A linkage error correction model for population size estimation with multiple sources

Daan Zult[1], Peter – Paul de Wolf[1], Bart Bakker[1,2], Peter van der Heijden[3]
[1]Statistics Netherland[1]
[2]VU University
[3]Utrecht University and University of Southampton

Abstract
A new method is described to do population size estimation, while linkage of sources occurs with errors. Our model is derived from a linkage error correction model introduced by Ding and Fienberg (1994). They show how to use linkage probabilities to correct the capture - recapture estimator for linkage errors, but only in the case of two sources and no covariates. A generalisation is proposed by incorporating the Ding & Fienberg model into the standard log - linear modelling approach used in multiple - recapture estimation. We show how the method performs in a simulation study with data that resemble real data.

Keywords
Multiple – recapture estimation; population size estimation; capture – recapture; record linkage; linkage errors

1. Introduction

This paper is a summary of Zult et al. (2019), which we refer to for a more extensive and elaborate discussion of this topic. The size of a partly observed population is often estimated with the capture – recapture (CR, for two sources) or multiple – recapture (MR, for multiple sources) method. An important assumption for these models is that records in different sources can be identified such that it is known whether these records belong to the same unit or not, i.e. records can be perfectly linked between sources. This assumption of perfect linkage is of particular relevance if identification is not obtained by some perfect identifier (like a tag or id-code) but by indirect identifiers (like name and address). In that case record are usually linked with probabilistic linkage (see Fellegi and Sunter, 1969, Winkler, 1988 or Jaro, 1989) and the perfect linkage assumption is often violated which generally leads to a biased population size estimate (PSE) (Gerritse et al., 2017).

A solution to this problem was provided by Ding and Fienberg (1994) (DF), Di Consiglio and Tuoto (2015) (DC&T_15) and De Wolf et al. (2018) (DW). These authors show how to use linkage probabilities to correct the capture -

[1] The authors like to thank Jan van der Laan from Statistics Netherlands for his review of the final version of this the paper.

recapture estimator for linkage errors. Recently, Di Consiglio and Tuoto (2018) (DC&T_18) extended their method to three sources.

In this paper we provide a general framework that allows us to extend this work further in two ways, with covariates and multiple sources. This is done by generalising the standard log - linear modelling approach used in multiple - recapture estimation such that it incorporates linkage error correction. This leads to the weighted multiple – recapture (WMR) model and is discussed in section 2. In section 3 we show the results of a simulation study that tests the WMR model.

2. Methodology

We first introduce some formal notation. s defines the source, where in standard CR = (1,2) and in MR = (1,2, ...). Next, we define the linked 'register' R_{t-1} as:

$$R_{t-1} = \begin{cases} R_{0=}\, S_1 \\ R_{1=}\, L_1(S_1, S_2) \\ R_{2=}\, L_2(R_1, S_3) \\ \quad \vdots \\ R_{t-1=}\, L_t(R_{t-2}, S_t) \end{cases},$$

where R_t refers to a set of t + 1 sequentially linked sources and L_t refers to the linkage process that links R_{t-1} with S_{t+1}. In case of CR this reduces to $R_1 = R = L(S_1, S_2)$. The *true* cell counts, *estimated* cell counts and *observed* cell counts (i.e. the counts of records that are linked and not linked between R_{t-1} and S_{t+1}) are denoted as $m_{ij} = (m_{11}, m_{10}, m_{01})$, $\hat{m}_{ij} = (\hat{m}_{11}, \hat{m}_{10}, \hat{m}_{01})$ and $n_{ij} = (n_{11}, n_{10}, n_{01})$ respectively. Here $i \in \{1,0\}$ corresponds to records in and not in R_{t-1} and $j \in \{1,0\}$ corresponds to records in and not in S_{t+1}. When there are no linkage errors, the true cell counts are equal to the observed cell counts, i.e. $m_{ij} = n_{ij}$. Furthermore, we define $m^*_{ij} = (m^*_{11}, m^*_{10}, m^*_{01})$ and $n^*_{ij} = (n^*_{11}, n^*_{10}, n^*_{01})$ as the true and observed cell counts in a random sample from R_{t-1} called a rematch or audit study (for a discussion on the difference between rematch and audit sample, which is small, we refer to Zult et al. (2019)). Beside that m^*_{ij} refers to a subsample, the difference between m^*_{ij} and m_{ij} is that in the presence of linkage errors m^*_{ij} is assumed to be known while m_{ij} is not. Finally, we introduce $p = 1, \ldots, P^t$ which are the records in R_t. Under perfect linkage this implies that all records refer to unique units/individuals, but in case of linkage errors two records in R_t might belong to different units/individuals or one record in R_t might represent two or more units.

The derivation of the WMR model follows three steps. First the D&F model is written as log – linear Poisson regression model. Second, the dependent variable in this model is corrected for linkage errors in case of two sources but

with covariates. These two steps are discussed in section 2.1. Third, this model is extended towards multiple – sources, which is discussed in section 2.2.

2.1 Capture - recapture estimation and linkage error correction

In the most basic case of CR the PSE is given by the standard Petersen (Petersen, 1986, Lincoln, 1930) formula:

$$\widehat{M}_{Petersen} = m_{11} + m_{10} + m_{01} + \frac{m_{10}m_{01}}{m_{11}} = \frac{(m_{11}+m_{10})((m_{11}+m_{01})}{m_{11}} = \frac{m_{1+}m_{+1}}{m_{11}} \quad (1),$$

where under the appropriate assumptions $\widehat{M}_{Petersen}$ is an unbiased estimate of the true population size (Wolter, 1986). The Petersen estimator is closely related to a fitted value obtained from a log - linear Poisson regression model with cell counts data (e.g. see Cormack, 1989), i.e.:

$$E[m_{ij}] = e^{(\beta_0 + \beta_1 i + \beta_2 j)} \text{ for } i, j \in \{1,0\} \quad (2),$$

where m_{ij} serves as the dependent variable in the log - linear regression model. The Poisson regression model uses maximum likelihood to obtain estimates $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$. An important difference between equation (1) and (2) is that (2) can be easily extended with additional sources or categorical covariates.

When the appropriate assumptions are not met, for instance records are not perfectly linked, $\widehat{M}_{Petersen}$ is biased. Therefore D&F developed a linkage error correction method that uses a rematch study from which they calculate the linkage error probabilities that are used to correct the PSE for linkage errors. DW show that this correction method can be written as:

$$\widehat{M}_{D\&F} = \frac{m_{1+}m_{+1}}{\hat{m}_{11}} \quad (3),$$

where $\hat{m}_{11}$ is the estimated number of links between both sources that takes linkage errors into account. Combining equation (1), (2) and (3) allows us to write:

$$E[\hat{m}_{ij}] = e^{(\beta_0 + \beta_1 i + \beta_2 j)} \text{ for } i, j \in \{1,0\} \quad (4)$$

where $\hat{m}_{11} = n_{11}\frac{m_{11}^*}{n_{11}^*}$, $\hat{m}_{10} = n_{1+} - \hat{m}_{11}$ and $\hat{m}_{01} = n_{+1} - \hat{m}_{11}$ (see Zult et al. (2019) for a more extensive derivation). In words, equation (4) constitutes the same model as equation (2), except the dependent variable m_{ij} is replaced by $\hat{m}_{ij}$, where $\hat{m}_{ij}$ is simply a vector of estimated cell counts that is based on the results of the audit study. Here we should note that the calculation of $\hat{m}_{11}$ is independent of the exact linkage procedure L . In fact, the only thing that matters is that the fraction $\frac{m_{11}^*}{n_{11}^*}$ is a consistent estimate of $\frac{m_{11}}{n_{11}}$, which implies that the audit study should be representative for R.

Equation (4) allows for the inclusion of covariates in the same way as in a regular log - linear Poisson regression, which implies that $\hat{m}_{ij}$ must be separated further into groups (e.g. male/female) and this categorical covariate can be added to the regression equation. We refer to this extension of the D&F model as the weighted CR (WCR) model. Why it is called 'weighted' will become clear in the next section.

2.2 The weighted – multiple recapture model

In section 2.1 we showed how the D&F model can be written as a log – linear Poisson regression model and how (categorical) covariates can be added to this equation by splitting - up $\hat{m}$ into smaller groups. This implies that after this procedure we have for each cell count both an estimated and observed cell count. Here we should note that each cell count consists of records, so for each record we can calculate its weighted contribution to its estimated cell count, i.e.:

$$w_p = \frac{\hat{m}_p}{n_p} \tag{5}$$

where $\hat{m}_p$ and n_p refer to the estimated and observed cell count of record p. E.g., when we ignore covariates and record p is linked between S_1 and S_2 , $\hat{m}_p$ and $n_p = 11$. Now w_p is a record level weight that sums up to the different elements in $\hat{m}$. Adding up over w_p is similar to the case of no linkage errors where each record has a weight of 1 and is added up to obtain (the true and observed) cell counts. However, when we want to extend the model such that it can deal with multiple – sources, we can write w_p as:

$$w_p^t = w_p^{t-1} \frac{\hat{m}_p^t}{n_p^t} \tag{6},$$

with $w_p^{t=0} = 1$, $n_p^t = \sum_{p \epsilon cell\ count} w_p^{t-1}$. Under equation (6) w_p^t is updated after every linkage procedure, which can be repeated for each new source. After the update of w_p^t the estimated cell count elements of $\hat{m}$ can be calculated by summing up w_p^t over the records p that belong to that cell, where $\hat{m}$ does not only distinguish between i and j but may distinguish between any number of sources and categorical covariates. The WMR model can then be written as:

$$[\hat{m}_{zt}] = e^{f(\beta, Z_t)} \tag{7},$$

where $\hat{m}_{zt}$ is the estimated cell count vector that depends on $Z_t = (R_{t-1}, X)$ with a set of categorical covariates, according to some function $f(\beta, Z_t)$ with β a parameter vector.

3. Results

We evaluate the WMR model with a simulation study. In this study the true population size (TPS) is known and will be compared with estimates of the population size. We use a (quasi – real) dataset that is a publicly available fictitious population dataset of 26 625 persons that is representative for the UK population census. It was created in a European project on data integration (McLeod, Heasman and Forbes, 2011) that ran from 2009 to 2011. The dataset has linkage keys such as address and birthdate but also covariates such as gender and age. By generating sources from this quasi - real dataset, outcomes may reflect reality to some extent.

The main goal of this simulation study is to evaluate the performance of the WMR model. The WMR model is applied within different scenarios, where scenarios differ with respect to three elements:

1. Covariate dependence of capture probabilities, which implies that the probability of a record to be in S_1, S_2 and S_3 may vary due to differences in the covariate values of records (e.g. a male may have a higher probability to be in S_1 and a lower probability to be in S_2).
2. Source dependence of capture probabilities, which implies that the probability of a record to be in S_1, S_2 and S_3 may depend on this record being in another source (e.g. a record in S_1, may have a different probability to be in S_2 than a record that is equal in all other aspects except being in S_1).
3. Linkage errors in the linkage procedure; sources are linked either with errors or are linked perfectly without errors.

These three elements are of particular interest, because they are the sources of bias that the WMR model aims to correct for while the alternative models should suffer from at least one of them. They lead to four different scenarios that can be seen in table 1.

Table 1: Simulation study scenarios.

Scenario	Linkage errors	Covariate dependence	Source dependence
1	Yes	No	No
2	Yes	Yes	No
3	Yes	No	Yes
4	Yes	Yes	Yes

Each scenario is replicated 1 050[2] times and in each replication a population of 10 000, together with three sources of approximately 8 000, 5 000 and 2000 records is generated, where the generation of sources differs between

[2] The number is 'only' 1 050 because we use a spark cluster of fifteen cores (available at Statistics Netherlands mainly for Big Data related computations) that each does 70 replications with different random seeds, in which each single replication takes about 10 minutes. In total it takes almost two days to run all four scenarios, which is mainly due to the computation time of the probabilistic linking the three sources.

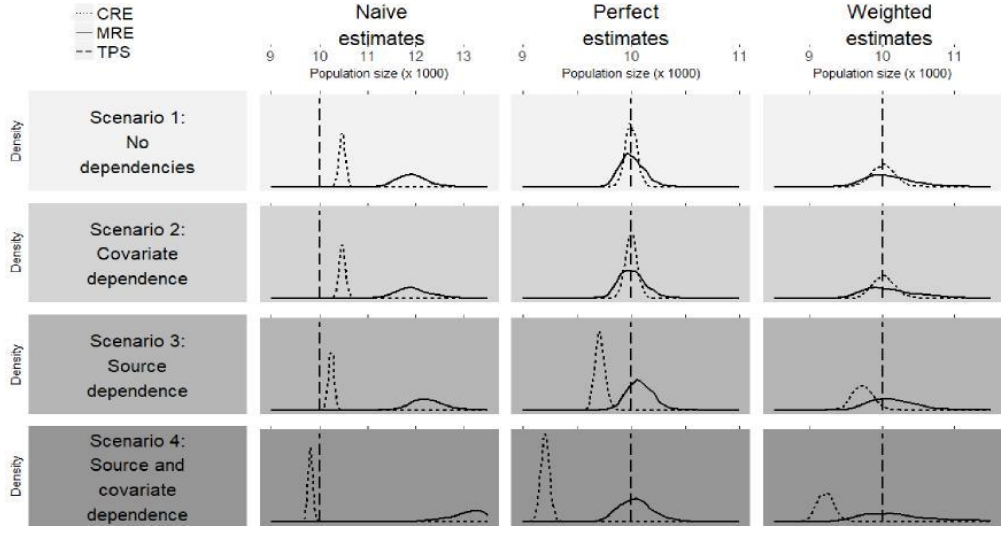
scenarios. For further details on the simulation setup we refer to Zult et al. (2019), in the next section we discuss the results.

3.1 Simulation outcome

In figure 1 below the simulation results of the four scenarios are presented as density plots where each density plot contains the CR estimate (CRE), MR estimate (MRE) and TPS that are calculated in three different ways, i.e. naïve (with linkage errors and without correction), perfect (without linkage errors and without correction) and weighted (with linkage errors and with correction). The results are in figure 1 on the next page.

Ideally the density of an estimate revolves around the TPS of 10 000. However, the first column shows that the densities of the naïve estimates do not, which implies that the linkage errors indeed lead to biased PSEs. Furthermore, in case of perfect linkage, in scenario 1 and 2 both the CREs and MREs revolve around the TPS. However, when source dependence is introduced in scenario 3 and 4 the CR model (necessarily) fails while the MR model still performs well. This failure of the CR model implies that it suffers from source dependence as intended by the simulation setup. Finally, the third column contains the weighted estimates. Here the (weighted) CR model performs well in scenario 1 and 2, which implies the WCR model is able to correct for both linkage errors and covariate dependence simultaneously. However, in scenario 3 and 4 the CR model logically fails, because it is unable to deal with source dependence. Fortunately, in these scenarios the density of the MREs still revolves around the TPS, which implies that the WMR model indeed corrects for linkage errors, covariate dependence and source dependence simultaneously.

Figure 1: Density plots of two PSEs with three dependent variables and four scenarios (table 1).



4. Discussion and Conclusion

In this paper we derived and tested the WMR model for population size estimation corrected for linkage error. The model is derived from the D&F model and is a more general extension than the models developed by DC&T (2015, 2018) and De Wolf et al. (2018) because it can deal with three or more sources and covariates. Furthermore, the WMR model is incorporated in the more general family of log - linear regression models and therefore no longer has to be studied as an isolated issue in CR and MR models. Finally, the WMR model was tested and approved in a simulation study.

In theory the WMR model might be an improvement on the D&F model, they both still require the availability of a rematch (for D&F) or audit (for WMR) study. The advantage of the WMR model is that an audit study might be easier to obtain, because it has lower requirements (it needs to be constructed on the cell count level instead of the much more detailed records matching pair level). However, the incorporation of covariates and additional sources in the WMR model also puts additional constraints on the audit study, in the sense that the audit study should include these same covariates and additional sources. Given that the sample that underlies the audit study must be representative for R^t, this might be more difficult for increasing t.

Also, we should note that we paid little attention to the impact of the exact linkage procedure. In section 2 we developed the WMR model in the context of the common sequential linkage approach, in which first two sources are linked and a third source is linked to this combined source. However, it is also possible that sources are linked pairwise or simultaneously. These approaches are less common because they suffer either from computational (i.e. the number of potential matches between multiple sources increases exponentially) or methodological (e.g. what to do with inconsistent matching patterns like $A \rightarrow B,\ B \rightarrow C,\ C \nrightarrow A$?). Furthermore, in the simulation study of section 3 we applied probabilistic linkage that uses techniques developed by Fellegi and Sunter (1969), Winkler (1988) and Jaro (1989) that aim to optimise the quality of matches on the matching pair level, while matching techniques that are designed to optimise the quality of the matches on the cell count level might already significantly reduce the problem of linkage errors in population size estimation.

Another point that deserves some discussion is the 'individual starting weight of 1'. Lists or registers of individuals sometimes also contain individual sample weights, which indicate the size of the group that this individual represents as part of the total population. There is no reason why these sample weights cannot replace the starting weights of 1 in the WMR model. Furthermore, when additional sources also contain sample weights they can be used to calculate n^t, n^{*t} and m^{*t} in a slightly different way, i.e. simply by adding up sample weights instead of counting. This way we would get 'linkage

error corrected sample weights'. However, we should note that the presence of sample weights usually implies that the source only covers a (very) small part of the population, so when multiple sources contain sample weights the probability of matches becomes low, leading to very low cell counts and an unreliable PSE. How exactly sample weights can be combined with linkage and linkage error correction requires further research.

References

1. Cormack, R. M. (1989). Log-linear models for capture-recapture. Biometrics, 45, 395 – 413.
2. De Wolf, PP., Van Der Laan, J. and Zult, D. (2018). Joining correction methods for linkage error in capture-recapture, 45, Discussion paper, Statistics Netherlands, The Hague/Heerlen. Available at: https://www.cbs.nl/en-gb/background/2018/18/connecting-correction-methods-for-linkage-error-in-crc. To appear in Journal of Official Statistics, September 2019.
3. Di Consiglio, L. and Tuoto, T. (2015). Coverage evaluation on probabilistically linked data. Journal of Official Statistics, 31, 415 – 429.
4. Di Consiglio, L. and Tuoto, T. (2018). Population Size Estimation and Linkage Errors: the Multiple Lists Case. Journal of Official Statistics, Vol. 34, No. 4, 2018, pp. 889–908.
5. Ding, Y. and Fienberg, S.E. (1994). Dual system estimation of Census undercount in the presence of matching error. Survey Methodology, 20, 149 – 158.
6. Fellegi, I. P. and Sunter, A. B. (1969). A Theory for Record Linkage. Journal of the American Statistical Association, 64, 1183 – 1210.
7. Gerritse, S.C., Bakker, B.F.M. and Van der Heijden, P.G.M. (2017). The impact of linkage errors and erroneous captures on the population size estimator due to implied coverage. Discussion paper 2017 - 16, Statistics Netherlands, The Hague/Heerlen. Available at: https://www.cbs.nl/en-gb/background/2017/39/impact-of-linkage-errors-and-erroneous-captures
8. Jaro, M. (1989). Advances in Record Linkage Methodology as Applied to Matching the 1985 Test Census of Tampa, Florida. Journal of American Statistical Association 84: 414–420.
9. McLeod, P., Heasman, D. and Forbes, I. (2011). Simulated data for the on the job training. Essnet DI. Available at: http://www.cros-portal.eu/content/job-training.
10. Lincoln, F. C. (1930). Calculating Waterfowl Abundance on the Basis of Banding Returns, U.S. Dept. Agric., 118: 1-4.

11. Petersen, C.G.J. (1896). The yearly immigration of young plaice into the Limfiord from the German Sea. Report of the Danish Biological Station, 6, 5 – 84.
12. Winkler, W. E. (1988). Using the EM algorithm for weight computation in the Fellegi - Sunter model of record linkage. Section on Survey Research Methods, 667 – 671.
13. Wolter, K.M. (1986). Some coverage error models for census data. Journal of the American Statistical Association, 81, 338 – 346.
14. Zult, D.B.,. De Wolf, PP., Bakker, B.F.M. and van der Heijden, P.G.M. (2019). A general framework for multiple - recapture estimation that incorporates linkage error correction. Discussion paper 2019 - 12, Statistics Netherlands, The Hague/Heerlen. Available at: https://www.cbs.nl/en-gb/background/2019/19/correcting-for-linkage-errors-in-the-multiple-capture