Report:

Capture-Recapture Models for Population Estimates

Antony Overstall and Ruth King

University of St Andrews

Summary

1. Standard capture-recapture models can be extended to allow for additional model complexities, removing some of the usual modelling assumptions.

2. New models are developed for three different problems/modelling assumptions:

   (a) false negatives (due to corrupted data entries) for one data source;

   (b) partially collated data where two data sources are not collated/matched;

   (c) non-target individuals observed by one data source.

3. Bayesian data augmentation techniques are developed to fit the proposed model and simulation studies undertaken to assess how the models perform (assuming independent data sources).

4. The following was observed from the simulation studies:

   (a) for false negatives a small negative bias was generally observed in the population estimate;

   (b) for partially collated data the estimates were reasonable, though some small negative bias appeared to be observed;

   (c) for non-target individuals observed by one source, the results appeared to be generally unbiased (with additional subsampling leading to additional improvements, albeit relatively small, in the precision of the estimates).

5. The problems identified and corresponding models developed are for the simplest cases, for example, only one source with non-target individuals; only one source with corrupted individuals. The proposed models can be extended further, for example, non-target individuals identified by more than one source, and combined to consider multiple data collection issues. However, this is beyond the remit here, but feasible future work.

Introduction

Capture-recapture (or multi-list) data are often used in order to obtain estimates of hidden or hard-to-reach populations (for example, see Fienberg (1972), Hook and Regal (1995) and Chao *et al* (2001) with recommendations on the use of such methods presented by Hook and Regal (1999) and Hook and Regal (2000)). Capture-recapture data are typically presented in the form of an incomplete contingency table, corresponding to the observed number of individuals by each distinct combination of data sources. The contingency table is incomplete, however, since the number of individuals that are in the population but not observed by any source is unknown. A common approach for estimating the total population size (or unobserved contingency table cell) is to fit log-linear models to the observed data (i.e. observed contingency table cells). [1]

This report describes a feasibility study for estimating population sizes using capture-recapture methods in the presence of:

(1) imperfect linkage between data source; or

(2) non-target individuals observed within the data sources.

These are issues that arise for multiple administrative data lists that may be used in the future within the NRS. Thus, the feasibility study will assess the possible implications of these issues using simulated datasets and propose and fit new log-linear capture-recapture methods to formally model these additional complications. Due to the nature of these problems, a Bayesian (data augmentation) approach will be considered[2]. This will facilitate the statistical analysis by permitting easier model fitting algorithms, given the "non-standard" log-linear models developed to allow for the above additional modelling assumptions. In addition, standard model-averaging algorithms can be incorporated, extending the models proposed within this study and potentially allow for external information to be incorporated into analyses, for example, relating to the total population size. However, note that these additional issues are not the focus of this study. All associated computational algorithms used within the study will be made available to NRS upon request (but no support will be able to be provided). Technical mathematical details are provided in the associated appendices (and references).

We note that for simplicity, in order to develop the modelling framework and assess the performance of the models, we will assume that there are four data sources such that they all are independent of each other (i.e. being observed by one source does not affect the probability of being observed by any other source and vice versa). We note that the modelling ideas proposed can be extended to a larger number of data sources and general log-linear models permitting interactions between different sources (and model-averaging), although further investigation of parameter identifiability may be needed (see Discussion). Notationally, we let A, B, C and D denote the four data sources and let the corresponding contingency table cells be denoted by $\{i, j, k, l\}$ such that $i, j, k, l \in \{0, 1\}$ where $i = 1/0$ corresponds to an individual being observed/not observed by source A, and similarly for j, k, l in relation to sources B, C and D, respectively. For example contingency cell $\{1, 0, 0, 0\}$ corresponds to individuals that are only observed by source A.

[1] Standard log-linear models and described further in Appendix A.

[2] For a brief summary of the Bayesian paradigm see Appendix B.

We now consider the following three separate modelling issues in turn:

(i) imperfect linkage leading to false negatives;

(ii) imperfect linkage leading to partially collated data;

(iii) non-target individuals observed.

For each problem we develop a modelling framework (and associated model-fitting algorithm) and undertake a simulation study. In particular we consider a total of 1000 individuals in the population of interest (the population size chosen here is for computational purposes). Two "scenarios" are considered, corresponding to different probabilities that each data source observes a given individual in the population: (1) "extreme" scenario, where the probability an individual is observed by each source is 0.1 (so that the probability of being unobserved is 0.66); and (2) "moderate" scenario, where the probability of being observed by each source is 0.4 (so that the probability of being unobserved is 0.13). A total of 1000 datasets (i.e. observed contingency tables) are simulated for each of the "moderate" and "extreme" scenarios, assuming that the data sources observe individuals independently of each other (and allowing for the additional modelling complexities for each problem, described separately for each problem).

(i) Imperfect linkage - false negatives

In standard capture-recapture models it is assumed that individuals are uniquely identifiable and perfectly matched across all different data sources (so that there are no false negatives/positives in matching individuals). However, imperfect data linkage can arise in a number of different ways as a result of, for example, imperfect matches between sources and/or partially collated data sources. We consider the former problem here (and consider the latter problem later). For imperfect matches we will focus on the possibility of false negatives, and assume that false positives are not possible. For capture-recapture data, a false negative corresponds to an individual that is observed by two different data sources, but is not "matched" between these different sources; whereas a false positive would correspond to two unique individuals being "matched" and incorrectly regarded as a single individual.

We consider the case of false negatives occurring due to corrupted data entries (or incorrect decisions made at clerical review). For simplicity, to develop the modelling framework, we assume that only one source (source A, say) may result in corrupted data entries. We assume that if an unique identifier is corrupted it cannot be incorrectly matched to a different individual (i.e. the corrupted identifier does not match a non-corrupted identifier of a different individual leading to a potential false positive). Mathematically, we define the set $\mathcal{S} = \{0,1\}^3 \setminus \{0,0,0\}$. Then, the count in cell $\{1,j,k,l\}$, for $(j,k,l) \in \mathcal{S}$, is (in general) an undercount since some individuals observed by source A have not been correctly matched to the other sources (i.e. they are false negatives). Conversely, the cells $\{1,0,0,0\}$ and $\{0,j,k,l\}$, are (in general) overcounts, due to the false negatives occurring by failing to match individuals observed by A with other sources due to their identifier being corrupted.

The corresponding likelihood of the data is no longer a simple product of Poisson probability mass functions as for standard capture-recapture data (see Appendix A) since the

observed cell entries have been "corrupted". This means that the observed contingency table is composed of cells entries that are either (potentially) undercounts or overcounts. (The likelihood can be written as a sum over all possible combinations of true counts - however this is very computationally expensive and for large tables will be infeasible). We propose a Bayesian (data augmentation) approach. In particular, we consider the true cell counts as parameters to be estimated and form the joint posterior distribution over the log-linear parameters and true cell counts (including the unobserved cell count). The formal mathematical approach is described in Appendix C.

Simulation study

Within each simulated dataset we initially simulate complete contingency tables assuming the independent model, which we denote by $\boldsymbol{z}$. We then "corrupt" this true contingency table. We define the probability an individual observed by source A has their data entry corrupted so that they cannot be correctly matched to any of the other data sources to be denoted by ρ_A. For the moderate scenario, we assume that $\rho_A = 0.1$ and for the extreme scenario we set $\rho_A = 0.5$. Let n_{jkl} (for $(j,k,l) \in \mathcal{S}$) denote the number of individuals who have their data entry corrupted by source A and are unable to be matched to the corresponding sources, but were observed by the combination of sources (j,k,l) for sources B, C and D, respectively. Assuming each individual has their record corrupted by source A, independently of each other and which sources they are observed by, we simulate

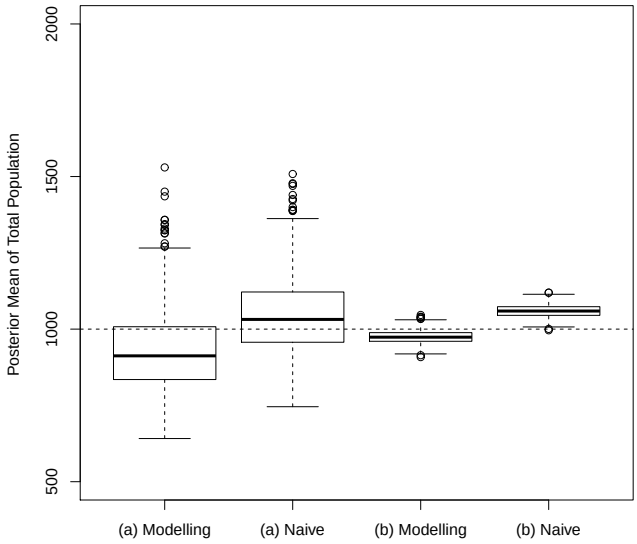
$$n_{jkl} \sim \text{Binomial}(z_{1jkl}, \rho_A),$$

for $(j,k,l) \in \mathcal{S}$ and for each simulated data set. Finally, the corresponding true cell entries, $\boldsymbol{z}$, are corrupted using equations (1), (2) and (4) in Appendix C to provide the observed contingency table entries, $\boldsymbol{y}$, (in the presence of false negatives).

Results

The Bayesian data augmentation approach described in Appendix C is implemented in order to (1) estimate the total population size and (2) infer the true cell counts (the log-linear parameters/individual data source capture probabilities are also estimated within the algorithm). We focus on the estimates of the total population size. For comparison we also fit the standard independent log-linear model to the observed contingency tables (ignoring the presence of false negatives). We refer to this approach as *naive*. Conversely, we refer to the approach that directly models the presence of false negatives as *modelling*. Figure 1 provides boxplots of the estimated posterior mean of the total population size for the 1000 simulated contingency tables for the naive and modelling approaches for (a) extreme and (b) moderate scenarios. It is immediate that the posterior means for the modelling approach generally appear to underestimate the total population size. For example, the median of the posterior means for the extreme and moderate scenarios have an overall bias of -88 (or -9%) and -26 (or -3%). Unsurprisingly, the bias is smaller under the moderate scenario (with higher capture probability and lower corruption) than the extreme scenario; and the variance of the posterior means is smaller under the moderate scenario when compared to the extreme scenario. Further, the coverage of the 95% highest posterior density intervals (HPDIs) for the total population size are 86.4% and 90.1%, for the extreme and moderate scenarios, respectively.

Figure 1: Boxplot of the 1000 posterior means of the total populations sizes from the simulated contingency tables for false negatives under the (a) extreme and (b) moderate scenarios using the modelling approach or naive approach (ignoring the presence of false negatives). In each case the true total population size is 1000.



Conversely, the naive approach appears to perform better for the extreme scenario, but worse for the moderate scenario. This is most likely due to the fact that for the extreme scenario, fewer individuals are observed - on average only 35% are observed, and in particular only 10% of individuals are observed by source A and less than 3% observed by source A and at least one other source. So that despite a relatively high corruption rate (of 50%), there is less data to be corrupted (on average there will be less than 14 false negatives). For the moderate scenario, the naive approach appears significantly worse than when the false negatives are explicitly modelled, with the true population size less than the lower 5% quantile of estimated population means. Note that for the moderate scenario, there are a larger number of false negatives due to an increased number of observed individuals - on average 87% of individuals are observed with 40% observed by source A and 31% are observed by source A and at least one other source. This leads to an average of 31 false negatives within the data (with a corruption rate of 10% for source A).

The false negative model does appear to lead to a small negative bias (the strength dependent on the underlying capture and corruption probabilities). However, the proposed model may be able to be improved in two particular aspects:

1. Within the model-fitting process a Uniform distribution is currently specified on the number of "corrupted" individuals (or false negatives) in each relevant contingency table cell. It should be possible to extend the model-fitting process to explicitly model more complex distributions leading to additional parameter(s) being estimated. (Note that there are similarities here with regard to issues associated with non-target individuals being observed).

2. Sub-sampling (or prior information) may be able to be introduced (again akin to ideas discussed for non-target individuals) in order to provide improved estimates of the corruption probabilities that can be formally incorporated within the model.

(ii) Partially collated data

Here we consider data sources that are only partially collated, leading to some individuals unable to be matched between sources. For example, consider the case where sources A and B have not been matched. Consequently not all cells are observable. For example, suppose that an individual is observed by source A but is not observed by either source C or D. Then it is not possible to match this individual to those observed by only source B (and vice versa). Alternatively if an individual is observed by source A and sources C and/or D, then this individual can be matched to individuals observed by source B since source B has been matched to sources C and D. In other words

- Cells $\{1,0,0,0\}$, $\{0,1,0,0\}$ and $\{1,1,0,0\}$ are not directly observable; instead the combined cells $(\{1,0,0,0\}+\{1,1,0,0\})$ and $(\{0,1,0,0\}+\{1,1,0,0\})$ are observed.

In order to fit the log-linear model(s) to the data we use a Bayesian data augmentation approach, described in Appendix D. This approach can be extended to allow for different combinations of sources being collated and further, where only a proportion of individuals have been collated between two sources (assuming that it is known which individuals have been matched and which have not, leading to (additional) lower bounds on the true cell entries).

Simulation study

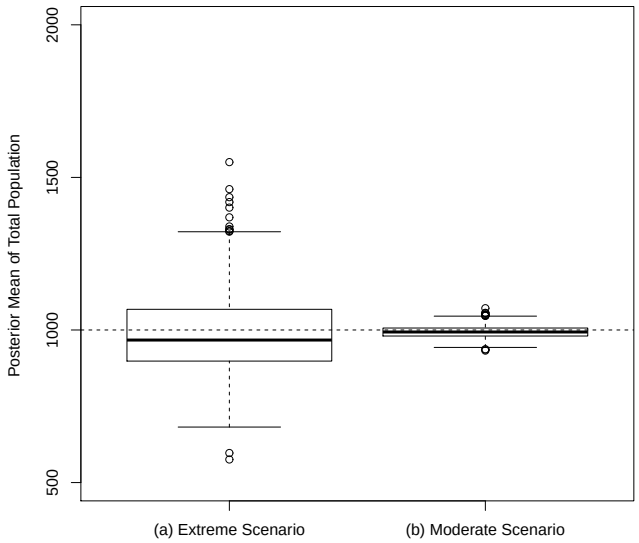
We consider the same 1000 contingency tables generated in the previous section for the moderate and extreme scenarios, but without corrupting the data (i.e. without incorporating false negatives). For these contingency tables, the observed data are the true cell entries excluding cells $\{0,0,0,0\}$, $\{1,0,0,0\}$, $\{0,1,0,0\}$ and $\{1,1,0,0\}$, but including the combined cell entries $(\{1,0,0,0\}+\{1,1,0,0\}$ and $(\{0,1,0,0\}+\{1,1,0,0\})$.

Results

The Bayesian (data augmentation) approach for fitting the model with partially collated data is provided in Appendix D. This approach once again permits the estimation of not only the total population size but also the true cell entries for $\{1,0,0,0\}$, $\{0,1,0,0\}$ and $\{1,1,0,0\}$ (and the log-linear parameters). Figure 2 provides a boxplot of the posterior means of the total population size for the 1000 simulated contingency tables under each scenario.

The true value of the total population size appears to be generally well estimated for both scenarios (though possibly some minor underestimation). The median of the posterior means lies close to the true value which is within the inter-quartile range of the estimated posterior means for each scenario. The median for the bias of the posterior mean is -32 (or -3.2%) or -6 (or -0.6%) for the extreme and moderate scenarios, respectively. Again, we can observe the smaller variability in the posterior mean for the total population size for the moderate scenario compared to the extreme scenario. Further, the coverage of the

Figure 2: Boxplot of the 1000 posterior means of the total populations sizes from the simulated contingency tables for partially collated data under the (a) extreme and (b) moderate scenarios. In each case the true total population size is 1000.



95% HPDIs for the true population size are 90.9% and 91.2% for the extreme and moderate scenarios, respectively.

(iii) Non-target individuals

Finally, we consider the problem where an individual may be observed by a source, yet the individual themselves is not a member of the target population of interest. For example, for GP records, an individual may be recorded on this data source, yet may not be a resident in the population (e.g. an individual may have visited the GP while on holiday). We consider the case where only a single source may observe a combination of target and non-target individuals (see also Overstall *et al* 2012). Without loss of generality, we assume that source A observes both target and non-target individuals; while each of the other sources B, C and D only observe members of the target population. This implies that:

- An individual observed by source A and any other source is a member of the target population (since only members of the target population are observed by sources B, C and D).

- Cell $\{1,0,0,0\}$ is an overcount of the number of target individuals only observed by source A, (as this cell contains a mixture of target and non-target individuals).

- Assuming that all individuals observed by source A are members of the target population (when they contain a mixture of target and non-target individuals) will lead to an over-estimate of the total population size.

The likelihood of the data relating to all cells, except cell $\{1,0,0,0\}$ are simply a product of Poisson probability mass functions. The final likelihood term, corresponding to cell $\{1,0,0,0\}$ can be written as a sum over the Poisson probability mass function for the true

number of target individuals in the given cell and the corresponding function specified on the *censoring* of the given cell (i.e. the distribution of the observed cell, given the true cell entry and model parameters). We consider a Bayesian data augmentation approach, where we consider the true cell count for cell $\{1, 0, 0, 0\}$ to be a parameter to be estimated, in addition to the log-linear parameters and total population size (or alternatively unobserved cell entry). The formal mathematical approach is described in Appendix E.

Simulation study

We use the same 1000 simulated true contingency tables under the extreme and moderate scenarios as for the previous studies, but then adjust these data to allow for non-target individuals to be observed by only source A. In particular we assume that the observed cell corresponding to individuals that are only observed by source A is double that of the true cell count. In other words, 50% of the individuals that are only observed by source A correspond to non-target individuals (and hence 50% are members of the target population).

Results

Initially we assume that there is no information relating to the proportion of non-target individuals (i.e. there is no subsampling) that are observed by source A (so that we assume a Uniform censoring distribution - see Appendix E). We term this approach *censoring*. For comparison, we re-analyse the observed contingency tables ignoring the fact that source A observes non-target individuals, assuming that all individuals observed are members of the target population. This approach is termed *naive*. A summary of the posterior results for the naive and censoring approaches are given in Figure 3 (the left-most plot in the upper and lower panels) for the extreme and moderate scenarios. Clearly, the naive approach consistently (and significantly) overestimates the total population size, whereas the censoring approach appears to generally provide consistent estimates (median bias for the posterior means of -15 (-1.5%) and -6.9 (-0.69%) for the extreme and moderate scenarios, respectively). Further, the corresponding coverage probabilities for the 95% HPDIs are given in Figure 4 (again the left-most values under "no subsampling" provides correspond to the naive and censoring approaches). The coverage probabilities under both scenarios are close to the nominal 95%, but the naive approach leads to very poor coverage probabilities. This again clearly demonstrates the effect on the total population size in the presence of non-target individuals being present within the data (and significant bias that is introduced if non-target individuals are treated as members of the target population).

We now extend the modelling approach to consider the case where additional information may be collected via subsampling the individuals observed by only source A (though this could be extended to observed by source A and not those *only* observed by source A). In particular we assume that the individuals only observed by source A are subsampled and their *status* determined (i.e. whether they are a member of the target population or not). This extends the above approach, in more than one way. Firstly, lower and upper bounds can be placed on the number of individuals only observed by source A (i.e. cell $\{1, 0, 0, 0\}$), since some individuals observed by only source A are identified as members of the target population or not. Incorporating these additional bounds and retaining the Uniform censoring distribution is referred to as "uninformative" censoring. Further, alternative, informative,

Figure 3: Boxplot of the 1000 posterior means of the total populations sizes from the simulated contingency tables under the various approaches applied to the non-target individual problem under the (a) extreme and (b) moderate scenarios. In each case the true total population size is 1000. Note that the y-axes are different for the extreme and moderate scenarios.

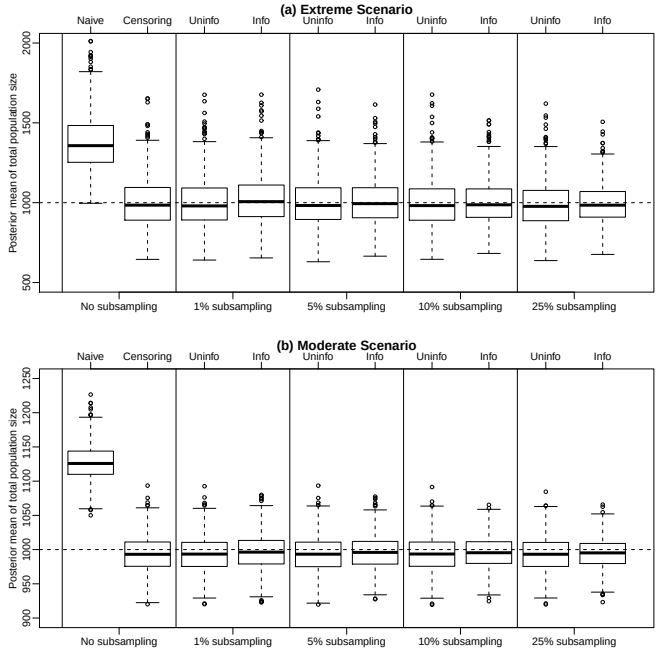


Figure 4: Coverages of the 95% highest posterior density intervals for the total population size under the various approaches applied to the non-target individual problem under the (a) extreme and (b) moderate scenarios.

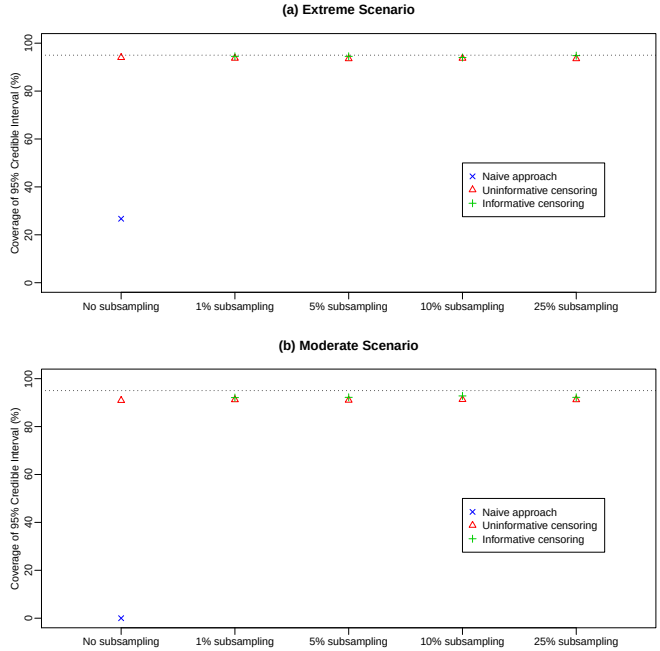
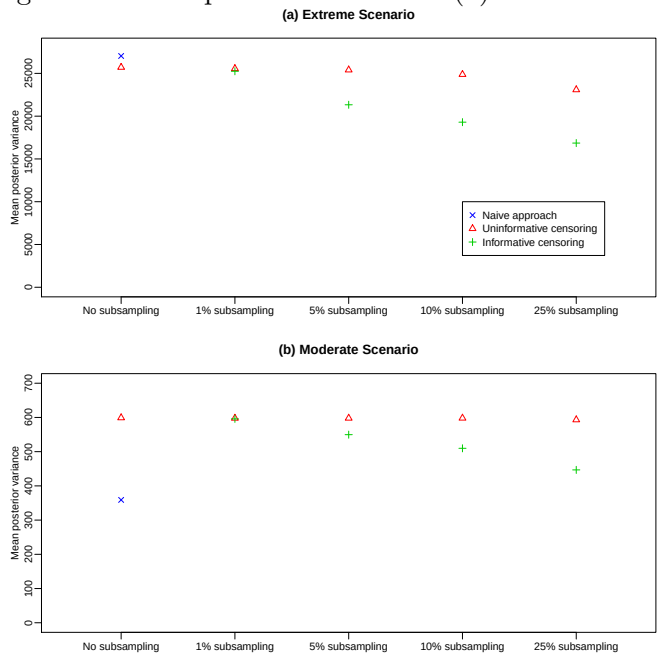


Figure 5: Mean posterior variance of the total population size under the various approaches applied to the non-target individual problem under the (a) extreme and (b) moderate scenarios.



distributions can be specified on the distribution of the observed cell entries given the true cell entries, such as the Negative-Binomial distribution, assuming that each individual is independently a member of the target population with some probability, to be estimated (see Appendix E for further details). Incorporating both the lower and upper bounds and the Negative-Binomial distribution on the censoring distribution is referred to as "informative" censoring. We consider a range of scenarios corresponding to subsampling 1%, 5%, 10% and 25% of those individuals only observed by source A.

The upper and lower panels of Figure 3 show the corresponding boxplots of the posterior means of the total population size for the simulated datasets for the non-informative and informative censoring approaches under differing subsampling rates and scenarios. In each case the posterior means appear to be generally clustered about the true value of 1000. The coverage probabilities for the 95% HPDIs lie within 93-95% for the extreme scenario and 91-93% for the moderate scenario, with consistently (very) slightly higher coverage probabilities for the informative censoring compared to the uninformative censoring (see Figure 4). Finally, Figure 5 shows the mean posterior variance for each case (including the naive approach under no subsampling). From this figure we can see that, unsurprisingly, as the amount of subsampling increases, the posterior variance of the total population size decreases, due to the increased level of information within the data. However, the decrease in variance is fastest when assuming informative subsampling. Thus, subsampling is able to improve the precision of the posterior estimates (with smaller 95% HPDIs), without adversely affecting the coverage probabilities of these credible intervals.

The modelling approach derived considers non-target individuals being observed by only one source. The approach (and model-fitting ideas) can be extended to allow for non-target members being observed by multiple sources - but further model development will be needed.

In such cases, it is envisaged that subsampling may have an even greater impact on the estimates of the total population size, since the observed data will contain less data, with an increased number of cells containing non-target individuals. For example, if non-target individuals can be observed by both sources A and B, then a total of 3 cells (namely, $\{1,0,0,0\}$, $\{0,1,0,0\}$ and $\{1,1,0,0\}$) will contain a mixture of target and non-target individuals.

Discussion

The report describes a small feasibility study into three different problems in relation to capture-recapture data - (i) false negatives; (ii) partially collated data; and (iii) non-target individuals. Each problem is taken independently of each other and "simple" cases considered, i.e. (i) false negatives for only one sources; (ii) two sources not collated; and (iii) non-target individuals observed by only one source. Statistical models are developed for each of these cases which can be fitted within a Bayesian (data augmentation) framework and simulation studies conducted to assess their performance. Overall, the models developed seem to perform reasonably well, although they could be developed further. For example, for false negatives, more informative distributions can be used and subsampling undertaken to improve the performance of the models in estimating total population sizes. In particular, modelling ideas presented for non-target individual problem may be able to be incorporated within the modelling framework. We note that it is not surprising that the false negatives problem results in the poorest estimates of total population size within the simulation studies, since for this problem, no observed cell entries are assumed to be the true values. In contrast for the problem relating to non-target individuals observed by one source, there is only one observed cell entry that is assumed not to be equal to the true cell entry.

The modelling approaches described here can be developed further. For example, extending the number of sources that can observe non-target individuals (and thus increasing the number of censored cells). Similarly, more than one source may have "corrupted" identifiers leading to additional forms of false negatives within the contingency tables. In addition, it is possible to combine these different problems, such as false negatives and non-target individuals being observed. The frameworks for such problems will build on those presented here, but may lead to additional issues (and computational algorithms) needing to be addressed. Note that for such models, the use of subsampling may be particularly important, due to the increasing complexity of the models, and "weaker" form of data due to the increased uncertainty in the true cell entries. Finally, only the independent log-linear model was considered. This can be extended to consider more general log-linear models, allowing for interactions between sources. "Standard" model selection (and model-averaging) tools are available to address the issue of model choice and estimation of total population size in the presence of model uncertainty See for example King and Brooks 2001 and Overstall *et al* 2012 in relation to log-linear models. However, additional issues should be investigated, including general parameter identifiability for more complex models to ensure that models being fitted are identifiable.

References

Brooks, S. P. (1998), Markov Chain Monte Carlo Method and its Application. *The Statistician* **47**, 69–100

Chao, A., Tsay, P. K., Lin, S., Shau, W. and Chao, D. (2001), The application of capture-recapture models to epidemiological data. *Statistics in Medicine* **20** 2123–57.

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

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

Hook, E. B. and Regal, R. R. (1999), Recommendations for presentation and evaluation of capture-recapture estimates in epidemiology *Journal of Clinical Epidemiology* **52** 917–926.

Hook, E. B. and Regal, R. R. (2000), On the need for a 16th and 17th recommendation for capture-recapture analysis *Journal of Clinical Epidemiology* **52** 1275–1277.

King, R. and S. P. Brooks (2001), On the Bayesian Analysis of Population Size. *Biometrika* **88**, 317–336

Overstall, A. M., King, R., Bird, S. M., Hutchinson, S. J. and Hay, G. (2012), Estimating the number of people who inject drugs in Scotland using multi-list data with left censoring. *University of St Andrews Technical Report.*

Appendices

A Standard log-linear models

Log-linear models are often fitted to capture-recapture data to estimate total population size (Fienberg 1972). Here, for notational simplicity we assume that there are a total of 3 sources, denoted by A, B and C (though note that within the simulation studies 4 sources are used). Data are typically presented in the form of an incomplete 2^3 contingency table, with cells $\{i, j, k\}$ for $i, j, k \in \{0, 1\}$, where 0/1 denotes absence/presence in each data source, respectively. We let y_{ijk} denote the number of individuals observed in cell $\{i, j, k\}$ corresponding to being observed/not observed by the given combination of sources. For example, y_{010} denotes the number of individuals only observed by source B. The cell y_{000} corresponding to the number of individuals in the population not observed by any of the sources is unknown. The set of all cells is denoted by $\boldsymbol{y} = \{y_{ijk} : i, j, k \in \{0, 1\}\}$.

A log-linear model is typically fitted such that,

$$y_{ijk}|\lambda_{ijk} \sim \text{Poisson}(\lambda_{ijk}),$$

independently for each i, j, k and where λ_{ijk} is of log-linear form. For example, the independent model is specified in the form,

$$\log \lambda_{ijk} = \theta + \theta_i^A + \theta_j^B + \theta_k^C,$$

where θ corresponds to the underlying mean and θ_i^A, θ_j^B and θ_k^C main effect terms for each distinct source (to allow for the probability of being observed by each source to be different). The independent model assumes that being observed by any source is independent of whether the individual is observed by any of the other data sources. Constraints are specified on the main effect terms for identifiability. In particular, we adopt the sum-to-zero constraints, so that $\theta_0^A + \theta_1^A = 0$, and similarly for all other main effect terms. An alternative specification of the log-linear form often used is to specify,

$$\log \lambda_{ijk} = \boldsymbol{x}_{ijk}^T \boldsymbol{\theta},$$

where $\boldsymbol{\theta}$ is a column vector corresponding to the identifiable log-linear parameters and $\boldsymbol{x}_{ijk}$ the corresponding design vector (i.e. the column of the design matrix) linking the identifiable parameters with λ_{ijk}. We note that for when using sum-to-zero constraints, each element of $\boldsymbol{x}_{ijk}$ is equal to ± 1.

We note that the addition of interaction terms removes the independent assumption. For example, the saturated model (for incomplete contingency tables) is specified such that,

$$\log \lambda_{ijk} = \theta + \theta_i^A + \theta_j^B + \theta_k^C + \theta_{ij}^{AB} + \theta_{ik}^{AC} + \theta_{jk}^{BC},$$

where θ_{ij}^{AB} corresponds to the interaction between sources A and B, and similarly for θ_{ik}^{AC} and θ_{jk}^{BC}. For identifiability, sum-to-zero constraints are again applied, such that, for example, $\theta_{00}^{AB} + \theta_{01}^{AB} = 0 = \theta_{00}^{AB} + \theta_{10}^{AB} = \theta_{01}^{AB} + \theta_{11}^{AB}$. A value of $\theta_{11}^{AB} > 0$ corresponds to a positive interaction, so that being observed by source A leads to an increased probability of being observed by source B (and vice versa); a value of $\theta_{11}^{AB} < 0$ corresponds to a negative

interaction, so that being observed by source A leads to an decreased probability of being observed by source B (and vice versa). Different log-linear models fitted to the data can lead to (vastly) different estimates for the total population size, so that model selection can be very important (model-averaging can be implemented to account for both parameter and model uncertainty). However, for simplicity within the simulation studies we consider the independent model.

Notationally we let $\boldsymbol{y}$ denote the contingency table cell entries. The parameters to be estimated are $\boldsymbol{\phi} = \{\boldsymbol{\theta}, y_{000}\}$, where $\boldsymbol{\theta}$ denotes the set of identifiable log-linear parameters. The joint probability mass function of the contingency table cell entries can be expressed as a product of independent Poisson probability mass functions, such that,

$$\pi(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i,j,k} \left[\frac{\exp(-\lambda_{ijk})\lambda_{ijk}^{y_{ijk}}}{y_{ijk}!} \right].$$

For further discussion see for example Overstall *et al* (2012).

We note that an alternative specification to the Poisson distribution is the multinomial (see for example, King and Brooks 2001). We let the total population size be denoted by,

$$N = \sum_{i,j,k} y_{ijk}.$$

An equivalent model specification is,

$$\boldsymbol{y}|N, \boldsymbol{\theta} \sim \text{Multinomial}(N, \boldsymbol{p}),$$

where $\boldsymbol{p} = \{p_{ijk} : i, j, k \in \{0, 1\}\}$. These probabilities are specified such that,

$$p_{ijk} = \frac{p^*_{ijk}}{\sum_{i,j,k} p^*_{ijk}},$$

where,

$$\log p^*_{ijk} = \theta^A_i + \theta^B_j + \theta^C_k.$$

Note that the intercept term θ is no longer necessary (essentially this term cancels within the calculation of the cell probabilities, $\boldsymbol{p}$, and is "replaced" by the parameter for total population size, N). Within this model specification, a prior on the total population size, N can be immediately incorporated within the Bayesian analysis.

B Bayesian Statistics (Brief Outline)

Suppose that we observe data $\boldsymbol{y} = \{y_1, \ldots, y_n\}$ on which we wish to make inference on the parameters $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_p\}$. The Bayesian paradigm can be described as follows:

1. Before observing any data there is some independent *prior* information relating to the parameters $\boldsymbol{\theta}$ - this is represented by the prior distribution on the parameters, $\pi(\boldsymbol{\theta})$.

2. Observe data $\boldsymbol{y}$ from some underlying system with given statistical model expressed as a function of the (unknown) parameters $\boldsymbol{\theta}$ - this is represented by the probability mass/density function of the data given the parameters $\pi(\boldsymbol{y}|\boldsymbol{\theta})$ (this is often referred to as the *likelihood* function).

3. In light of the observed data, update our initial prior beliefs, to form our *posterior* beliefs of the data (using Bayes' Theorem) - this is the posterior distribution, $\pi(\boldsymbol{\theta}|\boldsymbol{y}) \propto \pi(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$.

Notes:

- The prior distribution needs to be specified *independently* of observing the data.
- The parameters have a distribution (rather than regarded as having a fixed value as in classical statistics), and this distribution is typically summarised via summary statistics, such as:
  - Posterior mean/median - providing a point estimate in relation to the location of parameters;
  - Posterior variance/standard deviation - a point estimate of the spread of the distribution of each parameter;
  - Credible intervals (CIs) - an uncertainty interval providing an indication of the spread of the distribution for a given parameter (e.g. a symmetric 95% CI provides the lower and upper 2.5% quantiles of the posterior distribution; the 95% highest posterior density interval (HPDI) the shortest 95% credible interval);
  - Marginal density plots of each parameters - providing a graphical representation of the marginal distribution of the given parameter;
  - Posterior correlation between parameters - providing an estimate of the relationship between two parameters.
- For all log-linear models considered within the simulation studies, we assume uninformative priors. In particular, we specify independent $N(0, \sigma^2)$ priors, such that $\sigma^2 \sim \Gamma^{-1}\left(\frac{a}{2}, \frac{b}{2}\right)$, where $a = b = 0.001$.

The posterior distribution is typically too complex to obtain inference directly and computational techniques are used to (indirectly) sample from the posterior distribution and obtain estimates of the posterior summary statistics of interest. The most common computational algorithm used (and used within this feasibility study) is Markov chain Monte Carlo (MCMC). The basic idea of MCMC is to construct a Markov chain with stationary distribution equal to the posterior distribution of interest. The Markov chain is then run until the stationary distribution has been reached, so that further realisations of the Markov chain can be regarded as a (dependent) sample from the posterior distribution of interest. These realisations can be used to obtain estimates of the summary statistics of interest (e.g. posterior mean of a parameter is estimated as the sample mean of the parameter from the given realisations of the Markov chain). For further details on using MCMC for log-linear models, see for example, Overstall *et al* (2012).

C Bayesian Approach in the Presence of False Negatives

We consider the presence of false negatives, such that the unique identifier of an individual observed by source A may be corrupted so that it cannot be matched to other sources

correctly. When there are 4 data sources there are $2^4 = 16$ cells in the incomplete contingency table with cells $\{i, j, k, l\}$, such that $i, j, k, l \in \{0, 1\}$. We let $\boldsymbol{y} = \{y_{ijkl} : i, j, k, l \neq \{0, 0, 0, 0\}\}$ denote the set of *observed* contingency table cells. Further, we let $\boldsymbol{z} = \{z_{ijkl} : i, j, k, l \in \{0, 1\}\}$ denote the set of *true* contingency table cells. Recall that we set $\mathcal{S} = \{0, 1\}^3 \setminus \{0, 0, 0\}$. Finally, we let n_{jkl} for $(j, k, l) \in \mathcal{S}$ denote the number of individuals that are observed by source A that have not been correctly matched to the individuals observed by the combination of sources (j, k, l) for sources B, C and D, respectively (i.e. the true number of false negatives for each combination of (j, k, l)).

Using the model considerations for false negatives, we have that for $(j, k, l) \in \mathcal{S}$,

$$\begin{aligned} z_{1jkl} &= y_{1jkl} + n_{jkl}, \quad (1) \\ z_{0jkl} &= y_{0jkl} - n_{jkl}. \quad (2) \end{aligned}$$

It follows immediately, that,

$$z_{0jkl} + z_{1jkl} = y_{0jkl} + y_{1jkl}. \quad (3)$$

The true number of individuals observed only by source A, denoted z_{1000} is equal to the number of individuals recorded as only observed by source A, less the total number of false negatives (i.e. falsely recorded as only being observed by source A, when they were observed by at least one other source but unmatched due to their identifier being corrupted by source A). Mathematically,

$$z_{1000} = y_{1000} - \sum_{(j,k,l) \in \mathcal{S}} n_{jkl}. \quad (4)$$

The total number of individuals observed by source A is unaffected by false negatives, so that

$$y_{1000} + \sum_{(j,k,l) \in \mathcal{S}} y_{1jkl} = z_{1000} + \sum_{(j,k,l) \in \mathcal{S}} z_{1jkl}. \quad (5)$$

In other words the total number of individuals observed by source A in the observed contingency table, $\boldsymbol{y}$ is equal to the total number of individuals observed by source A in the true contingency table, $\boldsymbol{z}$.

We make the standard modelling assumption for the true contingency table and assume that for $i, j, k, l \in \{0, 1\}$,

$$z_{ijkl} \sim \text{Poisson}(\lambda_{ijkl}),$$

with the usual log-linear form

$$\log \lambda_{ijkl} = \boldsymbol{x}_{ijkl}^T \boldsymbol{\theta},$$

where $\boldsymbol{\theta}$ are the identifiable log-linear parameters and $\boldsymbol{x}_{ijkl}$ the design vector which links the elements of $\boldsymbol{\theta}$ with the given cell. Both $\boldsymbol{\theta}$ and true contingency table cell entries, $\boldsymbol{z}$, are unknown. Thus we treat $\boldsymbol{z}$ as parameters (or auxiliary variables) to be estimated and using Bayes' theorem form the joint posterior distribution over the log-linear parameters, $\boldsymbol{\theta}$, and true cell entries, $\boldsymbol{z}$, given by,

$$\begin{aligned} \pi(\boldsymbol{z}, \boldsymbol{\theta} | \boldsymbol{y}) &\propto \pi(\boldsymbol{y} | \boldsymbol{\theta}, \boldsymbol{z}) \pi(\boldsymbol{\theta}, \boldsymbol{z}), \\ &= \pi(\boldsymbol{z} | \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) \pi(\boldsymbol{y} | \boldsymbol{z}). \end{aligned}$$

Note that for simplicity we assume a Uniform distribution for $\pi(\boldsymbol{y}|\boldsymbol{z})$, given the constraints specified by (3) and (5). For an alternative (informative) distribution, see Appendix E for ideas that can be applied to these data (including subsampling).

Given the joint posterior distribution, $\pi(\boldsymbol{z}, \boldsymbol{\theta}|\boldsymbol{y})$, we can sample from the distribution using standard MCMC methods. (We note that the full conditional distribution for some elements of $\boldsymbol{z}$ is degenerate. Specifically, if we know the elements of $\boldsymbol{z}$ corresponding to not being observed by source A, then we know the elements corresponding to being observed by source A, from (3)). For further general discussion of MCMC methods see for example, Brooks (1998) and King and Brooks 2001 and Overstall *et al* 2012 for the particular application to incomplete contingency table data.

D Bayesian Approach for Partially Collated Data

We consider 4 data sources that are only partially collated, such that sources A and B have not been matched. Consequently not all cells are observable, instead sums of cells are observed. In this case the sum of the cells $(\{1,0,0,0\} + \{1,1,0,0\})$ and $(\{0,1,0,0\} + \{1,1,0,0\})$ are observed, rather than each of the individual cells. Let $\boldsymbol{z}$ denote the true cell counts, where $\boldsymbol{z}_O$ are the observed cell counts and $\boldsymbol{z}_U = (z_{0000}, z_{1000}, z_{0100}, z_{1100})$ are the missing or "partially" observed cell counts. Let $\boldsymbol{y}_S = (z_{1000} + z_{1100}, z_{0100} + z_{1100})$ be the sum of the partially collated cell counts. We again assume that

$$z_{ijkl}|\lambda_{ijkl} \sim \text{Poisson}(\lambda_{ijkl}),$$

with the usual log-linear form

$$\log \lambda_{ijkl} = \boldsymbol{x}_{ijkl}^T \boldsymbol{\theta},$$

where $\boldsymbol{\theta}$ are the identifiable log-linear parameters and $\boldsymbol{x}_{ijkl}$ the corresponding design vector for cell (i,j,k,l). We treat $\boldsymbol{z}_U$ as parameters and form the joint posterior distribution of $\boldsymbol{\theta}$ and $\boldsymbol{z}_U$ using Bayes' theorem, given by,

$$\begin{aligned}
\pi(\boldsymbol{z}_U, \boldsymbol{\theta}|\boldsymbol{z}_O, \boldsymbol{y}_S) &\propto \pi(\boldsymbol{z}_O, \boldsymbol{y}_S|\boldsymbol{\theta}, \boldsymbol{z}_U)\pi(\boldsymbol{\theta}, \boldsymbol{z}_U), \\
&= \pi(\boldsymbol{z}_O|\boldsymbol{\theta}, \boldsymbol{z}_U, \boldsymbol{y}_S)\pi(\boldsymbol{y}_S|\boldsymbol{\theta}, \boldsymbol{z}_U)\pi(\boldsymbol{z}_U|\boldsymbol{\theta})\pi(\boldsymbol{\theta}), \\
&= \pi(\boldsymbol{z}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\pi(\boldsymbol{y}_S|\boldsymbol{z}_U),
\end{aligned}$$

where $\pi(\boldsymbol{z}|\boldsymbol{\theta}) = \pi(\boldsymbol{z}_O|\boldsymbol{\theta})\pi(\boldsymbol{z}_U|\boldsymbol{\theta})$ is the complete data likelihood. The constraints on the elements of $\boldsymbol{z}_U$ given by $\boldsymbol{y}_S$ enter the posterior distribution through $\pi(\boldsymbol{y}_S|\boldsymbol{z}_U)$.

We can again sample from this posterior using MCMC methods. The full conditional distribution for some elements of $\boldsymbol{z}_U$ is degenerate. Specifically, given z_{1100}, then we can determine z_{1000} and z_{0100}.

E Bayesian Approach in the Presence of Non-Target Individuals

We consider the case where non-target individuals may be observed by (only) one source. We denote this source by source A. All other sources only observe members of the target population. We let $\boldsymbol{y} = \{y_{ijkl} : ijkl \neq \{0,0,0,0\}\}$ denote the set of *observed* cell entries and

$\boldsymbol{z} = \{z_{ijkl} : \{i,j,k,l\} \in \{0,1\}^4\}\}$ the *true* cell entries (for the target population). Thus cell z_{0000} corresponds to the number of individuals in the target population but unobserved by each of the sources; and z_{1000} the true number of individuals observed by only source A. We have that $y_{\boldsymbol{k}} = z_{\boldsymbol{k}}$ for all $\boldsymbol{k} \neq \{0,0,0,0\}, \{1,0,0,0\}$. Further, $z_{1000} \leq y_{1000}$ (i.e. the observed cell entry is an upper bound of the true cell entry, allowing for non-target individuals to be observed by source A). We note that we can regard cell $\{1,0,0,0\}$ as being a *censored* cell. Finally, for notational convenience we let $\boldsymbol{z}_O = \{y_{ijkl} : i,j,k,l \neq \{0,0,0,0\}, \{1,0,0,0\}\}$, denoting the set of observed true individuals; $y_C = y_{1000}$, the observed upper bound for the number of individuals in the target population only observed by source A; $z_C = z_{1000}$ the true number of individuals in the target population only observed by source A and $z_U = z_{0000}$.

We consider the joint posterior distribution of the log-linear parameters, $\boldsymbol{\theta}$, unobserved cell entry, z_U, and censored cell entry z_C, given by,

$$\begin{aligned}\pi(\boldsymbol{\theta}, z_U, z_C|\boldsymbol{z}_O, y_C) &\propto \pi(\boldsymbol{z}_O, y_C|\boldsymbol{\theta}, z_U, z_C)\pi(\boldsymbol{\theta}, z_U, z_C) \\ &= \pi(\boldsymbol{z}_O|y_C, \boldsymbol{\theta}, z_U, z_C)\pi(y_C|\boldsymbol{\theta}, z_U, z_C)\pi(z_U, z_C|\boldsymbol{\theta})p(\boldsymbol{\theta}).\end{aligned}$$

However, we note that $\boldsymbol{z}_O$ is independent of y_C, z_U and z_C, given log-linear parameters $\boldsymbol{\theta}$. Thus,

$$\pi(\boldsymbol{z}_O|y_C, \boldsymbol{\theta}, z_U, z_C)\pi(z_U, z_C|\boldsymbol{\theta}) = \pi(\boldsymbol{z}|\boldsymbol{\theta}),$$

and is the standard likelihood function for log-linear models (i.e. a product over Poisson probability mass functions). Thus, the posterior distribution can be written in the form,

$$\pi(\boldsymbol{\theta}, z_U, z_C|\boldsymbol{y}_{true}, y_C) \propto \pi(\boldsymbol{z}|\boldsymbol{\theta})\pi(y_C|\boldsymbol{\theta}, z_U, z_C)\pi(\boldsymbol{\theta}),$$

where $\pi(y_C|\boldsymbol{\theta}, z_U, z_C)$ denotes the model specification on the censored cell; and $\pi(\boldsymbol{\theta})$ the prior on the log-linear parameters. We note that additional subsampling of the data may be undertaken. In particular, we consider subsampling of those individuals that are only observed by source A to determine whether or not they are members of the target population. Notationally, let s_A denote the number of individuals that are subsampled from those individuals that are only observed by source A; such that of these individuals m_A are recorded as being members of the target population (so that $s_A - m_A$ are not members of the target population). We then consider two different specifications for the censored cell, corresponding to (i) non-informative censoring (with and without subsampling); and (ii) informative censoring (with subsampling). We discuss each in turn.

(i) Non-informative censoring

Assuming non-informative censoring (without subsampling), so that the true censored cell provides no information on the observed censored cell, we have that,

$$y_C|z_C \sim U[z_C, \infty),$$

i.e. the distribution for the total number of individuals observed in the cell is Uniform with a lower bound corresponding to the true number of target individuals observed only by source A. In other words there is an (uninformative) Uniform distribution on the number of non-target individuals observed by source A, i.e. $(y_C - z_C) \sim U[0, \infty)$.

In the presence of (uninformative) subsampling, further information is available regarding the number of individuals that are only observed by source A that are members (or not members) of the target population. This means that there is further information on the bounds of the (conditional) Uniform distribution of y_C. In particular we have that,

$$\begin{aligned} y_C|q_A, z_C, m_A, s_A &\sim \mathrm{U}\left[0, z_C + s_A - m_A\right], \\ m_A|z_C, s_A, q_A &\sim \mathrm{U}\left[0, \min(z_C, s_A)\right], \end{aligned}$$

(since the number of subsampled individuals that are observed to be members of the target population is at most the number sampled, s_A, or the true number of individuals observed by source A, z_C).

(ii) Informative censoring

We now consider informative censoring, assuming additional information is available. In particular, we consider additional data collected via subsampling of those individuals that are only observed by source A and determining whether or not they are members of the target population. Notationally, let s_A denote the number of individuals that are subsampled from those individuals that are only observed by source A; such that of these individuals m_A are recorded as being members of the target population (so that $s_A - m_A$ are not members of the target population). Thus the observed data corresponds to $\{\boldsymbol{z}_O, y_C, s_A, m_A\}$. The joint posterior distribution can be written in the form,

$$\pi(\boldsymbol{\theta}, q_A, z_U, z_C|\boldsymbol{z}_O, y_C, s_A, m_A) \propto \pi(\boldsymbol{z}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\pi(y_C|q_A, z_C, m_A, s_A)\pi(m_A|z_C, s_A, q_A)\pi(q_A).$$

The term $\pi(\boldsymbol{z}|\boldsymbol{\theta})$ corresponds to the joint probability density of the true contingency table cells given the log-linear parameters and is again of standard form (a product over Poisson likelihood terms). Assuming that each individual only observed by source A is a member of the target population with probability q_A, independently of each other, we can consider an informative censoring distribution. In particular,

$$\begin{aligned} y_C|q_A, z_C, m_A, s_A &\sim \mathrm{Neg-Bin}(z_C, q_A)\mathrm{I}\left(y_C \geq z_C + s_A - m_A\right), \\ m_A|z_C, s_A, q_A &\sim \mathrm{Bin}(s_A, q_A)\mathrm{I}\left(m_A \leq z_C\right), \end{aligned}$$

where I is the indicator function and in this context defines a truncated distribution. We note that the negative binomial distribution can be used for the probability distribution of the number of trials given the probability of success and the number of successes.