the data to simply consider the $K' = 2$ independent lists, S_3 and S_4. See Appendix B for further description of the model and associated parameter values. We again simulated 50,000 datasets, given the specified model, before reducing the model to simply the two independent lists and estimate the associated total population size. From the simulations, we first note that the MLE of the total population size using the $K' = 2$ lists was less than the number of *observed* individuals from the $K = 4$ lists (which is a lower bound of the total population size) for 37% of the simulated datasets; and again none of the 95% CI contained the true value of the population size. These results are despite the model used to simulate the data assuming that the two $K' = 2$ lists are independent.

The results obtained from these limited simulation studies suggest that additional investigation is warranted to determine the appropriateness of considering reduced K' NHOI models.

5. Concluding remarks

The authors highlight the untestable identifying assumptions in their paper relating to multiple systems estimation. However, to date, an explicit identifying assumption has only been specified for NHOI models; yet a number of practical issues can arise when considering the saturated NHOI model as discussed above and demonstrated for the given case study. In particular, within this discussion we have focused on potential issues which can arise when considering NHOI models without due care, and explicitly highlight the potential risks of over-fitting and discarding data, thus losing potentially valuable information, and leading to poor population size estimates.

The set of hierarchical log-linear models are commonly fitted within multiple systems estimation, which are simply nested sub-models of the NHOI model. These nested models are specified such that given interaction terms are simply set to the value zero. Following this intuitive line of thought (of simply setting interaction terms to be equal to zero), readers of the paper may wonder why the identifying assumptions do not follow in the analogous manner; and also whether the subsequent interpretation of the previous log-odds ratio changes.

Considering datasets beyond the Kosovo case study presented, many such multiple systems estimation datasets are sparse in nature, with (multiple) cell entries equal to 0. The sparsity of the data typically increase as K increases. The presence of 0 cell entries can lead to parameters that are non-estimable (e.g. Fienberg and Rinaldo, 2012; Sharifi Far et al., 2021) or are estimated to be infinite (e.g. Chan et al., 2021). This issue will be most acute for the NHOI model, but can also affect many nested sub-models. Do the authors have any thoughts or insights in to the implications of such 0 cell entries for their results? For example, how infinite (or non-estimable) estimates affect either the interpretation and/or sensitivity of the identifying assumption.

Finally, we hypothesise that consideration of ideas associated with the (related) field of ecological capture-recapture modelling may provide some additional insight into the general thorny problem regarding the validity of underlying assumptions (though it is unavoidable that assumptions relating to the unobserved individuals are unverifiable). For example, within the ecological models, the validity of some assumptions can be considered through a series of diagnostic goodness-of-fit tests derived through the factorisation of the likelihood function: one component of which is used to estimate parameters, whilst the other can be used to assesses model adequacy (Pollock et al., 1985). Similar ideas can be applied to investigate the case of suspected independence between two lists, where we may wish to consider the $K' = 2$ case. For example, list independence can be examined within the observed data, conditional on being observed by each other list in turn (or combination of lists).

The need to be able to provide robust estimates of population size is of key importance across many different fields. Consequently it is critical that limitations of multiple systems estimation are fully understood and explored. This paper has articulated explicit identifying assumptions relating to the unobserved individuals that can be applied to the NHOI models (and challenges readers to derive these for alternative models), although the interpretation is challenging for $K > 3$ lists. Practical issues also arise when applying the specified NHOI model to data, with some of these briefly investigated above. In general, the particular approach taken, will be dictated by the specific question(s) of interest. However, due to the many practical issues that arise for the NHOI model, we suspect that practitioners will continue to focus on using model selection techniques and/or model-averaged results within multiple systems estimation, rather than the saturated model which has a nice theoretical result relating to an explicit untestable assumption for the unobserved individuals. However a better understanding of these issues within multiple systems estimation, and additional theoretical results, are likely to be ongoing areas of future research.

Received XXXXX 20XX. *Revised XXXXX* 20XX. *Accepted XXXXX* 20XX.

Appendix A

The estimates of the model parameters for the Kosovo data for the NHOI model and model deemed optimal via AIC statistic are provided in Table 2.

[Table 2 about here.]

Appendix B

For the additional data that are simulated in Section 4 we consider four sources labelled S_1, S_2, S_3 and S_4. The data are generated from the model with all two-way interactions, excluding the interaction between lists S_3 and S_4, i.e. model $\{S_1 \times S_2, S_1 \times S_3, S_2 \times S_3, S_1 \times S_4, S_2 \times S_4\}$. The parameters specified for the log-linear parameters are given in Table 3. Within the specification of the model, we use corner point constraints for all parameters, such that the parameter values given above relate to the "upper" level of the associated log-linear parameter.

[Table 3 about here.]

References

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

Chan, L., B. Silverman, and K. Vincent (2021). Multiple systems estimation for sparse capture data: Inferential challenges when there are non-overlapping lists. *Journal of the American Statistical Association 116*, 1297–1306.

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

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky (1999). Bayesian model averaging: A tutorial. *Statistical Science 14*(4), 382 – 417.

Hook, E. B. and R. Regal (1995). Capture-recapture methodology in epidemiology: Methods and limitations. *Epidemiological Reviews 17*, 243–64.

King, R. and S. P. Brooks (2001). On the Bayesian analysis of population size. *Biometrika 88*, 317–336.

Pollock, K. H., J. E. Hines, and J. D. Nichols (1985). Goodness-of-fit tests for open capture-recapture models. *Biometrics 41*(2), 399–410.

Sharifi Far, S., R. King, S. M. Bird, A. M. Overstall, H. Worthington, and N. P. Jewell (2021). Multiple systems estimation for modern slavery: Robustness of list omission and combination. *Crime and Delinquency 67*, 2213–36.

Sharifi Far, S., M. Papathomas, and R. King (2021). Parameter redundancy and the existence of the maximum likelihood estimates in log-linear models. *Statistica Sinica 31*, 1125–1143.

Silverman, B. (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.

Table 1: The minimum, lower 2.5% quantile, lower 25% quantile; upper 25% quantile; upper 2.5% quantile and maximum MLE of the total population size when perturbing the observed data cell entries by (i) 10%; (ii) 5%; and (iii) 1% for the NHOI model; model deemed optimal for the Kosovo data via the AIC criterion; and reduced 2 list model using 10,000 simulated datasets. The MLE for the observed data was 16941 for model NHOI; 12741 for the model deemed optimal; and 9691 for the reduced 2 list model.

	NHOI model			Optimal model			2-list NHOI model		
	Perturbation level			Perturbation level			Perturbation level		
	10%	5%	1%	10%	5%	1%	10%	5%	1%
Minimum	7986	10658	14046	10268	11389	12261	8052	8730	9283
2.5%	11910	13843	15278	10970	11766	12346	8702	9140	9448
25%	14799	15705	16304	12028	12356	12547	9317	9487	9601
75%	19393	18269	17599	13491	13129	12945	10063	9896	9780
97.5%	25486	21175	18915	14945	13836	13162	10802	10280	9945
Maximum	35338	25537	20172	16113	14322	13261	11911	10886	10137

Table 2: The MLEs (standard errors) for the model parameters for the NHOI model and model deemed optimal via the AIC statistics for the Kosovo data.

Parameter	NHOI model	Optimal model
Intercept	9.437 (0.473)	9.029 (0.192)
ABA	-2.695 (0.472)	-2.290 (0.189)
EXH	-2.406 (0.472)	-1.999 (0.194)
HRW	-3.713 (0.470)	-3.305 (0.182)
$OSCE$	-2.595 (0.472)	-2.190 (0.194)
$ABA \times EXH$	0.843 (0.465)	0.443 (0.204)
$ABA \times HRW$	0.408 (0.433)	–
$ABA \times OSCE$	1.234 (0.466)	0.842 (0.201)
$EXH \times HRE$	1.346 (0.459)	0.949 (0.206)
$EXH \times OSCE$	0.994 (0.467)	0.603 (0.207)
$HRW \times OSCE$	1.684 (0.460)	1.299 (0.202)
$ABA \times EXH \times HRW$	-0.326 (0.340)	–
$ABA \times EXH \times OSCE$	0.388 (0.448)	0.745 (0.234)
$ABA \times HRW \times OSCE$	-0.293 (0.377)	–
$EXH \times HRW \times OSCE$	-1.008 (0.416)	-0.722 (0.258)

Table 3: Parameters specified for the log-linear model in Section 4.

Parameter	Intercept	S_1	S_2	S_3	S_4	$S_1 \times S_2$	$S_1 \times S_3$	$S_2 \times S_3$	$S_1 \times S_4$	$S_2 \times S_4$
Value	6.0	-2.3	-2.8	-2.4	-1.9	1.5	1.9	1.8	1.2	2.7