sciendo

Journal of Official Statistics, Vol. 37, No. 3, 2021, pp. 699–718, http://dx.doi.org/10.2478/JOS-2021-0031

A General Framework for Multiple-Recapture Estimation that Incorporates Linkage Error Correction

Daan Zult[1], *Peter-Paul de Wolf*[1], *Bart F. M. Bakker*[1], *and Peter van der Heijden*[2]

The size of a partly observed population is often estimated with the capture-recapture model. An important assumption of this chat model is that sources can be perfectly linked. This assumption is of relevance if the identification of records is not obtained by some perfect identifier (such as an id code) but by indirect identifiers (such as name and address). In that case, the perfect linkage assumption is often violated, which in general leads to biased population size estimates. Initial suggestions to solve this use record linkage probabilities to correct the capture-recapture model. In this article we provide a general framework, based on the standard log-linear modelling approach, that generalises this work towards the inclusion of additional sources and covariates. We show that the method performs well in a simulation study.

Key words: Population size estimation; capture-recapture; dual-system estimation; multiple-system estimation; record linkage.

1. Introduction

Capture-recapture (CR) estimation provides a standard approach to estimate the size of a population, including the unobserved part (Petersen 1896; Fienberg 1972; Bishop et al. 1975). These models are also known under other names, such as dual-system, multiple-system and mark-recapture estimation models (see e.g., IWGDMF 1995). Dual-system (DS) estimation uses two sources and multiple-system (MS) estimation uses three or more sources (e.g., Fienberg 1972). A source refers to a set, list or register of records. We assume that each *record* represents a *unit* that is unique to that source and belongs to the target population. When the combination of available sources does not cover the full target population, under specific assumptions as described in Wolter (1986), CR models can be used to estimate the size of the missing part of the population.

One of the assumptions in CR models is that records can be perfectly identified over sources as belonging to the same unit, or not. This allows an accurate linkage of records and sources into one combined source. If a perfect identification of units is not possible, 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) is generally biased (Wolter 1986; Chao 2001; Chen and Kuo 2001; Cadwell 2005; Gerritse et al. 2017).

[1] Statistics Netherlands – Methodology, P.O. Box 24500, 2490 HA The Hague, the Netherlands. Emails: db.zult@cbs.nl,pp.dewolf@cbs.nl, and bfm.bakker@cbs.nl
[2] Utrecht University – Methodology and Statistics, Padualaan 14, Utrecht 3508 TC, the Netherlands. Email: P.G.M.vanderHeijden@uu.nl

A first step in a solution to this problem was provided by D&F (Ding and Fienberg 1994). For the linkage of two sources S^1 and S^2 they define five different linkage error types:

(1) A missed link between the same unit that is in both S^1 and S^2,
(2a) A false link between two different units that are both in S^1 and S^2,
(2b) A false link between a unit that is in S^1 and S^2 and a different unit that isonly in S^2,
(2c) A false link between a unit that is in S^1 and S^2 and a different unit that is only in S^1, and
(2d) A false link between two different units that are in S^1 and S^2.

Linkage error type (1) concerns a missed match while types (2a−d) concern different types of mismatches. To simplify the model, D&F assume that linkage error types (2a−c) are negligible because they require a double linkage error. Therefore, they derive a model that corrects for the two remaining linkage error types (1a) and (2d).

The D&F + model requires a rematch study. This is a study that checks whether or not a subset of record linkages and non-linkages is correct, and is usually carried out by a clerical review. This subset is assumed to be representative for the entire population. The D&F + model uses the rematch study to obtain different sorts of linkage error probabilities.

Note that linkage errors refer to *record* linkage errors that occur during *source* linkage. A record linkage is the linkage between two records in two sources. Source linkage refers to the linkage of records in two or more sources.

The D&F model is extended by DC&T_15 and DC&T_18 (Di Consiglio and Tuoto 2015, 2018). They showed that D&F only explicitly consider the probability of a record in S^1 to be falsely linked to a record in S^2, while a record in S^2 can just as well be falsely linked to a record in S^1. Therefore, DC&T_15 derive a model that takes both options into account. Further progress is presented by WLZ (De Wolf et al. 2019), who showed that both D&F and DC&T_15 implicitly assume that S^1 and S^2 are of equal size. This is important, because the probability of a false link increases or decreases when the number of potential record linkages increases or decreases, which depends on the size of both sources. Therefore, WLZ derive a model that takes these different source sizes into account. This progress in DC&T_15 and WLZ is restricted to linkage error type (1) and (2d). WLZ take one more step and derive a (what we refer to as D&F +) model that takes all five linkage error types into account. We will see in Section 3 that this D&F + model is also less complex and can be used to generalize the model even further.

Despite the progress, the D&F + model still suffers from two major shortcomings:

(1) It is unclear how to perform statistical inference with respect to covariates in the model, and
(2) The D&F + model is only defined for two sources and not for three or more.

These two shortcomings are important where captures are covariate and/or source-dependent, while they cannot be modelled explicitly in case of recapture-prone or recapture-adverse populations (e.g., see Chatterjee and Mukherjee 2018). If there are two sources, these linkage error probabilities can be incorporated in the derivation of a closed form maximum likelihood estimator when the sources are independent. However, this

derivation becomes increasingly complicated when covariates and additional sources are added, and it is unclear how to perform statistical inference in this situation.

In this article we propose to use the rematch study in a different way than the existing linkage error correction models. Where these existing models first estimate linkage error probabilities and use these probabilities to correct the DS estimate, we directly correct the cell counts in the contingency table for linkage errors. In this way, linkage error correction is integrated in the general framework of CR estimation. A cell count represents the size of a group in the combined source, where a group is defined by its source(s). This linkage error-corrected contingency table may include multiple sources and covariates and underlies the CR model. Using the log-linear Poisson regression model, statistical inference on this table can be accomplished in the same way as in this model without linkage errors. In this way, we derive a CR estimation procedure that corrects for linkage errors but can deal with any number of linked sources and covariates.

In Section 2 we introduce some notation and discuss the general problem of linkage errors in CR models. In Section 3 we first discuss CR models in general and corresponding linkage error correction methods known in literature. In the same section we combine these to derive a general CR model framework that corrects for linkage errors and can deal with covariates and multiple sources. We refer to this model as the weighted multiple-recapture (WMR) model. The expression 'weighted' comes from the individual record weights that we will introduce in Subsection 3.6. Section 4 presents a simulation study that shows that the model works, and Section 5 concludes and discusses the results.

2. Notation and An Illustration of Linkage Errors

In this section we introduce the notation that we use to describe our model. Because our model involves linkage errors, we discuss source linkage first. Imagine there is some linkage procedure ℓ that links a set of sources with linkage keys. A linkage key can either be a perfect identifier γ, like a flawless ID number, or some set of Z imperfect identifiers $\tilde{\gamma} = \tilde{\gamma}_1, \ldots, \tilde{\gamma}_Z$, such as non-unique names or names that are not spelled flawlessly. Where γ is available, linkage can be performed without linkage errors, and where $\tilde{\gamma}$ is available, source linkage might contain errors.

Where more than two sources are available, sources can be linked either simultaneously, pairwise, or sequentially. Simultaneously means that all sources are linked in one step. Pairwise means that different sets of sources are linked first, after which these linked sources are linked again until all sources are linked into one linked source. Sequential linkage means that first two sources are linked, then the next is linked to this source, and so on, until no sources remain. Each step of sequential linkage could be considered a special version of pairwise linkage.

Where γ is available, there is no difference between simultaneous, pairwise, and sequential linkage; they lead to the same result. However, if only $\tilde{\gamma}$ is available this equality does not necessarily hold. For instance, in the case of pairwise linkage, records might be linked inconsistently (e.g., $A \rightarrow B, B \rightarrow C, C \nrightarrow A$). This inconsistency is not possible in simultaneous and sequential linkage. However, simultaneous linkage has the problem that it can become computationally very intensive, because the number of potential matches increases exponentially with every source. Therefore, as was also

argued by DC&T_18, sequential linkage is usually preferred in practice. Therefore, we assume that source linkage by ℓ is performed sequentially.

2.1. Linkage With Perfect Identifiers

We first discuss the situation for perfect identifiers. Let there be K sources S^k ($k = 1, \ldots, K$). Each source S^k contains s^k records that represent a set of population units. We assume that the units in each source are a subset of units from the population that has unknown size m. We assume that each source contains a *perfect* matching key γ that can be used in the (sequential) linkage procedure ℓ. ℓ starts with linking S^1 and S^2 and so on until S^K is linked, which implies a total of $K - 1$ linkages. After each step, the resulting linked source is referred to as N^k after each step. This can be written as:

$$N^k = \begin{cases} N^1 = S^1 \\ N^2 = \ell\left(N^1, S^2, \gamma\right) \\ \vdots \\ N^K = \ell\left(N^{K-1}, S^K, \gamma\right) \end{cases}, \tag{1}$$

where N^k consists of n^k records with $n^k < m$.

2.2. Linkage Without Perfect Identifiers

When instead of a perfect, an *imperfect* linkage key $\tilde{\gamma}$ is available, linkage errors can occur. The number of linkage errors may be reduced with probabilistic linkage models (see e.g., Fellegi and Sunter 1969; Winkler 1988; Jaro 1989). Probabilistic linkage models generally use imperfect linkage keys to minimise both the probability of mismatches and missed matches and find the optimal balance between these two. They estimate, for each possible pair of records, a probability of this pair being a match. For example, when two records have almost the same and unique name, the probabilistic linkage model estimates this pair to have a high probability of being a match and links them. The concepts behind these estimated probabilities will be discussed in more detail in section 3, because they are at the base of the D&F model and its successors.

We defined N^k as the combined source that is obtained in case a perfect linkage key γ is available. With the imperfect linkage key $\tilde{\gamma}$, N^k is replaced by $\tilde{N}^k$ and can be written as:

$$\tilde{N}^k = \begin{cases} \tilde{N}^1 = S^1 \\ \tilde{N}^2 = \ell\left(\tilde{N}^1, S^2, \tilde{\gamma}^1\right) \\ \vdots \\ \tilde{N}^K = \ell\left(\tilde{N}^{K-1}, S^K, \tilde{\gamma}^K\right) \end{cases}, \tag{2}$$

where $\tilde{\gamma}^k$ refers to the set of imperfect linkage key variables that is available in linkage k and $\tilde{N}^k$ has $\tilde{n}^k$ records and may contain mismatches and/or missed matches.

While with perfect linkage it is certain that the number of records n^k is a lower bound for the population size m, this does not hold for $\tilde{n}^k$. Due to imperfect linkage, $\tilde{n}^k$ can be smaller, equal, or larger than m, but also smaller, equal or larger than n^k, because a missed match increases the number of records and a mismatch reduces the number of records in $\tilde{N}^k$.

2.3. *Records and Cell Counts*

N^k and $\tilde{N}^k$ are combined sources with n^k and $\tilde{n}^k$ records, where a record may represent multiple individuals. A single record r is referred to as N_r^k and $\tilde{N}_r^k$ with $r = 1, \ldots, n^k$ and $r = 1, \ldots, \tilde{n}^k$ respectively. Each of these records contains a string of k binary indicators $S^1 \ldots S^k$ that indicates in which source a record occurs, denoted as $N^k_{r,S^1 \ldots S^k}$ (e.g., $N^k_{r,S^1S^2} = N^k_{r,11}$ means the subset of records r that are in both S^1 and S^2). Each subset $N^k_{r,S^1 \ldots S^k}$ has a corresponding cell count denoted as $n_{S^1 \ldots S^k}$, which is simply the number of records in subset $N^k_{r,S^1 \ldots S^k}$. These binary indicators are a fundamental part of CR models because they define the cell count levels in the contingency table and serve as explanatory variables. Corresponding with the combined sources N^k and $\tilde{N}^k$ we define A^k and $\tilde{A}^k$ with the unique set of strings and the observed cell counts adding up to n^k and $\tilde{n}^k$. For example, under perfect linkage the unique set of strings collected in $A^3 = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ has a corresponding vector of counts $n_{S^1S^2S^3}$ where the subscripts $S^1S^2S^3$ refer to a row in A^3. For instance, $n_{S^1S^2S^3} = n_{111}$ is the count that belongs to the first row in A^3.

The observed cell counts $n_{S^1S^2S^3}$ can be considered realisations of a random process, so they also have an expectation that we refer to as $m_{S^1S^2S^3}$. For $m_{S^1S^2S^3}$ we have the equality $\sum_{S^1S^2S^3} m_{S^1S^2S^3} + m_{000} = m$, where m_{000} is the expected number of units in the population missed by S^1, S^2 and S^3. Estimates of $m_{S^1S^2S^3}$, m_{000} and m based on $n_{S^1S^2S^3}$ are denoted with a '$\wedge$', for example $\hat{m}_{S^1S^2S^3}$ and based on $\tilde{n}_{S^1S^2S^3}$ are denoted with a '$\vee$', for example $\breve{m}_{S^1S^2S^3}$.

Finally we note that the definition of A^k above allows for a straightforward extension when categorical covariates are to be included in the process, by adding dummy variables as columns and adding rows such that $S^1 \ldots S^k$ are represented separately for the distinct levels of the covariates. Interactions between the sources, and between sources and covariates, can be included by adding columns appropriately.

2.4. *An Illustration of Source Linkage, Linkage Errors and the Contingency Table*

Figure 1 illustrates the simple case of the linkage of $k = 2$ sources with the imperfect linkage key $\tilde{\gamma}^1$ and the five linkage error types (1–2d) discussed in Section 1.

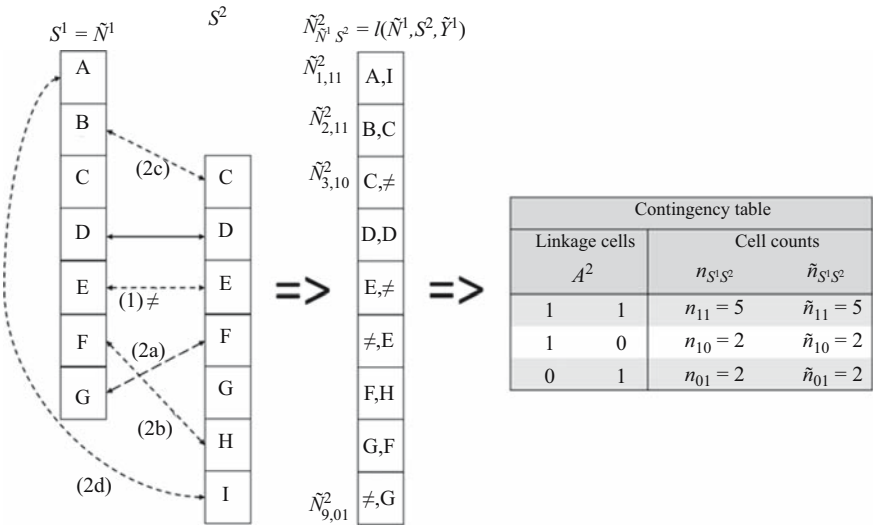

Fig. 1. Illustration of linkage of two sources and different types of linkage errors.

The illustration in Figure 1 presents two imperfectly linked sources of equal size $s^1 = s^2 = 7$. The total number of units in S^1 or S^2 is nine, and the units are labelled A to I. The solid line arrow represents a correct record linkage while the dashed line arrows represent five other linkages that all correspond to one of the linkage error types (1 – 2d). The resulting combined source $\tilde{N}^2$ contains the nine records $\tilde{N}^2_r \left(r = 1, \ldots, 9\right)$ and each record belongs to one of the subsets $\tilde{N}^2_{r,S^1S^2}$. Under perfect linkage each record in N^2 should correspond to one unique unit in S^1 and S^2. This does not hold in case of linkage errors. In fact, in this artificial example the only correct match is [D, D] while all other records represent missed or mismatches. Despite the linkage errors, in this case it (coincidentally) does not lead to errors in the cell counts. The reason is that in this artificial example the five different linkage error types cancel each other out. Obviously, ignoring linkage errors generally leads to a difference between $n_{S^1S^2S^3}$ and $\tilde{n}_{S^1S^2S^3}$. The question we deal with in Section 4 is how we can correct $\tilde{n}_{S^1S^2S^3}$ in such a way that this correction is an unbiased estimate of $n_{S^1S^2S^3}$. But to see why this is useful, first we discuss CR models in Section 3.

3. Linkage Error Correction in Capture-Recapture Estimation

In this section we describe and discuss DS models and the linkage error correction method introduced by D&F. We first describe the most basic DS model which was introduced by Petersen (1896) and is also known as the Lincoln-Petersen model (Lincoln 1930). Next, we show how D&F improve this model so that it corrects for linkage errors. We further discuss DC&T_15, DC&T_18 and WLZ, because they provide the tools that help us to show why correction of the contingency table also corrects for linkage errors.

3.1. Relation Between the Basic Dual-System and the Log-Linear Poisson Regression Model

In the DS model A^2 has three rows, with associated expected cell counts. The maximum likelihood (ML) estimates $\left(\hat{m}_{11}, \hat{m}_{10}, \hat{m}_{01}\right)$ are equal to $\left(n_{11}, n_{10}, n_{01}\right)$ because the DS

model is saturated. Under the appropriate assumptions (Wolter 1986), including perfect linkage, the basic DS estimate can be obtained by:

$$\hat{m}_{\mathrm{DS}} = \hat{m}_{11} + \hat{m}_{10} + \hat{m}_{01} + \hat{m}_{00} = n_{11} + n_{10} + n_{01} + \frac{n_{10}n_{01}}{n_{11}} = \frac{s^1 s^2}{n_{11}}, \tag{3}$$

where $\hat{m}_{00}$ represents an estimate of the unobserved part of the population and $\hat{m}_{\mathrm{DS}}$ is the estimate for the population size. The expression $\frac{s^1 s^2}{n_{11}}$ simply follows from $s^1 = n_{11} + n_{10}$ and $s^2 = n_{11} + n_{01}$. This expression will become important, because it contains only one value (n_{11}) that can be affected by linkage errors because S^1 and S^2 are simply the size of S^1 and S^2, which are unaffected by linkage errors.

The population size can also be estimated using the log-linear Poisson model (e.g., see Cormack 1989). The log-linear Poisson regression model for A^2 can be written as:

$$m_{S^1S^2} = e^{(\beta_0 + \beta_1 S^1 + \beta_2 S^2)}. \tag{4}$$

Using the estimate of the intercept an estimate $\hat{m}_{00}$ can be obtained as $\hat{m}_{00} = e^{\hat{\beta}_0}$. Because the ML estimate $\hat{\beta}_0$ in (4) is $\hat{\beta}_0 = \log(n_{10}) + \log(n_{01}) - \log(n_{11})$, the equality $e^{\hat{\beta}_0} = \frac{n_{10}n_{01}}{n_{11}}$ also holds. This equality shows why Equations (3) and (4) lead to the same result. However, an important advantage of the log-linear formulation is that it can be easily extended with additional sources or categorical covariates and the interaction between them. For instance, with a third source and a categorical covariate x with levels 1 and 0, then $m_{S^1S^2}$ becomes $m_{S^1S^2S^3X}$ and the model might for instance be:

$$m_{S^1S^2S^3X} = e^{(\beta_0 + \beta_1 S^1 + \beta_2 S^2 + \beta_3 S^3 + \beta_4 S^1S^2 + \beta_5 S^1S^3 + \beta_6 X)}.$$

Extending the Petersen formula in this way would be non-trivial at best, while for each category in x a PSE of the unobserved population can be obtained by $\hat{m}_{0000} = e^{\hat{\beta}_0}$ and $\hat{m}_{0001} = e^{\hat{\beta}_0 + \hat{\beta}_6}$.

3.2. *Impact of Linkage Errors on the Dual-System Model*

We provide a simple numerical example that illustrates the problem of linkage errors in the DS model. We take $s^1 = 300$, $s^2 = 150$ and $n_{11} = 100$. Due to linkage errors $\tilde{n}_{11} = 90$. This difference between n_{11} and $\tilde{n}_{11}$ implies that the number of missed links is 10 more than the number of false links. This simple case is represented in Table 1.

An estimate for m_{00}, $\hat{m}_{00} = \frac{n_{10} * n_{01}}{n_{11}} = \frac{200*50}{100} = 100$. However, due to linkage errors not n_{ij}, but $\tilde{n}_{ij}$ is observed and when this is naively ignored the DS estimate becomes: $\breve{m}_{00} = \frac{\tilde{n}_{10}\tilde{n}_{01}}{\tilde{n}_{11}} = \frac{210*60}{90} = 140$, leading to a linkage error bias of 40, something better not left ignored. Note that the '$\breve{}$' on $\breve{m}_{00}$ means that $\breve{m}_{00}$ is an estimate based on cell counts that are subject to linkage errors.

3.3. *The D&F and D&F + Model*

The D&F model is a DS model that aims to correct the population size estimate for linkage errors type (1) and (2d) (cf. Section 1). We refer to this estimate as $\hat{m}_{\mathrm{D\&F}}$. To estimate the linkage error probabilities of these two error types, they use a rematch study. A rematch study aims to confirm whether a subset of matches and non-matches were correct or not. The rematch study can be summarized as in Table 2.

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

A^2		$n_{S^1S^2}$	$\tilde{n}_{S^1S^2}$
1	1	100	90
1	0	200	210
0	1	50	60

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 in the rematch study were correctly matched (a_{11}), correctly not matched (a_{00}), incorrectly matched (a_{10}) and incorrectly not matched (a_{01}). They define the probability of linkage error type (1) by α and of type (2d) by θ. Thus, $\alpha = \frac{a_{11}}{a_{11}+a_{01}}$ and $\theta = \frac{a_{10}}{a_{10}+a_{00}}$ and D&F show how to use these probabilities to obtain $\breve{m}_{D\&F}$ that corrects for linkage errors (1) and (2d). The D&F model recently received more attention from DC&T_15 and DW. DC&T_15 write $\breve{m}_{D\&F}$ as:

$$\breve{m}_{D\&F} = \frac{\tilde{n}_{11} + \tilde{n}_{10} + \tilde{n}_{01}}{\hat{p}_1 + \hat{p}_2 - (\alpha - \theta)\hat{p}_1\hat{p}_2 - \theta\hat{p}_1} \tag{5}$$

with

$$\hat{p}_1 = \frac{-N_{11} + \theta(\tilde{n}_{11} + \tilde{n}_{10})}{(\theta - \alpha)(\tilde{n}_{11} + \tilde{n}_{01})},\ \hat{p}_2 = \frac{-\tilde{n}_{11} + \theta(\tilde{n}_{11} + \tilde{n}_{10})}{(\theta - \alpha)(\tilde{n}_{11} + \tilde{n}_{01})}.$$

These equations show that the D&F model is complex and hard to interpret. The formulas become even more complex when DC&T_15 introduce their so called two-way linkage errors. This model is further extended by WLZ, who show that, in the calculation of the two-way linkage errors, it is implicitly assumed that the sizes of source 1 and 2, s^1 and s^2, are equal. Therefore, they extend the model with asymmetrical two-way errors, which allows for s^1 and s^2 to be different. Unfortunately, this implies introducing more notational complexity (as θ is separated into θ_1 and θ_2). In DC&T_18 the linkage error correction model is extended from two to three sources. DC&T_18 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 model. This is 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.

Beside WLZ's asymmetrical two-way errors extension, they provide us with another useful contribution. They show that the D&F model, the DC&T_15 model and their own extension all give identical outcomes when not only the formula of $\hat{m}_{D\&F}$ but also of α and θ are chosen appropriately. They also show that in this case the model corrects for all five linkage error types introduced in Section 1. We refer to this model as the D&F + model.

WLZ also show that this can be written much more comprehensively as:

$$\breve{m}_{\mathrm{D\&F+}} = \frac{s^1 s^2}{\breve{n}_{11}}, \tag{6}$$

where $\breve{n}_{11}$ is an estimate for m_{11} based on $\tilde{n}_{11}$ and the rematch study, instead of the directly observed n_{11} used in the DS model. Equation (6) shows that the models derived in D&F, DC&T_15 and WLZ are all equal and a generalisation of the DS estimator. In the next section we will show that $\breve{n}_{11}$ can be derived in a straightforward way when the rematch study is used in a slightly different way. Unlike in WLZ it will no longer depend on α and θ altogether.

3.4. Further Simplification of the D&F + Model

The D&F + model as defined in Equation (6) contains only one element that is susceptible to linkage errors; that is, $\breve{n}_{11}$. WLZ derive $\breve{n}_{11}$ starting with the a's in Table 2. These are used to estimate α and θ that in turn are used to estimate $\breve{n}_{11}$. In this section we propose to simplify this procedure by using the rematch study differently. The rematch study concerns a representative subsample of the population of which the matches and non-matches were clerically reviewed. This means that, for the records in this subset, it is quite simple to count the number of matches before and after clerical review. We refer to the set of records that are subject to clerical review with a '*'. This implies $\tilde{N}^{k^*}$ and $\breve{N}^{k^*}$ are the sets of linked records between $\breve{N}^{k-1}$ and S^k, that were under clerical review, before and after clerical review. The overlap count of the records in the clerical review study before and after clerical review are denoted as $\tilde{n}_{11}^{k^*}$ and $\breve{n}_{11}^{k^*}$. Then the ratio $\frac{\breve{n}_{11}^{k^*}}{\tilde{n}_{11}^{k^*}}$ can be used to estimate $\breve{n}_{11}$ with:

$$\breve{n}_{11} = \tilde{n}_{11} \frac{\breve{n}_{11}^{k^*}}{\tilde{n}_{11}^{k^*}}. \tag{7a}$$

For $k = 2$ the elements $\breve{n}_{10}$ and $\breve{n}_{01}$ can be obtained by:

$$\breve{n}_{10} = s^1 - \breve{n}_{11}, \tag{7b}$$

$$\breve{n}_{01} = s^2 - \breve{n}_{11}. \tag{7c}$$

Note that we write $\breve{n}_{11}^{k^*}$ instead of $n_{11}^{k^*}$, although for $k = 2$ they are equal. As we will see later, for $k > 2$ this equality no longer holds, because then $n_{11}^{k^*}$ is no longer a simple count but a sum of weights unequal to 1. Elements (7a–c) serve as input for the saturated model as defined in (4); that is, $\breve{n}_{S^1S^2} = e^{(\breve{\beta}_0 + \breve{\beta}_1 S^1 + \breve{\beta}_2 S^2)}$, which gives $\breve{m}_{00} = e^{\breve{\beta}_0}$. This implies that, by combining (4) with (7a–c) in the basic DS model, we have obtained the PSE $\breve{m}_{\mathrm{D\&F+}}$, with a simple set of formulas. In the next section we show how these formulas can be extended such that they can deal with covariates and additional sources.

3.5. Covariates in the D&F + Model

We proceed by a further development of DS model in the context of the log-linear Poisson regression model with categorical covariates. When there is only one categorical covariate X

with $X \in (0, 1)$, then n_{110} is the number of records in S^1 and S^2 with $X = 0$. Note that while without covariates we had $n_{10} = s^1 - n_{11}$, with covariates this can be replaced by $n_{10X} = s^1_X - n_{11X}$ where s^1_X refers to the number of records in S^1 for each level in X. This gives us a straightforward way to incorporate covariates in the D&F + model, because we can simply replace the subscript S^1S^2 in Equation (7a–c) with the subscript S^1S^2X, which gives:

$$\hat{n}_{11X} = \tilde{n}_{11X} \frac{\breve{n}^*_{11X}}{\tilde{n}^*_{11X}}, \tag{8a}$$

$$\breve{n}_{10X} = s^1_X - \breve{n}_{11X}, \tag{8b}$$

$$\breve{n}_{01X} = s^2_X - \breve{n}_{11X}. \tag{8c}$$

Equations (8a–c) yield $(\breve{n}_{11X}, \breve{n}_{10X}, \breve{n}_{01X})$ that can be used as values of the dependent variable in the log-linear Poisson regression model that includes the covariate X as explanatory variable. This can be extended in a straightforward way for more explanatory variables, as was described in Subsection 3.1. This approach has the advantage that it allows for parsimonious models. For example, it may turn out that some parameters that estimate the effect of covariates do not depart significantly from zero and the model can therefore further ignore this covariate. This option of hypothesis testing is an important improvement over the D&F + model.

Working with a saturated model will induce redundant noise in the DS model, when a more parsimonious model fits adequately. Therefore, significance testing of covariates is important, and becomes increasingly so when the number of covariates in the CR model increases.

Without discussing technical details, we elaborate on the role of X. It is important to include X in the CR model when the capture probabilities are heterogeneous over S^1 and S^2, and X takes this into account. However, it is not necessarily the case that the levels of X differ with respect to linkage error probabilities as well. For instance, records with $X = 1$ might be more likely to be in S^1; however, they are not necessarily also more likely to be falsely linked or not linked to S^2. In this case, despite the significance of X in the CR model, the ratios $\frac{n^*_{S^1S^21}}{\tilde{n}^*_{S^1S^21}}$ and $\frac{n^*_{S^1S^20}}{\tilde{n}^*_{S^1S^20}}$ will not differ significantly and X can be ignored in the linkage error correction step. The cell counts of records with $X = 1$ and $X = 0$ can both be corrected with the same ratio $\frac{n^*_{S^1S^2}}{\tilde{n}^*_{S^1S^2}}$. Therefore, in practice one may first test whether the ratios $\frac{n^*_{11X}}{\tilde{n}^*_{11X}}$ differ significantly from each other for different levels within X.

3.6. *Additional Sources in the D&F + Model*

Equation (7) can be applied on the contingency table of the combined source $\tilde{N}^2$ (this also holds for Equation (8), but we further ignore this to keep the presentation simple). When a third source is involved, it must be linked to $\tilde{N}^2$ again. However, $\tilde{N}^2$ was not affected by Equation (7), so simply linking S^3 to $\tilde{N}^2$ would ignore the linkage error correction in Equation (7). Therefore, before the next source is linked, the information obtained in this linkage error correction step must somehow be transferred to (the records in) $\tilde{N}^2$. A straightforward way to do this, is to disaggregate $\breve{n}_{S^1S^2}$ back to the record level, simply by

distributing $\breve{n}_{S^1S^2}$ evenly over the corresponding records. In other words, each record $\tilde{N}^2_r$ in $\tilde{N}^2$ receives a weight $w^2_r = \frac{\breve{n}_{S^1S^2}}{n_{S^1S^2}}$. We refer to the combination of $\tilde{N}^2$ and the corresponding vector of linkage error correction weights w^2 as $\breve{N}^2$. $\breve{N}^2$ may now be linked to S^3, giving $\tilde{N}^3$, which may introduce new linkage errors. $\tilde{N}^3$ can be used to obtain $\tilde{m}_{\breve{N}^2S^3}$ by summing up over w^2_r for the records in $\breve{N}^2$ while (new) records in S^3 receive a weight $w^2_r = 1$. This gives cell counts that are corrected for linkage errors in going from S^1 to S^2 but not yet in going from $\breve{N}^2$ to S^3. To correct for these new linkage errors the linkage error correction step in Equation (7) can be repeated to transform $\tilde{m}_{\breve{N}^2S^3}$ into $\breve{m}_{\breve{N}^2S^3}$. Where more sources are linked, this linkage error correction procedure can be repeated after each new source.

This procedure of linking two sources, aggregating this combined source to a contingency table, correcting the cell counts for linkage errors, disaggregation of the contingency table back to the combined source and again linking a new source, is quite cumbersome. This procedure becomes more straightforward when the linkage error correction step in Equation (7) is performed directly on the record level weights w^k_r. Then, only after the last source is linked, a contingency table that is corrected for linkage errors is produced by summing up over the weights for the corresponding categories. This can be written more formally by an updating scheme for w^k_r with $w^1_r = 1$:

$$w^k_r = \begin{cases} w^{k-1}_r \dfrac{\breve{n}^{k^*}_{11}}{\tilde{n}^{k^*}_{11}} & \text{for } r = \in \tilde{N}^k_{r,11} \\ w^{k-1}_r \dfrac{\breve{n}^{k^*}_{10}}{\tilde{n}^{k^*}_{10}} = \dfrac{\breve{n}^{k-1^*} - \breve{n}^{k^*}_{11}}{\breve{n}^{k-1^*} - \tilde{n}^{k^*}_{11}} & \text{for } r = \in \tilde{N}^k_{r,10} \\ 1 \dfrac{\breve{n}^{k^*}_{01}}{\tilde{n}^{k^*}_{01}} = \dfrac{s^{k^*} - \breve{n}^{k^*}_{11}}{s^{k^*} - \tilde{n}^{k^*}_{11}} & \text{for } r = \in \tilde{N}^k_{r,01} \end{cases} \tag{9}$$

Where $S^{k^*} = \sum_{r \in (\breve{N}^{k^*}_{r,11}, \breve{N}^{k^*}_{r,01})} w^{k-1}_{r,}$, $\breve{n}^{k-1^*} = \sum_{r \in (\breve{N}^{k^*}_{r,11}, \breve{N}^{k^*}_{r,10})} w^{k-1}_r$, $\breve{n}^{k^*}_{11} = \sum_{r \in \breve{N}^{k^*}_{r,11}} w^{k-1}_r$ and $\tilde{n}^{k^*}_{11} = \sum_{r \in \tilde{N}^{k^*}_{r,11}} w^{k-1}_r$. Note that records with $r \in \tilde{N}^k_{r,01}$ are always new records that were not linked in the $k-1$ previous linkage steps. Therefore, their (individual starting) weight is simply (still) equal to 1, because they were not updated in any of the previous updating steps. Furthermore, note that, if there is reason to believe some covariate groups may be more susceptible to linkage errors than others, Equation (9) may be applied for these groups separately.

Generally, the record-level linkage error correction weight w^k_r is a weight that can be interpreted in a similar way as well-known individual sample weights in survey models. In survey models, individual sample weights allow a researcher to correct for over- and underrepresentation of specific groups in a survey. A record with a higher-than-average weight belongs to a group that is underrepresented and vice versa for a record with a low weight. Similarly, a record with a higher or lower-than-average linkage error correction weight belongs to a group with a cell count that is under- or overestimated, respectively. With individual sample weights, it is quite common to sum up over these weights to obtain representative totals. For instance, when the number of men is underrepresented, summing up over their sample weights gives the number of men that is corrected for this underrepresentation. The same reasoning holds for the record-level linkage error correction weights.

By applying Equation (9) after each source linkage, a contingency table that is corrected for linkage errors can be constructed after every source linkage. This contingency table is different from a regular contingency table that simply counts the number of records in each linkage cell. The linkage error-corrected contingency table is constructed by summing up the weights of these records over these linkage cells instead of counting records. Therefore, we refer to the models based on this contingency table as the weighted dual-system (WDS) model for two sources and the weighted multiple-recapture (WMR) model for more than two sources. Where there are no linkage errors, the models reduce to the standard DS and MR models.

4. Simulation Study

We evaluate the WMR model with a simulation study. The main goal of this simulation study is to examine whether our new WDS and WMR models behave under different conditions, such as (no) linkage errors, (no) covariate dependence, (no) source dependence and combinations thereof. In Subsection 4.1 we describe the setup of this simulation study and in Subsection 4.2 we discuss the results.

4.1. Simulation Study Setup

For the simulation study to reflect reality as well as possible, we use a quasi-real data set that is publicly available and represents a fictitious population dataset of 26,625 persons. It is constructed such that it is representative for the UK census population. It was created in the ESSnet DI (McLeod et al. 2011), a European project on data integration (Record Linkage, Statistical Linking, Micro integration Processing) that ran from 2009 to 2011. The data set has linkage keys such as address and birthdate but also covariates such as gender and age.

In each replication of the simulation study, a random population of 10,000 is generated. This size of 10,000 is chosen because the Poisson regression estimators have known finite sample bias (see e.g., Chapman 1951, Menkens and Anderson 1988; Chen and Giles 2009). This bias goes to zero when the sample increases to infinity.

For, say, a population size of 1,000, this bias may still play a role, so then it will be hard to determine whether a CR model corrects for linkage error bias. A probable example of this finite sample bias can be found 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 when we experimented with a TPS of 1,000.

Unfortunately, the population size can also not be too large because probabilistic record linkage is computationally highly intensive. A population size of 10,000 is a balanced choice that leaves the finite sample bias practically ignorable and leaves the probabilistic linkage procedure computationally feasible.

This population of 10,000 serves to generate three sources that each cover part of the population. These sources are generated under different conditions where conditions vary with respect to covariate and source dependence. This leads to four scenarios:

1. Three randomly generated sources (no dependencies),
2. Three sources in which covariates affect the probability of a record to be in a source (covariate dependence),
3. Three sources where the probability of a record to be in a source is affected by this record being in other sources (source dependence), and
4. Three sources where records are subject to both covariate and source dependence.

Next, in each replication the sources are linked both with and without linkage errors. The linked sources allow us to apply both the regular (referred to as naïve) and weighted DS (using only the first two sources) and MR (using all three sources) model. By replicating this procedure many times (i.e., 1,050 for each scenario) we can obtain a distribution of estimates that, in the case of the model providing asymptotically unbiased estimates, will evolve around the TPS of 10,000. In this way, we can see whether the WDS and WMR model can deal with covariate and source dependencies while suffering from linkage errors, conditions under which the regular DS and MR model fail. A more detailed description of the simulation setup can be found in Appendix, Section 6.

4.2. Simulation Results

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

Figure 2 contains twelve density plots that each contain distributions/densities of DS and MR estimates. In the rows there are the four scenarios and the first two columns distinguish naïve estimation with and without linkage errors. The third column shows the WDS and WMR model in the case of linkage errors.

The graph clearly shows that the estimates that can be expected to be biased, are biased. Most importantly, however, it shows that, in the case of linkage errors, the weighted estimates are on target while naïve estimates are biased. Furthermore, the presence of covariate dependence is no problem for the weighted estimates, even in combination with source dependence. A numerical calculation example of one of the replications can be found at the end of the Appendix.

5. Discussion

In this article 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. (2019) 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. Thus linkage error no longer has to be studied as an isolated issue in CR 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. For instance, it still requires a rematch study in which, for a share of records, clerical review is required to check whether they were correctly linked or not linked. Ideally these records are representative for the records in both sources, both with respect to covariates but also the quality of linkage keys. This last element should not be underestimated. When for instance the

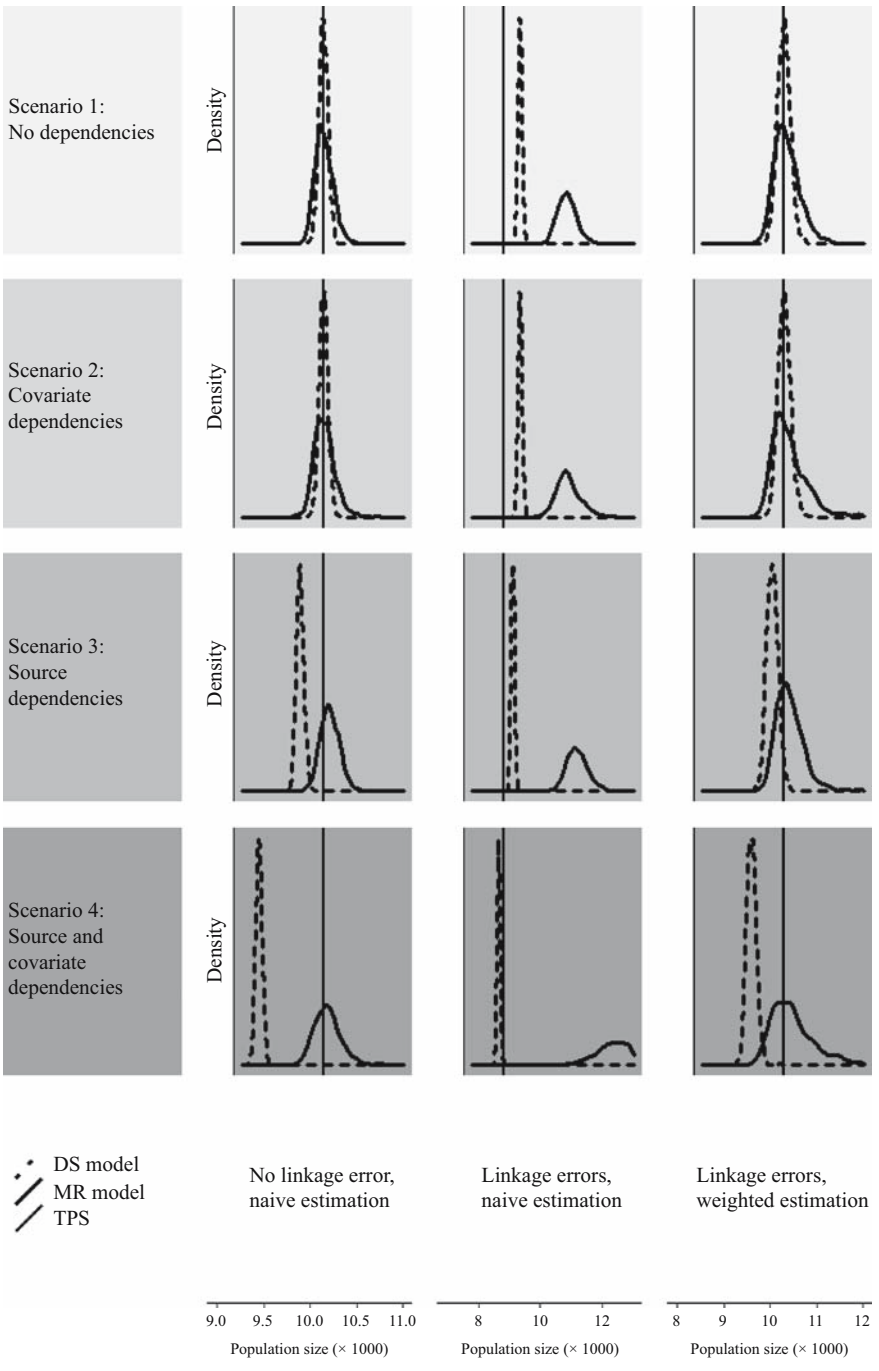

Fig. 2. Density plots of two PSEs with three dependent variables and four scenarios.

records in the rematch study are based on their high-quality linkage keys (which makes clerical review easier), they might suffer less from linkage errors than other records. This will lead to a biased correction. Another issue is the size of the rematch study, when the sources

contain some small groups of records, it might be hard to find enough records of this group to perform clerical review. The extent of the impact of such issues 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 the 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 practice, 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. The proportion of the sample weights of these new records in relation to the weights of records that were already known from previous records can be used to improve these starting weights. Furthermore, when additional sources also contain sample weights, they can be used to construct the cell counts in the contingency table by adding up over weights instead of counting the records. In this way we would get 'linkage error corrected sample weights'. How and when sample weights can be combined with linkage and linkage error correction requires further research.

6. Appendix

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. ('PERNAME2', 'DOB_DAY' and DOB_MON served as linkage variables, which corresponds to the 'bronze scenario' in DC&T_15). 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 of being selected. Furthermore, the list contains several covariates, of which we use 'SEX' as covariate X to affect capture probabilities.

The population and sources generation, record linkage and estimation procedure is replicated 1,050 times. This number is 'only' 1,050 because we use a Spark cluster of 15 cores (available at Statistics Netherlands mainly for Big Data related computations) that each do 70 replications with different random seeds, in which each single replication takes about ten minutes. In total it took almost two days to run all four scenarios, which is mainly due to the computation time of the probabilistic linking of the three sources.

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 independent of these other sources. Therefore, before the first source is generated, we first generate three so called latent sources $U^k = \big(k = 1, 2, 3\big)$ of 8,000 units each, 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 S^k is independent of the other sources. This is done by giving each unit $u = 1, \ldots, 10,000$ a probability to be in each source k by:

$$P_u^k\big[S^k = 1\big] = \frac{1}{1 - \exp\big(-\mu_u^k\big)} \tag{10}$$

where $\mu_u^k = \delta_{U^1}^k U_u^1 + \delta_{U^2}^k U_u^2 + \delta_{U^3}^k U_u^3 + \delta_X^k X_u$. Given Equation (10) we can vary $\delta' s$ and thereby control dependencies between any source in S^k and the other two sources in S^k and the covariate. For instance, when $\delta_{U^1}^1, \delta_{U^1}^2, \delta_{U^2}^1, \delta_{U^2}^2 \neq 0$, 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 are in Table 3.

Because the varying of $\delta' s$ affects the capture probabilities of units, different $\delta' s$ also correspond to different estimates of the $\hat{\beta}' s$ from the Poisson regressions. To assure that by varying $\delta' s$ we introduce a substantial source and covariate dependence. To show that this works, Table 4 presents the (corresponding) mean values of estimated $\hat{\beta}' s$ over all replications of the benchmark case of no linkage errors.

Table 4 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 do

Table 3. Parameter values of the four different scenarios.

Scenario 1	$\delta_{U^1}^k$	$\delta_{U^2}^k$	$\delta_{U^3}^k$	δ_X^k
μ_u^1	6.3	0	0	0
μ_u^2	0	3.5	0	0
μ_u^3	0	0	1.9	0
Scenario 2	$\delta_{U^1}^k$	$\delta_{U^2}^k$	$\delta_{U^3}^k$	δ_X^k
μ_u^1	5.6	0	0	2
μ_u^2	0	4.6	0	−2
μ_u^3	0	0	0.42	2
Scenario 3	$\delta_{U^1}^k$	$\delta_{U^2}^k$	$\delta_{U^3}^k$	δ_X^k
μ_u^1	4.8	1.8	0	0
μ_u^2	0	3.5	0	0
μ_u^3	0	−0.5	2.3	0
Scenario 4	$\delta_{1,4}$	$\delta_{2,4}$	$\delta_{3,4}$	$\delta_{4,4}$
μ_u^1	3.9	1.5	0	2
μ_u^2	1.5	3.3	0	−2
μ_u^3	0	−0.5	1.8	1

Table 4. Average estimated $\hat{\beta}$'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'.

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.

Finally, the last necessary elements of the simulation study are $\breve{N}^{2*}$ and $\breve{N}^{3*}$, which are generated by first selecting a random 10% (without replacement) of the population and within this selection only keeping those records that are also in S^1 and S^2 (for $\breve{N}^{2*}$) or S^2 and S^3 (for $\breve{N}^{3*}$).

We compare three types of PSEs, naïve, perfect, and weighted. Naïve PSEs are estimates based on $\tilde{n}$, so linkage errors are present but ignored. Perfect PSEs are based on n, so linkage errors are not present (and ignored). Weighted PSEs are based on $\breve{n}$, so linkage errors are present but if the model works it should correct for them. Finally, for each scenario and PSE type, the DS and MR model are applied.

6.1. Numerical Calculation Example

As an illustration of the method, we present one of the replications generated under scenario 4 in the simulation study. In Table 5 we show the total cell counts with linkage

Table 5. Contingency table and correction of weights after linking S^1 and S^2

Contingency table source S^1 and S^2 ($\tilde{N}^2$) and weight correction						
Linkage cells		Covariate	Cell counts			Weight correction
S^1	S^1	X	$\tilde{n}_{S^1S^2}$	$n^*_{S^1S^2}$	$n^*_{S^1S^2}$	w_r^2
1	1	0	2,784	264	222	$r \in N^2_{r,110} : w_r^2 = 232/264$
1	0	0	1,080	138	180	$r \in N^2_{r,100} : w_r^2 = 180/138$
0	1	0	164	12	64	$r \in N^2_{r,010} : w_r^2 = 64/12$
1	1	1	2,030	226	152	$r \in N^2_{r,110} : w_r^2 = 152/226$
1	0	1	2,292	240	314	$r \in N^2_{r,101} : w_r^2 = 314/240$
0	1	1	82	10	44	$r \in N^2_{r,011} : w_r^2 = 44/10$

Table 6. Contingency table and correction of weights after linking $\tilde{N}^2_{S^1S^2}$ and S^3

Contingency table source $\tilde{N}^2_{S^1,S^2}$ and S^3 ($\tilde{N}^3$) and weight correction						
Linkage cells $\tilde{N}^2$	S^3	Covariate X	Sum of weights w_r^2 $\tilde{n}_{\tilde{N}^2S^3}$	$\tilde{n}^*_{\tilde{N}^2S^3}$	$\tilde{n}^*_{\tilde{N}^2S^3}$	Weight correction w_r^2
1	1	0	502.20	62.74	36	$r \in \tilde{N}^3_{S,110} : w_r^3 = w_r^{2*}232/264$
1	0	0	4,122.25	410.34	412	$r \in \tilde{N}^3_{S,100} : w_r^3 = w_r^{2*}412/410.34$
0	1	0	344	38	2	$r \in \tilde{N}^3_{S,010} : w_r^3 = 38/2$
1	1	1	1,789.00	194.14	174	$r \in \tilde{N}^3_{S,111} : w_r^3 = w_r^{2*}174/194.14$
1	0	1	2,935.81	272.39	296	$r \in \tilde{N}^3_{S,101} : w_r^3 = w_r^{2*}296/272.39$
0	1	1	1,798	188	14	$r \in \tilde{N}^3_{r,011} : w_r^3 = 14/188$

Table 7. Contingency table and correction of weights after linking $\tilde{N}^2$ and S^3.

Contingency table source S^1, S^2 and S^3 (N^3,$\tilde{N}^3$ and $\breve{N}^3$)						
Linkage cells A^3			Covariate X	Cell counts and sums of weights w_r^2 $n_{S^1S^2S^3}$	$\tilde{n}_{S^1S^2S^3}$	$\breve{n}_{S^1S^2S^3}$
1	1	1	0	200	344	165.99
1	1	0	0	2,328	2,440	2,060.10
1	0	1	0	82	106	79.34
1	0	0	0	1,254	974	1,275.56
0	1	1	0	44	14	42.85
0	1	0	0	704	150	803.23
0	0	1	0	18	344	18.11
1	1	1	1	452	766	461.74
1	1	0	1	872	1,264	910.45
1	0	1	1	1,102	866	1,015.47
1	0	0	1	1,896	1,426	1,998.07
0	1	1	1	150	32	126.19
0	1	0	1	310	50	235.61
0	0	1	1	94	1,798	133.89
Total				**9,506**	**10,574**	**9,327**
0	0	0	1	378.34	4,774.67	439.70
0	0	0	0	173.25	2,803.45	249.46
Total + Poisson regression estimates				**10,057.63**	**18,152.12**	**10,015.76**

errors together with the audit study cell counts. In the last column we show the correction of groups of individual weights.

Table 6 is similar to Table 5 but shows the linkage of $\tilde{N}^2_{S^1S^2}$ and S^3, together with the audit study. The last column shows the second update of weights.

Finally, Table 7 shows the contingency tables that underly the MR models. In the three rows at the bottom there are the different Poisson regression estimates of the unobserved parts of the population, for $X = 0$ and $X = 1$, together with the total population size estimate.

7. References

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

Cadwell, B.L., P.J. Smith, and A.L. Baughman. 2005. "Methods for capture-recapture analysis when cases lack personal identifiers." *Statistics in Medicine,* 24(13): 2041–2051. DOI: https://doi.org/10.1002/sim.2081.

Chao, A. 2001. "An Overview of Closed Capture-Recapture Models." *Journal of Agricultural, Biological, and Environmental Statistics* 6: 158–175. DOI: https://doi.org/10.1198/108571101750524670.

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

Chatterjee, K., and D. Mukherjee. 2018. "A new integrated likelihood for estimating population size in dependent dual-record system." *Can J Statistics* 46: 577–592. DOI: https:/doi.org/10.1002/cjs.11477.

Chen, Q., and D.E. Giles. 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 L. Kuo. 2001."A Note on the Estimation of the Multinomial Logit Model with Random Effects." *The American Statistician* 55: 89–95. DOI: https://doi.org/10.1198/000313001750358545.

Cormack, R.M. 1989. "Log-linear models for capture-recapture." *Biometrics,* 45: 395–413. DOI: https://doi.org/10.2307/2531485.

De Wolf, PP., J. van Der Laan, and D. Zult. 2019. "Connecting Correction Methods for Linkage Error in Capture-Recapture". *Journal of Official Statistics.* 35 (3): 577–597. DOI: https://doi.org/10.2478/jos-2019-0024.

Di Consiglio, L., and T. Tuoto. 2015. "Coverage evaluation on probabilistically linked data." *Journal of Official Statistics,* 31: 415–429. DOI: https://doi.org/10.1515/jos-2015-0025.

Di Consiglio, L., and T. Tuoto. 2018. "Population Size Estimation and Linkage Errors: the Multiple Lists Case." *Journal of Official Statistics.* 34 (4): 889–908. DOI: https://doi.org/10.2478/jos-2018-0044.

Ding, Y., and S.E. Fienberg. 1994. "Dual system estimation of Census undercount in the presence of matching error." *Survey Methodology* 20: 149–158. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1994002/article/14422-eng.pdf.

Fellegi, I.P., and A.B. Sunter. 1969. "A Theory for Record Linkage." *Journal of the American Statistical Association* 64: 1183–1210. DOI: https://doi.org/10.1080/01621459.1969.10501049.

Fienberg, S.E. 1972. "The multiple recapture census for closed populations and incomplete contingency tables." *Biometrika,* 59(3): 591–603. DOI: https://doi.org/10.1093/biomet/59.3.591.

Gerritse, S.C., B.F.M. Bakker, and P.G.M. van der Heijden. 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 (accessed 2016).

IWGDMF (International Working Group for Disease Monitoring and Forecasting). 1995. "Capture-recapture and multiple-record systems estimation I: history and theoretical

development." *American Journal of Epidemiology;* 142: 1047–1058. DOI: https://doi.org/10.1093/oxfordjournals.aje.a117558.

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. DOI: https://doi.org/10.1080/01621459.1989.10478785.

McLeod, P., D. Heasman, and I. Forbes. 2011. "Simulated data for the on the job training." *Essnet DI*. Available at: *http://www.cros-portal.eu/content/job-training* (accessed 2017).

Lincoln, F.C. 1930. *Calculating Waterfowl Abundance on the Basis of Banding Returns*, U.S. Dept. Agric., 118: 1–4. Available at: https://openlibrary.org/books/OL14861353M/Calculating_waterfowl_abundance_on_the_basis_of_banding_returns.

Menkens, G.E., and S.H. Anderson Jr. 1988. "Estimation of Small-Mammal Population Size." *Ecology* 69 (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. DOI: https://doi.org/10.2307/1941172.

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. DOI: https://courses.cs.washington.edu/courses/cse590q/04au/papers/WinklerEM.pdf.

Wolter, K.M. 1986. "Some coverage error models for census data." *Journal of the American Statistical Association* 81: 338–346. DOI: https://doi.org/10.1080/01621459.1986.10478277.

Received June 2019
Revised July 2020
Accepted November 2020