Discussion paper

A general framework for multiple - recapture estimation that incorporates linkage error correction

Daan Zult
Peter-Paul de Wolf
Bart Bakker
Peter G.M. van der Heijden

May 2019

A general framework for multiple - recapture estimation that incorporates linkage error correction[1]

Daan Zult[2], Peter - Paul de Wolf[2], Bart Bakker[3] and Peter G.M. van der Heijden[4]

Abstract

The size of a partly observed population is often estimated with the capture – recapture (for two sources) or multiple – recapture (for multiple sources) estimation method. An important assumption of 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 or animal's skin patterns). In that case the perfect linkage assumption is often violated, which in general leads to biased population size estimates. A solution to this problem was provided by Ding and Fienberg (1994), Di Consiglio and Tuoto (2015) and De Wolf et al. (2018). These authors show how to use linkage probabilities to correct the capture - recapture estimator for linkage errors. Recently, Di Consiglio and Tuoto (2018) 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. First, we extend this work further to any number of sources. Second, our framework allows to incorporate covariates in a better way. We do this by generalising the standard log - linear modelling approach used in multiple - recapture estimation such that it incorporates linkage error correction. We show how the method performs in a simulation study with data that resemble real data.

Keywords: dual – system estimation, multiple – recapture estimation, population size estimation, capture - recapture, record linkage, linkage errors.

1. Introduction

Capture – recapture (CR) estimation and multiple – recapture (MR) estimation provide a standard approach to estimate the size of the unobserved part of a population (Petersen, 1896, Fienberg, 1972, Bishop et al., 1975). These models are also known

[1] Daan Zult is the corresponding author, db.zult@cbs.nl. The others are collaborative authors. Bart Bakker and Peter van der Heijden are also research supervisors. The authors like to thank Jan van der Laan for his review of the final version of this the paper.

[2] Statistics Netherlands

[3] Statistics Netherlands and VU University

[4] Utrecht University and University of Southampton

under other names, such as dual - system, multiple – system and mark – recapture estimation (see e.g. IWGDMF, 1995). CR estimation uses two sources while in order to correct for source dependence / heterogeneous catchability, MR estimation uses three or more sources (e.g. Fienberg, 1972). Here, a source refers to a set, list or register of records in which each record represents a unit or individual that belongs to the target population, where no unit is represented more than once in each source. Furthermore, in general sources do not cover the full target population, so there are units that belong to the target population but are not represented by any of the records in any of the sources. Both CR and MR models can be written as a specific formulation of a log – linear model, which connects them to the general framework of log – linear Poisson regression analysis. One of the main assumptions that underlie the capture – recapture estimate (CRE) and multi – recapture estimate (MRE) is that records can be accurately identified over sources as being the same unit or not. If not, there is a non - zero probability that records will be falsely linked (a mismatch), or falsely not linked (a missed match) and the resulting population size estimate (PSE) *may* be biased (Wolter, 1986, Chao, 2001, Chen and Kuo, 2001, Cadwell, 2005, Gerritse et al., 2017). We write ‘may’ because in theory both types of linkage errors may cancel out each other, which we will elaborate on later. First we provide some notation and give an example of linkage error bias.

Imagine there are two sources S_1 and S_2 that are linked by some linkage procedure L into a combined source R, denoted as $R = L(S_1, S_2) = [S_1, S_2]$. Each (population) unit may then be (captured) only in S_1, only in S_2, in both or in neither of the sources (they are unobserved). The occurrences of these four events can be represented as counts, which we refer to as ‘cell counts’. We denote the true cell counts as $m = (m_{11}, m_{10}, m_{01})$ that are the cell counts that would occur without linkage errors. We denote the observed cell counts $n = (n_{11}, n_{10}, n_{01})$ that are the cell counts that are observed after linkage. In case of perfect linkage $n = m$ and in case of non - perfect linkage, which generally implies records are linked based on probabilities, they may differ. Elements in m and n are indicated as m_{ij} and n_{ij} where $i \in \{1,0\}$ corresponds to records in S_1 and not in S_1 and $j \in \{1,0\}$ corresponds to records in S_2 and not in S_2 respectively. Furthermore, i and j can both have the value $+$, where $i = +$ indicates all records in S_2 and $j = +$ indicates all records in S_1. This notation implies that the cell counts m_{00} and n_{00} are both unobserved and $n_{1+} = m_{1+}$ and $n_{+1} = m_{+1}$ are equal because they represent the size of the sources S_1 and S_2, which are unaffected by L. Finally we refer to cells indexed by $ij \in \{11,10,01\}$ as linkage cells.

As an illustration of this notation and linkage error bias we introduce a simple example. When $n_{1+} = m_{1+} = 300$, $n_{+1} = m_{+1} = 150$ and $m_{11} = 100$ and due to linkage errors the observed cell count is $n_{11} = 90$. This implies that the number of missed links is 10 more than the number of false links. This case is represented in table 1.

Table 1: Example of true and observed cell counts table of two sources.

Linkage cell		Cell count	
i	j	m	n
1	1	100	90
1	0	200	210
0	1	50	60

When all the CR model assumptions (see Wolter, 1986) are met, an unbiased CRE for the unobserved part of the population m_{00} can be obtained by: $E[m_{00}|m] = \frac{m_{10}*m_{01}}{m_{11}} = \frac{200*50}{100} = 100$, an estimator that is also known as the Petersen (1896) or Lincoln - Petersen (Lincoln, 1930) estimator. However, due to linkage errors not m but n is observed and when this is naively ignored the CRE becomes: $E[n_{00}|n] = \frac{n_{10}*n_{01}}{n_{11}} = \frac{210*60}{90} = 140$, leading to a linkage error bias of 40%, something better not left ignored.

Linkage errors are especially prone to occur if a perfect identifier that perfectly identifies units over different sources is not available. In this case records may be linked by means of indirect identifiers called linkage keys, such as surname, address, animal skin patterns or a combination of different identifiers. Linkage models that use such keys to link records in different sources are referred to as probabilistic linkage models (e.g. see Fellegi and Sunter, 1969, Winkler, 1988 or Jaro, 1989), which in general come with non - zero probabilities of false links. In general these probabilistic linkage models are designed to minimise the number of false matches. In practice this implies there is a threshold probability that tells the model what linkage probability is acceptable in linking two records, which in practice is often set at 50%. This implies, for instance, when two individuals are both in two sources and are only known under their name, probabilistic linkage will fail to link any of these records, because each single match would have at least a 50% probability of being false, which is too high. This implies that instead of adding 2 to n_{11}, the probabilistic linkage model will add 2 to n_{10} and 2 to n_{01}, even when a person's name is a strong identifier. From the CR perspective this result is unfortunate, because when name is a strong identifier, it is quite likely that these four records represents the same two units and are better added to n_{11}.

In order to correct for linkage error bias, Ding and Fienberg (1994) (D&F) propose a linkage error correction method. The basic idea behind their method is that when the probabilities of mismatches and missed matches are known, they can be used to correct the CRE for linkage errors. D&F show how these probabilities can be calculated with the help of a (small) study, which is a survey that takes a random portion of the matches and non - matches in the two sources under clerical review, hereby establishing which matches were correct and which were not. The drawback of the D&F method is twofold. First, the D&F method does not explicitly concern covariates and second the correction can only be applied to the CRE and not the MRE. Because the implications of these drawbacks are not straightforward we will discuss both of them in more detail in section 3.1 and 3.2.

The model that is presented in this paper is both a simplification and a generalisation of the D&F model. It is simpler in the sense that instead of calculating probabilities of true and false matches on the record level it evaluates the accuracy on the cell counts level. This simplification allows us to easily generalize the CR and MR with the extension of multiple sources and covariates. This general framework is referred to as the weighted multiple – recapture (WMR) model. This model owns its name to the record level weights that are updated each time a new source is linked. The sums of these record weights replace the dependent variable in the specific log – linear regression formulation of the CR and MR model. This replacement leads to the weighted capture – recapture estimate (WCRE) and weighted multiple – recapture estimate (WMRE), which are corrected for linkage errors.

The outline of this paper is as follows. In section 2 we discuss the classic CRE or Petersen estimator and its relation to the D&F model and later developments by Di

Consiglio and Tuoto (2015, 2018) (DC&T_15, DC&T_18) and De Wolf et al. (2018) (DW). In particular the theoretical results by DW we use to derive the WMR model. In section 3 we derive the WCRE and WMRE, in section 4 we present a simulation study to show they work and in section 5 we conclude and discuss the results.

2. Capture - recapture estimation

In this section we describe and discuss CR models and the linkage error correction method introduced by D&F, which we will use in section 3 to derive the WCRE and WMRE. We first describe the most basic CRE which was introduced by Petersen (1896) and next show how D&F correct this estimate for linkage errors. We further discuss DC&T_15, DC&T_18 and DW, because they provide a deeper understanding of the correction method. This is useful because it provides the intuition that helps to understand the derivation of the WCRE and WMRE.

2.1 The Petersen estimator

Under the appropriate assumptions (Wolter, 1986), including perfect linkage, a CRE can be obtained by the standard Petersen formula:

$$\widehat{M}_{\text{Petersen}} = n_{11} + n_{10} + n_{01} + \frac{n_{10}n_{01}}{n_{11}} = \frac{(n_{11}+n_{10})(n_{11}+n_{01})}{n_{11}} = \frac{n_{1+}n_{+1}}{n_{11}} \qquad (1),$$

where $\widehat{M}_{\text{Petersen}}$ is an estimate of the true population size M based on the observed cell counts n. 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.:

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

where m_{ij} serves as the dependent variable in the log - linear regression model. The Poisson regression model uses maximum likelihood to obtain estimates $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$. The unknown portion of the population is represented by the unobserved m_{00} of which an estimate can be obtained by $E[n_{00}|\hat{\beta}_0] = e^{\hat{\beta}_0} = \frac{n_{10}n_{01}}{n_{11}}$. This equality illustrates how equation (1) and (2) lead to the same result. However, an important difference is that the log – linear formulation in equation (2) can be easily extended with additional sources or categorical covariates and the interaction between them. For instance, let

S_3 be a third source in which presence is indicated by $k \in \{0,1\}$ and let x be a categorical covariate with levels $x \in \{0,1\}$. Then an example of a model is:

$E\left[n_{ijkX}\right] = e^{(\beta_0+\beta_1 i+\beta_2 j+\beta_3 k+\beta_4 ij+\beta_5 ik+\beta_6 jk+\beta_7 x+\beta_8 xi)}$, where e.g. ij is the product of i and j. Extending the Petersen formula in this way would be non - trivial at best, while for each value in x a PSE of the unobserved population can be relatively easily obtained, i.e.:

$E\left[n_{0000}|\hat{\beta}_0\right] = e^{\hat{\beta}_0}$ and $E\left[n_{0001}|\hat{\beta}_0, \hat{\beta}_7\right] = e^{\hat{\beta}_0+\hat{\beta}_7}$.

Our reason to emphasize on the relation between the Petersen estimator and the Poisson regression is that in the next section we will use this relation to extend the D&F estimator in the same way.

2.2 The D&F model

The D&F model uses a rematch study to estimate probabilities of specific linkage errors. This is a survey that is assumed to be representative for R and where, after sources were probabilistically linked, a subset of records are put under further scrutiny and it is checked whether these records were correctly matched or not. The outcome of such a study can be summarized as in table 2.

Table 2: Rematch study with D&F structure.

		Rematch study	
		Matched	**Not matched**
Probabilistic linkage	**Matched**	a_{11}	a_{10}
	Not matched	a_{01}	a_{00}

In table 2 we see how many records that are in the rematch study were correctly matched (a_{11}), correctly not matched (a_{00}), incorrectly matched (a_{10}) and incorrectly not matched (a_{01}). Based on this rematch study, D&F define a probability of a missed link $1 - \alpha$ and a probability of a mismatch θ by $\alpha = \frac{a_{11}}{a_{11}+a_{01}}$ and $\gamma = \frac{a_{10}}{a_{10}+a_{00}}$. With the help of these probabilities D&F define their so - called linkage error corrected CRE, i.e. $\widehat{M}_{\mathrm{D\&F}}$. The D&F model recently received more attention from DC&T_15 and DW. DC&T_15 write the $\widehat{M}_{\mathrm{D\&F}}$ as:

$$\widehat{M}_{\mathrm{D\&F}} = \frac{n_{11}+n_{10}+n_{01}}{\hat{p}_1+\hat{p}_2-(\alpha-\theta)\hat{p}_1\hat{p}_2-\theta\hat{p}_1} \quad (3)$$

with $\hat{p}_1 = \frac{-n_{11}+\theta(n_{11}+n_{10})}{(\theta-\alpha)(n_{11}+n_{01})}$, $\hat{p}_2 = \frac{-n_{11}+\theta(n_{11}+n_{10})}{(\theta-\alpha)(n_{11}+n_{10})}$ and the observed cell counts n as defined in table 1. These equations show that the D&F model is complex and hard to interpret. The formulas become more complex when DC&T_15 introduce their so called two - way linkage errors. DW further extend this by allowing the sources to be of different size (i.e. θ becomes (θ_1, θ_2)). For this purpose they propose so called asymmetrical two – way errors that use the size of the second source as additional parameter. In their 2018 paper, DC&T_18 extend their linkage error correction model from two to three sources also by using α and θ. They introduce a so called transition matrix that allows one to transform the observed cell counts into estimates of the true cell counts, which can serve as input for the Poisson regression. This is in itself a useful extension on their earlier model, but it is still limited in the sense that the method is not generic with respect to covariates and it is unclear how to add yet an additional source. Fortunately, beside DW's asymmetrical two – way errors extension, DW provide us with another useful contribution. They show that in fact the D&F model, the DC&T_15 model and their own extension all give identical outcomes when not only the formulas of $\widehat{M}$ but also of α and θ are chosen appropriately. In fact, they show that all three models can be written much more comprehensively as:

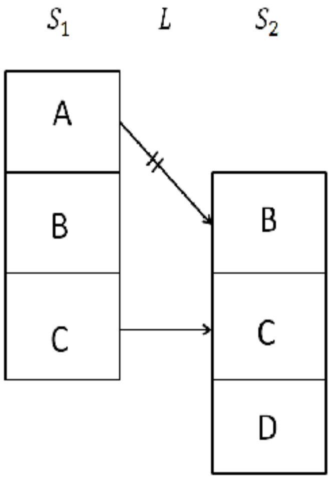
$$\widehat{M}_{\mathrm{D\&F}} = \widehat{M}_{\mathrm{DC\&T\_15}} = \widehat{M}_{\mathrm{DW}} = \frac{m_{1+}m_{+1}}{\widehat{m}_{11}} = \frac{n_{1+}n_{+1}}{\widehat{m}_{11}} \quad (4),$$

where we define $\widehat{m} = (\widehat{m}_{11}, \widehat{m}_{10}, \widehat{m}_{01})$ as *estimates* of the true cell counts m which are the observed cell counts n corrected for linkage errors. Equation (4) shows that the D&F, the DC&T_15 and the DW model are all equal and generalisations of the Petersen estimator, in which the observed cell count n_{11} is replaced by $\widehat{m}_{11}$. Implicitly or explicitly, D&F, DC&T_15, DC&T_18 and DW all derive an expression for $\widehat{m}_{11}$ where $\widehat{m}_{11}$ depends on n, α and θ. However, in the next section we will show that $\widehat{m}_{11}$ can be derived in a more simple and straightforward way, that no longer depends on α and θ altogether. This becomes clear when in the next section we reconstruct the rematch study into a structure that we will refer to as the 'audit study' structure.

2.3 Reformulation of the D&F model

In order to reformulate the D&F model into a more simple and generalizable model, we first note that D&F evaluate linkage errors on the record level, while CR uses cell counts as input. As discussed by D&F, DC&T_15, DC&T_18 and DW, on the record level there may be several types of linkage errors. In contrast, on the cell count level there is basically only one, i.e. the observed count n_{11} may be either too large or too small. Also because n_{1+} and n_{+1} are insensitive to linkage errors, n_{11} automatically gives n_{10} and n_{01} as well. The difference in complexity between linkage errors on the record and cell counts level is illustrated with a simple example in figure 1.

Figure 1: Simple example of two sources and linkage procedure.



In figure 1 we see two sources S_1 and S_2 that are linked with some linkage procedure L. S_1 contains units A, B and C while S_2 contains units B, C and D. We see that the first record in S_1, unit A, is falsely matched to the first record in S_2, unit B. Also the second record in S_1, unit B, is falsely not matched to the first record in S_2. In other words, on the record level there are two types of linkage errors. On top of these linkage errors, there is also a correct match of unit C and a correct non – match of unit D, which are also taken into account by D&F's model. However, when we would look at this example from the cell counts level perspective, we see that both the observed cell counts n and true cell counts m are equal, i.e. $n_{11} = m_{11} = 2$, $n_{10} = m_{10} = 1$ and $n_{01} = m_{01} = 1$, so on the cell level there are no linkage errors at all. This implies linkage errors on the record level may cancel out on the cell counts level. When we extend this cell counts point of view to the D&F method, we can restructure the rematch study in table 2 into a more simple structure that only contains observed and true cell counts. This structure we refer to as the audit study R^* can be seen in table 3.

Table 3: Audit study with true and observed cell counts.

Linkage cell		Cell count	
i	j	n^*	m^*
1	1	n_{11}^*	m_{11}^*
1	0	n_{10}^*	m_{10}^*
0	1	n_{01}^*	m_{01}^*

In general throughout this paper symbols that refer to elements related to the audit study are complemented with a $*$. In table 3, instead of counting correct and incorrect matches as in table 2, we present the observed and true cell counts in S^*, both according to probabilistic (i.e. $n^* = (n_{11}^*, n_{10}^*, n_{01}^*)$) and perfect (i.e. $m^* = (m_{11}^*, m_{10}^*, m_{01}^*)$) linkage in R. Notice we write 'in R', which is important and implies that the observed cell counts in R^* are based on the linkage of S_1 and S_2, not S_1^* and S_2^*. We see that the audit study has the same structure as the cell counts table in table 1. The difference is that in the audit study, beside the observed cell counts also the true cell counts are known. Finally note that we can write n_{11}^* and m_{11}^* as functions of a_{11}, a_{10} and a_{01} by $n_{11}^* = a_{11} + a_{10}$ and $m_{11}^* = a_{11} + a_{01}$ while the remaining parameters $n_{10}^*, n_{01}^*, m_{10}^*$ and m_{01}^* are simply residuals of the audit study sample size (e.g. $n_{10}^* = n_{1+}^* - n_{11}^*$).

The main difference between the rematch and the audit study is the way in which they are presented. However, in practice it might also be slightly easier to compile an audit study, because in a rematch study each match has to be studied separately while in the audit study groups of matches may be studied at once. For instance, when there are two individuals with the same name, it might be hard to verify which person is exactly which person, but it might be easier to verify that both individuals concern the same two individuals, which is enough for the audit study.

Because the rematch and therefore the audit study is assumed to be representative for R, we can simply use the ratio of n_{11}^* and m_{11}^* to obtain unbiased estimates for the true cell counts, i.e.:

$$\widehat{m}_{11} = n_{11} \frac{m_{11}^*}{n_{11}^*} \qquad \text{(5a),}$$

$$\widehat{m}_{10} = n_{1+} - \widehat{m}_{11} \qquad \text{(5b),}$$

$$\widehat{m}_{01} = n_{+1} - \widehat{m}_{11} \qquad \text{(5c).}$$

Equation (5a) simply defines $\widehat{m}_{11}$ larger as or smaller than n_{11}, depending on whether the number of links are over- or underestimated by the probabilistic linkage process. Equation (5b) and (5c) simply show $\widehat{m}_{10}$ and $\widehat{m}_{01}$ as residuals of the size of S_1, S_2 and $\widehat{m}_{11}$. When $m_{11}^* = n_{11}^*$ (i.e. no linkage errors on the cell counts level), $\widehat{m}_{11} = n_{11}$ and $\widehat{N}_{\text{D\&F}} = \widehat{N}_{\text{Petersen}}$. Further note that plugging in equation (5a) into equation (4) gives a formula that contains substantially less parameters than equation (3).

As an illustration in table 4 we extend the cell counts table in table 1 with the audit study. For simplicity we assume that we have an audit study that constitutes about 10% of the original linked sources and is perfectly representative.

Table 4: Example of true and observed cell counts with audit study.

Linkage cell		Cell count			
i	j	m	n	n^*	m^*
1	1	100	90	9	10
1	0	200	210	21	20
0	1	50	60	6	5

Under perfect linkage the Petersen estimator would be: $\widehat{N}_{\text{Petersen}} = \frac{300*150}{100} = 450$ while under linkage errors and by using the audit study, DW showed that both D&F and DC&T_15 can be reduced to: $\widehat{N}_{\text{D\&F}} = \widehat{N}_{\text{DC\&T\_15}} = \widehat{N}_{\text{DW}} = \frac{(90+210)*(90+60)}{90*\frac{10}{9}} = 450$, where we can see that $\widehat{m}_{11} = 90 * \frac{10}{9} = 100$. This implies that in this case the estimated cell count $\hat{m}_{11}$ is equal to the true cell count m_{11}. Note here that the correction method only requires the additional expression $\frac{m_{11}^*}{n_{11}^*}$, while n_{10}^*, n_{01}^*, m_{10}^* and m_{01}^* are not required.

Finally we note that equation (4) and (5) are not only a pleasant simplification of the D&F model, they also make the D&F correction method much easier to understand because it shows that all it does is replacing the observed cell counts n by the estimated cell counts $\hat{m}$, which can be estimated with the help of audit study R^*. In the next section we show how this insight can be used to obtain the more general WMRE model.

3 Derivation of the weighted multiple – recapture estimator

In this section we derive the WMRE model by combining equation (2), (4) and (5) in the previous section. The derivation is performed in two phases. First, in section 3.1, the D&F model is extended with covariates, leading to the more general WCR model. Next, in section 3.2, this WCR model is further extended with additional sources, leading to the WMR model.

3.1 Transformation of the D&F model into the WCR model with covariates.

The D&F model does not explicitly concern covariates. However, just like the Petersen estimator the D&F model could be applied on groups separately, which might solve this issue in some cases. However, in the context of covariates and linkage errors there are two problems with this approach. The first problem is related to (sparse) model selection. When capture probabilities are potentially related to a larger number of covariates (and maybe also the interaction between them), in the D&F model there is no standard approach available that selects those variables that fit the data best while they are selected sparsely. This in contrast to regular CR models, because different log – linear Poisson regression models can be conveniently compared by for instance their Aikake Information Criterion (Akaike, 1974) or Bayesian Information Criterion (Schwarz, 1978). The second problem of covariates and linkage errors is that they may lead to spurious correlations. For instance, when there is no relation between gender and capture probabilities, gender should be considered an irrelevant covariate. However, if there is a relation between linkage errors and gender, for instance men have a lower probability to be correctly linked, this leads to a spurious correlation between gender and capture probabilities. In this case the D&F model will (unnecessarily) be applied on men and women separately, which comes with additional noise in the resulting estimates. A better way would be to correct the cell counts for linkage errors first, hereby also removing the spurious correlation, and then estimate the population size without using gender as a covariate. In this section we will derive such a model.

Equation (2) shows that the Petersen estimator can also be written as a fitted value from a log - linear Poisson regression with table 1 as input and m as the dependent variable. As we have seen this Poisson regression can easily be extended with covariates. The first step is therefore to construct a new table that can serve as input for a log - linear Poisson regression such that it can be used to obtain the D&F estimator, which in table 5 can be seen to be pretty straightforward.

Table 5: Estimated cell counts table of two sources.

Linkage cell		Cell count
i	j	$\widehat{m}$
1	1	$\widehat{m}_{11} = n_{11}\dfrac{m_{11}^*}{n_{11}^*}$
1	0	$\widehat{m}_{10} = n_{1+} - \widehat{m}_{11}$
0	1	$\widehat{m}_{01} = n_{+1} - \widehat{m}_{11}$

In table 5 we have replaced the observed cell counts n from table 1 with the estimated cell counts $\widehat{m}$. We can now obtain a $\widehat{N}_{\mathrm{WCR}}$ without covariates, which results from the log - linear Poisson regression equation $E[\widehat{m}_{ij}] = e^{(\beta_0 + \beta_1 i + \beta_2 j)}$. The only difference with equation (2) is that n_{ij} is replaced by $\widehat{m}_{ij}$. This $\widehat{N}_{\mathrm{WCR}}$ is identical to the $\widehat{N}_{\mathrm{D\&F}}$ in equation (4). This can be shown by: $\frac{(\widehat{m}_{11}+\widehat{m}_{10})(\widehat{m}_{11}+\widehat{m}_{01})}{\widehat{m}_{11}} = \frac{n_{1+}n_{+1}}{\widehat{m}_{11}} = \widehat{N}_{\mathrm{D\&F}}$ (which further implies that $\widehat{m}_{00} = \frac{\widehat{m}_{10}\widehat{m}_{01}}{\widehat{m}_{11}}$ is the WCR estimate for m_{00}).

This regression formulation allows us to simply add a covariate x to the equation, i.e.:

$E[\widehat{m}_{ijx}] = e^{(\beta_0 + \beta_1 i + \beta_2 j + \beta_3 x + \beta_4 ix + \beta_5 jx)}$, where e.g. ix is the product of i and x. This can be further illustrated with a simple example. When we assume x is a binary covariate and we define γ_{ij0} as the proportion of $x = 0$ within R and γ_{ij1} as the proportion $x = 1$ (so $\gamma_{ij0} + \gamma_{ij1} = 1$) within R, table 6 shows the estimated cell counts table for this case.

Table 6: Estimated cell counts and one binary covariate.

Linkage cell			Cell count
i	j	x	$\widehat{m}$
1	1	1	$\widehat{m}_{111} = \gamma_{111}\widehat{m}_{11}$
1	1	0	$\widehat{m}_{110} = \gamma_{110}\widehat{m}_{11}$
1	0	1	$\widehat{m}_{101} = \gamma_{101}(n_{1+} - \widehat{m}_{11})$
1	0	0	$\widehat{m}_{100} = \gamma_{100}(n_{1+} - \widehat{m}_{11})$
0	1	1	$\widehat{m}_{011} = \gamma_{011}(n_{+1} - \widehat{m}_{11})$
0	1	0	$\widehat{m}_{010} = \gamma_{010}(n_{+1} - \widehat{m}_{11})$

Table 6 provides an intuitively easy to grasp cell counts table for a D&F model including a covariate. By means of the log – linear Poisson regression on $\widehat{m}$ it can be used to obtain the estimates $\widehat{m}_{001} = e^{(\widehat{\beta}_0 + \widehat{\beta}_3)}$ and $\widehat{m}_{000} = e^{(\widehat{\beta}_0)}$.

However, in order to add additional sources, which we will discuss in the next section, instead of splitting - up $\widehat{m}$ on the cell count level as in table 6, we split - up $\widehat{m}$ further into record level weights that add up to the cell counts $\widehat{m}$ in table 6. In order to keep the notation of this split - up as simple as possible, we specify i and j on the record level in R, i.e.:

$i_p = \begin{cases} 1 \text{ if record } p \text{ is in } S_1 \\ 0 \text{ if not} \end{cases}$ and

$j_p = \begin{cases} 1 \text{ if record } p \text{ is in } S_2 \\ 0 \text{ if not} \end{cases}$

where $p = 1, \dots, P$ with $P = n_{11} + n_{10} + n_{01}$ is the number of records in R. This implies that m and n can be calculated by summing up $i_p's$ and $j_p's$ in R. In fact, under probabilistic linkage we get $\sum_p i_p = n_{1+}$, $\sum_p j_p = n_{+1}$ and $\sum_p i_p j_p = n_{11}$ while under perfect linkage we get $\sum_p i_p = m_{1+}$, $\sum_p j_p = m_{+1}$ and $\sum_p i_p j_p = m_{11}$. The same can be done for n^* and m^*. The introduction of i_p and j_p allows us to write:

$$w_p = \frac{\widehat{m}_p}{n_p} \qquad (6),$$

with $\widehat{m}_p = \widehat{m}_{i_p j_p} \in (\widehat{m}_{11}, \widehat{m}_{10}, \widehat{m}_{01})$ and $n_p = n_{i_p j_p} \in (n_{11}, n_{10}, n_{01})$. In words, when a record in S_1 is linked with a record in S_2 they become a single record in R that is part of linkage cell $i_p j_p = 11$ and so $\widehat{m}_p = \widehat{m}_{11}$ and $n_p = n_{11}$. When covariates are involved $\widehat{m}_p$ and n_p should be determined on the covariate level (e.g. if record p is a man, $\widehat{m}_p$ and n_p should refer to the linkage of men between sources). With equation (6) this implies that summing up the weights w_p over the records within linkage cells returns our original $\widehat{m}$, i.e. $\sum_{i_p j_p \epsilon ij} w_p = \widehat{m}_{ij}$. In case of covariates, the definition of $\widehat{m}_p$ can be slightly extended to $\widehat{m}_p = \widehat{m}_{i_p j_p X}$ with $X = (x_{p1}, \dots, x_{pC})$ is a set of C categorical covariates. In table 7 we combine the example from table 6 with both this extended notation and equation (6).

Table 7: Example of R with w_p.

Record	Linkage cell			Cell count		Weight
p	i_p	j_p	x	n_p	$\widehat{m}_p$	w_p
1	1	1	1	n_{111}	$\widehat{m}_{111}$	$w_1 = \widehat{m}_{111} / n_{111}$
2	0	1	0	n_{010}	$\widehat{m}_{010}$	$w_2 = \widehat{m}_{010} / n_{010}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
p	i_p	j_p	x_p	$n_{i_p j_p x_p}$	$\widehat{m}_{i_p j_p x_p}$	$w_p = \widehat{m}_{i_p j_p x_p} / n_{i_p j_p x_p}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
P	0	1	1	n_{011}	$\widehat{m}_{011}$	$w_P = \widehat{m}_{011} / n_{011}$

Here we should emphasize that p is not a unit index but a record index in R, which may be larger or smaller than the number of true population units in R, depending on whether n_{11} is an over- or underestimation of m_{11}, i.e. $n_{11} + n_{10} + n_{01} = P \neq m_{11} + m_{10} + m_{01}$ and $\widehat{m}_{11} + \widehat{m}_{10} + \widehat{m}_{01} = \sum_p w_p \neq P$. We can interpret w_p as an indicator that indicates whether a record is part of a linkage cell that is under- or overrepresented by the linkage procedure. Here $w_p < 1$ implies that the linkage cell of record p occurs more frequent and $w_p > 1$ implies that the linkage cell of record p occurs less frequent than observed cell counts suggests. For instance, if the linkage procedure underestimates the number of true matches, i.e. $n_{11} < m_{11}$, then records in linkage cell $ij = 11$ should have weights larger than 1, in order to compensate for the low observed cell count n_{11}. When w_p is used to obtain the estimated cell counts of the different linkage cells as in table 6 and these estimated cell counts are used as the dependent variable in the log – linear Poisson regression model, this gives us a CRE that is corrected for linkage errors and includes x as a covariate, hence the WCRE and the WCR model.

3.2 Transformation of the WCR model into the WMR model.

In section 3.1 the D&F model without covariates was transformed into the WCR model with covariates. However, the WCR model still only concerns two sources, which is insufficient in case of source dependence (i.e. when captures in S_1 are related to captures in S_2). The MR model is less sensitive to source dependence, because it can correct for it by using multiple sources (e.g. Fienberg, 1972). It is therefore desirable to have a model that can correct for linkage errors, covariate dependence and source dependence simultaneously. However, it is not straightforward to extend the WCR model with addition sources, which is the subject of this section.

Considering both simplicity and practice, we assume that linkage of sources occurs sequentially, which implies that first two sources are linked and consecutively a new source is linked to this linked pair of sources as if they were one source. It was also argued by DC&T_18 that this linking approach is currently the most reasonable approach to consider because the alternatives of simultaneous or pairwise linkage 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 $A \to B, B \to C, C \nrightarrow A$?). Not coincidentally, the sequential approach is also quite common in practice.

In order to add additional sources we extend w_p, i_p, j_p, n, m, $\widehat{m}$, S_1, S_2, R, L and P with the additional sub- or superscript t that indicates the number of linkage procedures, i.e. $w_p \to w_p^t,\ i_p \to i_p^t,\ j_p \to j_p^t, n \to n^t,\ \ m \to m^t,\ \widehat{m} \to \widehat{m}^t,\ S_1 = S_t,\ S_2 = S_{t+1},\ R \to R_t$, $L \to L_t$ and $P \to P_t$, where R_{t-1} is further defined as:

$$R_{t-1} = \begin{cases} R_0 = S_1 \\ R_1 = L_1(S_1, S_2) \\ R_2 = L_2(R_1, S_3) \\ \qquad \vdots \\ R_{t-1} = L_t(R_{t-2}, S_t) \end{cases}$$

This introduction of t further implies a more nuanced interpretation of n^t, m^t and $\widehat{m}^t$. Beside a count variable, n^t can also be interpreted as the observed sum of weights after t linkages. Then, for $t = 1$, $n^{t=1}$ is simply the original observed count variable n, but for $t > 1$, n^t is the observed sum of weights after t probabilistic linkages and $t - 1$ updates. m^t is the true sum of weights after $t - 1$ probabilistic linkages and updates and a t^{th} perfect linkage. Finally $\widehat{m}^t$ is the estimate of m^t. Here we should emphasize that the difference between n^t, m^t and $\widehat{m}^t$ is only due to the last linkage and update of weights, because all three variables are based on the same sums of weights at $t - 1$, i.e. $\widehat{m}^{t-1}$, obtained over the previous periods.

When for the moment we ignore covariates, this extended notation allows us to write equation (6) as a record level update function that contains t, i.e.:

$$w_p^t = w_p^{t-1} \frac{\widehat{m}_p^t}{n_p^t} \tag{7}$$

where $\widehat{m}_{11}^t = n_{11}^t \frac{m_{11}^{*t}}{n_{11}^{*t}}$, $\widehat{m}_{10}^t = n_{1+}^t - \widehat{m}_{11}^t$, $\widehat{m}_{01}^t = n_{+1}^t - \widehat{m}_{11}^t$ and $w_p^{t=0} = 1$ that can be considered ‘an individual starting weight of 1’. When covariates are also in the equation n_p^t and $\widehat{m}_p^t$ must be determined on the covariate level, e.g. the number of observed and estimated links between sources of both men and women. The starting weight of 1 also applies to records that are new in the last source S_{t+1}, i.e. $w_{i_p^t j_p^t \in 01}^t = 1$. For $t = 1$, equation (7) reduces to equation (6), but for $t > 1$ equation (7) allows us to update weights after linking additional sources. As an illustration table 8 gives the dependent variable $\widehat{m}^{t=2}$ as a result of the updating of weights after linking three sources (i.e. $t = 2$).

Table 8: Update of weights for $t = 2$.

Linkage cell		Cell count
$i^{t=2}$	$j^{t=2}$	$\widehat{m}^{t=2}$
1	1	$\widehat{m}_{11}^{t=2} = \widehat{m}_{11}^{t=1} \frac{\widehat{m}_{11}^{*t=2}}{n_{11}^{*t=2}}$
1	0	$\widehat{m}_{10}^{t=2} = \widehat{m}_{1+}^{t=2} - \widehat{m}_{11}^{t=2}$
0	1	$\widehat{m}_{01}^{t=2} = \widehat{m}_{+1}^{t=2} - \widehat{m}_{11}^{t=2}$

Table 8 shows that in order to obtain the elements in $\widehat{m}^{t=2}$ all that is required is a value for $\widehat{m}_{11}^{t=2}$. This is sufficient because the total sums of weights in R_2 and S_3, i.e. $\widehat{m}_{1+}^{t=2}$ and $\widehat{m}_{+1}^{t=2}$, are both unaffected by the updating of weights, i.e. $\widehat{m}_{1+}^{t=2} = \widehat{m}_{1+}^{t=1}$ and $\widehat{m}_{+1}^{t=2} = \widehat{m}_{+1}^{t=1}$. Finally, summing up weights over records in different linkage and covariate linkage cells give the estimated cell counts $\widehat{m}_{ijkX}^{t}$, where $X = (x_1, \dots, x_r, \dots, x_C)$ represents a set of C categorical covariates. This updating of weights can be repeated after the linkage of each new source and using the sums of these weights per linkage cell as dependent variable in the log - linear Poisson regression constitutes the WMR model and the WMRE, which is corrected for linkage errors.

A general formulation of this log - linear Poisson regression of the WMR model can be written as:

$$E\left[\widehat{m}_{Z_t}\right] = e^{f(\beta, Z_t)} \qquad (8),$$

where $\widehat{m}_{Z_t}$ is the linkage errors corrected cell count vector that depends on $Z_t = (R_t, X)$, e.g. for three sources and one binary covariate we get:

$$\widehat{m}_{Z_t} = \widehat{m}_{ijkx} = \begin{pmatrix} \widehat{m}_{1111}, \widehat{m}_{1110}, \widehat{m}_{1101}, \widehat{m}_{1100}, \widehat{m}_{0111}, \widehat{m}_{0110}, \widehat{m}_{1011}, \widehat{m}_{1010}, \\ \widehat{m}_{1001}, \widehat{m}_{1000}, \widehat{m}_{0101}, \widehat{m}_{0100}, \widehat{m}_{0011}, \widehat{m}_{0010}, \widehat{m}_{0001}, \widehat{m}_{0000} \end{pmatrix}.$$

Furthermore, $f(\beta, Z_t)$ transforms Z_t into an equation that contains all linear combinations of linkage cells, e.g. when there are three sources and one binary covariate we get:

$$f(\beta, Z_t) = f(\beta, i, j, k, x) =$$

$$\beta_0 + \beta_1 i + \beta_2 j + \beta_3 k + \beta_4 x + \beta_5(ix) + \beta_6(jx) + \beta_7(kx) +$$

$$\beta_8(ij) + \beta_9(ik) + \beta_{10}(jk) + \beta_{11}(ijx) + \beta_{12}(ikx) + \beta_{13}(jkx)$$

where e.g. ix is the product of i and x. This equation will become larger quickly for more sources and covariates but this extension is straightforward while standard model selection techniques can be used to reduce the number of parameters.

It is interesting to compare table 8 to the transition matrix given by DC&T_18. Due to the number of matching patterns of a record, their matrix contains four unknown parameters, has a dimension of seven by seven (i.e. $2 \text{ x } 2 \text{ x } 2 - 1 = 7$) and will grow to six unknown parameters and a dimension of fifteen by fifteen (and still needs to be mathematically derived) if another source would be added. In fact, these numbers would grow exponentially with each newly added source and the exact transition matrix would have to be mathematically derived anew for each source. In contrast, because in the WMR model the updating of weights occurs after each new linkage, table 8 only requires one to know $\hat{m}_{11}^{*t}$ and n_{11}^{*t}, which both can be directly obtained from the audit study.

Finally it might be clarifying to show in table 9 how w^t looks for an example of R with three sources and a binary covariate x.

Table 9: Example of R with three sources and a binary covariate x

	Linkage cell					Weight
Record	$t = 1$		$t = 2$			$t = 2$
p	$i_p^{t=1}$	$j_p^{t=1}$	$i_p^{t=2}$	$j_p^{t=2}$	x	$w^{t=2}$
1	1	1	1	1	0	$w_1^{t=2} = \frac{\widehat{m}_{110}^{t=1}}{n_{110}^{t=1}} \frac{\widehat{m}_{110}^{t=2}}{n_{110}^{t=2}}$
2	0	1	1	0	1	$w_2^{t=2} = \frac{\widehat{m}_{011}^{t=1}}{n_{011}^{t=1}} \frac{\widehat{m}_{101}^{t=2}}{n_{101}^{t=2}}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
p	$i_p^{t=1}$	$j_p^{t=1}$	$i_p^{t=2}$	$j_p^{t=2}$	x_p	$w_p^{t=2} = \frac{\widehat{m}_{i_p^{t=1} j_p^{t=1} x_p}^{t=1}}{n_{i_p^{t=1} j_p^{t=1} x_p}^{t=1}} \frac{\widehat{m}_{i_p^{t=2} j_p^{t=2} x_p}^{t=2}}{n_{i_p^{t=2} j_p^{t=2} x_p}^{t=2}}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
P	0	0	0	1	1	$w_P^{t=2} = 1 \frac{\widehat{m}_{011}^{t=2}}{n_{011}^{t=2}}$

The weight column in table 9 might look quite complicated, but on closer inspection each $w_p^{t=2}$ is just a $w_p^{t=0} = 1$ that is multiplied twice with its true sums of weights that correspond to record $p's$ linkage cells and divided twice by its observed sums of weights that correspond to record $p's$ linkage cells, which all follow directly from the audit studies at $t = 1$ and $t = 2$. This shows why this updating of weights can be easily continued with additional sources.

4. Model evaluation

We evaluate the WMRE 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. In order to make the simulation study slightly less artificial, we use a quasi - real dataset to simulate from. This quasi - real dataset is a publicly available fictitious population dataset of 26 625 persons that is representative for the UK population census. It was created in the ESSnet DI (McLeod, Heasman and Forbes, 2011), a European project on data integration (Record Linkage, Statistical Linking, Micro integration Processing) 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 compare the performance of the WMRE with the performance of other estimators that were discussed in section 2. The different performances are compared under different scenarios, where scenarios differ with respect to three elements:

(i) Covariate dependence of capture probabilities, which implies 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).

(ii) Source dependence of capture probabilities, which implies 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).

(iii) 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 the WMRE model aims to correct for while the alternative models should suffer from at least one of them. In section 4.1 we describe the setup of this simulation study and in section 4.2 we discuss the results.

4.1. Simulation study setup

Before we discuss the details of the simulation study, we first discuss some general considerations. In a simulation study, in order to compare different estimators in a fair way there are two important elements to bear in mind, i.e. violations of assumptions and randomness due to sampling. Violations of assumptions that underlie the CR and MR model lead to bias. When multiple assumptions are violated simultaneously, the multiple sources of bias may occur simultaneously (and may even coincidentally cancel out each other). In such a case, a bias in the estimate cannot be attributed to a single violation. Therefore, in order to know whether a model corrects for a specific source of bias, it is important to introduce the different sources of potential bias separately. In this way the sources of bias can also be identified separately. In table 10 we show the four simulation scenarios of our interest.

Table 10. 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

The four scenarios differ with respect to covariate and source dependence while all scenarios suffer from linkage errors. As a benchmark, for each scenario we also calculate the CRE and MRE that (falsely) assume that there are no linkage errors.

Second, it is important to realise that even under perfect conditions a CRE or MRE may differ from its true value for two reasons. First, simply due to randomness in the sources. Sometimes a PSE overestimates and sometimes underestimates the TPS, which, if all the underlying model assumptions are met, should on average be close to zero if the sampling is repeated many times. Therefore, we replicate the sampling and estimation procedure a large number of times (i.e. 1 050)[5], where in every replication both a new population and three population sources S_1, S_2 and S_3 are generated. This gives, for each model and each scenario, 1 050 PSEs, which allows us to compare their mean and the distribution of each scenario. Second, in the context of the CRE, Poisson regression estimators have known finite sample bias (see e.g. Chapman, 1951, Menkens and Anderson, 1988 or Chen and Giles, 2009), which goes to zero when the sample increases to infinity. That is why we set the

[5] 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.

population size on 10 000, because this makes the finite sample bias practically ignorable.

In is interesting to note that in our original simulation setup we first considered a TPS of 1 000, because due the computational intensity of linkage procedures this is much more convenient. However, as it turned out, our estimation results were slightly but statistically significantly biased and we could not eliminate the possibility that this bias was due to a mistake in our derivation of the WMRE. Fortunately, with a population size of 10 000 we found that this bias was indeed due to finite sample size bias, because after the scale - up the bias disappeared. Probably, an example of this finite sample bias can also be seen in DC&T_18 who present a simulation study with similar data and setup but with a TPS of 1 000. In this study, the mean of the PSEs that were unaffected by linkage errors deviates slightly but statistically significantly (i.e. by 1.05%) from the TPS. This small bias is similar to the finite sample bias that we encountered.

From the available dataset we use the file ‘person_list.csv’. This list contains both a perfect identifier (id - code) and linkage keys (e.g. surname, address) and can therefore be used to link records both perfectly (i.e. deterministically without any errors) and probabilistically. In this simulation study we use a set of three linkage keys[6]. In order to have a certain degree of linkage errors, in each linkage key in each source, 3% of the records is replaced by a random value from the population, where in each source, each record has the same probability to be selected. Furthermore, the list contains several covariates, of which we use ‘SEX’ as covariate X to affect capture probabilities.

For each replication first a random population of 10 000 records is generated (without replacement) from the person list. Our aim is then to generate three sources of different sizes from this population (approximately 8 000, 5 000 and 2 000 records) that may suffer from source and covariate dependence. The introduction of source dependence is not straightforward, because source dependence implies that no single source may be independent of other sources. However, when the first source would be generated while other sources do not yet exist, this first source is by definition independent of these other sources. Therefore, before the first source is

[6] ‘PERNAME2’, ‘DOB_DAY’ and ‘DOB_MON’ served as linkage variables, which corresponds to the ‘bronze scenario’ in DC&T_15.

generated, we first generate three so called latent sources $\tilde{S} = (\tilde{S}_1, \tilde{S}_2, \tilde{S}_3)$ of 8 000 records, which are simply random samples from the population of 10 000. These three latent sources allow us to introduce dependencies between sources such that no source in S is independent of the other sources. This is done by giving each unit a probability to be in each source by:

$$P_{pl}[S_l = 1] = \frac{1}{1 - \exp(-\mu_{pl})} \qquad (9),$$

where $\mu_{pl} = \delta_{1,l}\tilde{S}_{1,p} + \delta_{2,l}\tilde{S}_{2,p} + \delta_{3,l}\tilde{S}_{3,p} + \delta_{4,l}X_p$, $p = 1, \dots, 10\ 000$ refers to the records in the true population and $l = 1,2,3$ refers to the three latent sources. Given equation (9) we can vary $\theta's$ and hereby control dependencies between any source in S and the other two sources in S and the covariate. For instance, when $\delta_{1,1}, \delta_{2,1}, \delta_{2,2}, \delta_{2,2} \neq 0$, the probability of a record to be in S_1 depends on it being in S_2 while the probability to be in S_2 also depends on it being in S_1. Furthermore, the $\delta's$ control the size of each source. The values for the $\delta's$ in the simulation study can be found in table 11 in appendix A. Because the varying of $\delta's$ affects the capture probabilities of records, different $\delta's$ also correspond to different estimates of the $\hat{\beta}'s$ from the Poisson regressions. Therefore, in order to assure that by varying $\delta's$ we introduce a substantial source and covariate dependence, in table 12 in appendix A we also present the mean values of estimated $\hat{\beta}'s$ over all replications of the benchmark case of no linkage errors. Here it is important to look at the case of no linkage errors, because otherwise we might be looking at dependencies that are the result of linkage errors, while we want the source and covariate dependency to occur also without linkage errors.

Finally, the last necessary element of the simulation study is the audit study R^*, which is generated by first selecting a random 10% (without replacement) of the population and within this selection only keeping those records that are also in one of the three sources S. With this audit study all ingredients are available to obtain the PSEs of interest in this simulation study, which is discussed next.

In order to compare the PSEs we compare the CRE and MRE (with 3 sources) with three different dependent variables (n, m and $\hat{m}$) under the four different scenarios from table 10. We refer to the use of n as the naïve estimates, the use of m as the perfect (but in practise unobtainable) estimates and $\hat{m}$ as the weighted estimates.

4.2. Simulation results

In figure 2 below the simulation results of the four scenarios are presented as density plots.

Figure 2: Density plots of two PSEs with three dependent variables and four scenarios (table 10).

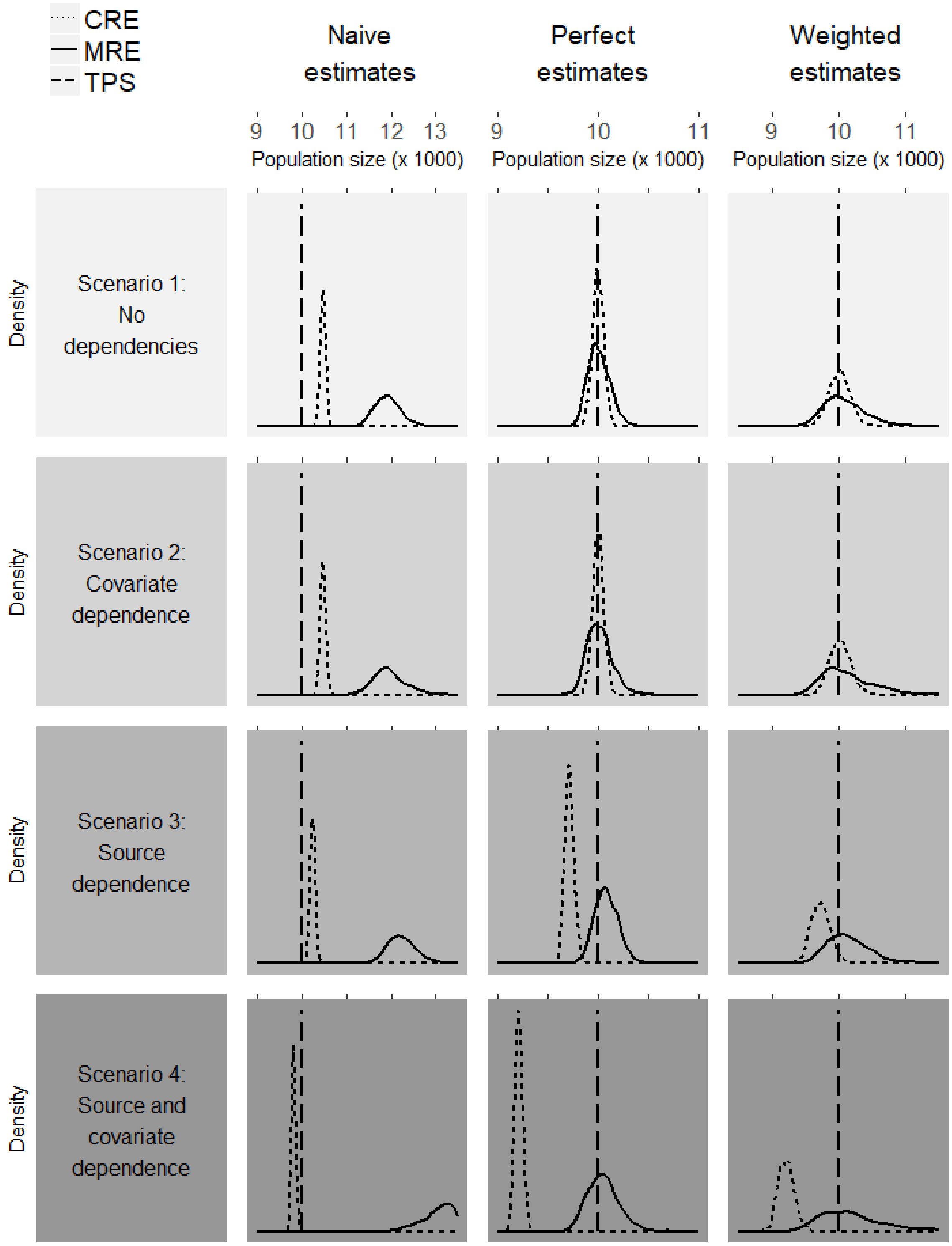


Figure 2 contains twelve density plots that each contains two PSE densities. In the row direction there are the four scenarios from table 10 and in the column direction there are the three estimates, i.e. naïve, perfect and weighted. Ideally a density revolves around the TPS of 10 000. However, the first column shows that the densities of the naïve estimates do not. From the simulation study perspective, this is a good sign, because it shows that the estimates suffer from linkage errors in all scenarios, as intended by the simulation setup. 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 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 again (necessarily) fails, because it is unable to deal with source dependence. Fortunately the density of the MREs revolve around the TPS in all scenarios, which shows that the WMR model corrects for linkage errors, covariate dependence and source dependence simultaneously.

5. Discussion

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 includes three or more sources and covariates, which are often necessary to correct for other sources of bias. The linkage error correction model we developed is incorporated in the more general family of log - linear regression models. It 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 practise the WMR model does not solve all the linkage error problems. Although it might be easier to construct an audit study as in table 3 than a rematch study as in table 2, because table 3 does not require closer scrutiny on the record level but on the linkage cell level, the main practical problem is still the selection of records that constitute the audit study R^*? Ideally the records in R^* are representative for the records in R, both with respect to covariates and the quality of linkage keys. This last element should not be underestimated, because when for instance the records in the audit study are based on their high quality linkage keys (what makes it easier to audit them), they might suffer less from linkage errors than the rest of the population. This will lead to a biased correction. Another issue is the size of R^*, in particular when the population contains small specific groups with low capture probabilities, in practise it might not be easy to have this small group represented sufficiently in R^*. How large the impact of such issues is, requires further research.

Also we should note that we paid little attention to the impact of the exact linkage procedure. We developed the WMR model in the context of sequential linkage, in which first two sources are linked and a third source is linked to this combined source. We think that in theory the order of linkage does not matter and also pairwise linkage (link each pair and then combine them into one) or simultaneous linkage (link all sources at once) can be incorporated into the WMR model, although this would require further research. In practise the exact linkage strategy may play a role, mainly because linkage is also often used to enrich sources. When, for instance one source contains data on say gender and another on income, the combined source usually contains both, which will probably affect the quality of linkage with a third source that also contains gender and income.

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

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19* (6): 716–723*.*

Bishop Y.M.M., Fienberg S.E. and Holland P.W. (1975). Discrete Multivariate Analysis: Theory and Practice. *MIT Press: Cambridge, Mass.*

Cadwell, B.L., Smith, P.J. and Baughman A.L. (2005). Methods for capture–recapture analysis when cases lack personal identifiers. *Statistics in Medicine, 24*(13): 2041–2051.

Chao, A. (2001). An Overview of Closed Capture-Recapture Models. *Journal of Agricultural, Biological, and Environmental Statistics 6*: 158–175.

Chapman, D.G. (1951). Some properties of the hypergeometric distribution with applications to zoological sample censuses. *Berkeley, University of California Press.*

Chen, Q. and Giles, D.E. (2009). Finite-Sample Properties of the Maximum Likelihood Estimator for the Poisson Regression Model With Random Covariates*. Econometrics Working Paper EWP0907*, University of Victoria.

Chen, Z. and Kuo, L. (2001). A Note on the Estimation of the Multinomial Logit Model with Random Effects. *The American Statistician 55*: 89–95.

Cormack, R. M. (1989). Log-linear models for capture-recapture. *Biometrics, 45, 395 – 413*.

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.

Di Consiglio, L. and Tuoto, T. (2015). Coverage evaluation on probabilistically linked data. *Journal of Official Statistics, 31*, 415 – 429.

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.

Ding, Y. and Fienberg, S.E. (1994). Dual system estimation of Census undercount in the presence of matching error. *Survey Methodology, 20*, 149 – 158.

Fellegi, I. P. and Sunter, A. B. (1969). A Theory for Record Linkage. *Journal of the American Statistical Association, 64*, 1183 – 1210.

Fienberg, S.E. (1972), The multiple recapture census for closed populations and incomplete 2^{k} contingency tables, *Biometrika, 59*(3), 591–603.

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*.

International Working Group for Disease Monitoring and Forecasting (IWGDMF, 1995). Capture-recapture and multiple-record systems estimation I: history and theoretical development. *American Journal of Epidemiology; 142*, 1047–1058.

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.

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.

Lincoln, F. C. (1930). Calculating Waterfowl Abundance on the Basis of Banding Returns*, U.S. Dept. Agric., 118*: 1-4.

Menkens, G.E. and Anderson Jr., S.H. (1988). Estimation of Small - Mammal Population Size*, Ecology, Vol. 69, No. 6*, 1952-1959.

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.

Schwarz, G.E. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2): 461–464

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.

Wolter, K.M. (1986). Some coverage error models for census data. *Journal of the American Statistical Association, 81*, 338 – 346.

Appendix A

Table 11: Parameter values of the four different scenarios.

Scenario 1	$\delta_{1,1}$	$\delta_{2,1}$	$\delta_{3,1}$	$\delta_{4,1}$
score_{i1}	6.3	0	0	0
score_{i2}	0	3.5	0	0
score_{i3}	0	0	1.9	0

Scenario 2	$\delta_{1,2}$	$\delta_{2,2}$	$\delta_{3,2}$	$\delta_{4,2}$
score_{i1}	5.6	0	0	2
score_{i2}	0	4.6	0	-2
score_{i3}	0	0	0.42	2

Scenario 3	$\delta_{1,3}$	$\delta_{2,3}$	$\delta_{3,3}$	$\delta_{4,3}$
score_{i1}	4.8	1.8	0	0
score_{i2}	0	3.5	0	0
score_{i3}	0	-0.5	2.3	0

Scenario 4	$\delta_{1,4}$	$\delta_{2,4}$	$\delta_{3,4}$	$\delta_{4,4}$
score_{i1}	3.9	1.5	0	2
score_{i2}	1.5	3.3	0	-2
score_{i3}	0	-0.5	1.8	1

Table 12 below shows for each scenario the mean value of the $\hat{\beta}'s$ in case there would be no linkage errors that have a value significantly different from zero.

Table 12: Average estimated $\widehat{\beta}'s$ per scenario without linkage errors

Scenario	**1***	**2***	**3***	**4***
Variable\Estimate	$\hat{\beta}$	$\hat{\beta}$	$\hat{\beta}$	$\hat{\beta}$
Constant	13,0	12,8	13	13,3
S_1	1,3	1,1	1,2	0,3

S_2	.	0,7	-0,2	.
S_3	-1,3	-2,7	-1,3	-1,6
X	.	-0,1	.	-0,6
S_1X	.	0,6	.	1,5
S_2X	.	-1,5	.	-1,9
S_3X	.	2	.	0,8
S_1S_2	.	.	0,4	1,1
S_1S_3	.	.	.	.
S_2S_3	.	.	-0,2	-0,2
S_1S_2X	.	.	.	0,2
S_1S_3X	.	.	.	.
S_2S_3X	.	.	.	0,1
* indicates 'scenario without linkage errors'.				

Table 12 clearly shows that the estimated $\hat{\beta}'s$ correspond to the four scenarios. In scenario 1 neither covariate X nor another source plays a significant role in describing the observed frequencies. In scenario 2 the observed frequencies do not depend on other sources but does depend on X. In scenario 3 the covariate X is not significant while the other sources have significant explanatory power. In scenario 4 both X and the other sources play a significant role. This is in accordance with the four intended scenarios presented in table 10.

Explanation of symbols

Empty cell	Figure not applicable
.	Figure is unknown, insufficiently reliable or confidential
*	Provisional figure
**	Revised provisional figure
2018–2019	2018 to 2019 inclusive
2018/2019	Average for 2018 to 2019 inclusive
2018/’19	Crop year, financial year, school year, etc., beginning in 2018 and ending in 2019
2016/’17–2018/’19	Crop year, financial year, etc., 2016/’17 to 2018/’19 inclusive

Due to rounding, some totals may not correspond to the sum of the separate figures.

Colophon

Publisher
Statistics Netherlands
Henri Faasdreef 312, 2492 JP The Hague
www.cbs.nl

Prepress: Statistics Netherlands
Design: Edenspiekermann

Information
Telephone +31 88 570 70 70
Via contact form: www.cbs.nl/information