**Weighted capture - recapture model to correct for linkage errors**

Daan Zult[[1]](#footnote-1), Peter-Paul de Wolf[[2]](#footnote-2), Bart Bakker[[3]](#footnote-3), Statistics Netherlands and Peter van der Heijden[[4]](#footnote-4), University of Utrecht, db.zult@cbs.nl

**Abstract**

*The size of a partly observed population is often estimated with the capture-recapture (CRC) model. An important requirement for the CRC model to hold is that records over different samples can be identified such that the observer knows whether they are the same unit or not (i.e. no linkage errors). This is particularly relevant when identification is not obtained by some direct identifier (like a tag or id-code) but by indirect identifiers (e.g. name, address or animal’s skin patterns). Linkage based on indirect identifiers is often probabilistic and may lead to records being falsely linked or identical records being falsely missed as a link, which gives biased CRC population size estimates. A partial solution to this problem was provided by Ding and Fienberg (1994) and was later extended by Di Consiglio and Tuoto (2015). These authors show how to use linkage probabilities to correct the CRC estimator. De Wolf et al. (2018) further extend this model by allowing for linkage error correction with different sized samples. However, all these correction methods are designed for only two samples and do not include the use of covariates (e.g. sex or age), which implies that correcting for linkage errors cannot be combined with other correction methods that are often required to correct for other sources of bias. In this paper we therefore reformulate the general model from De Wolf et al. into our Weighted-CRC model. This new extended model can deal with multiple samples and covariates and we show how it works with simulation data.*

**Keywords:** population size, capture-recapture, linkage errors

**1. Introduction**

Capture-recapture (CRC) models provide a standard approach to estimate the size of unobserved parts of the population (Petersen, 1896). The general idea behind CRC models is that the number of unobserved individuals can be estimated from observational patterns over multiple samples. This estimate can be written as a prediction from a log-linear Poisson (or Gamma) regression model (e.g. see Cormack, 1989) for count data. One of the important conditions for the CRC model to work well is that the same units can be accurately recognised over samples. If not, individuals will be falsely (not) linked between samples and the resulting CRC estimates will be biased (Wolter, 1986, Gerritse et al., 2017). In this case the population size estimate (PSE) suffers from so called linkage error bias. Linkage errors are especially prone to occur in case perfect identifyers are not available. In that case individuals are often linked by means of indirect identifyers, such as name, address, animal’s skin patterns or a combination of different identifyers. Linkage models that use such indirect identifyers to link different samples are referred to as probabilistic linkage models (e.g. see Fellegi and Sunter (1969)). Probabilistically linked samples in general contain linkage errors, especially in the case of large samples. For instance, two different individuals might have the same name and are therefore falsely linked.

In order to correct for this linkage error bias, Ding and Fienberg (1994) propose a linkage error correction method (D&F method) that uses a small subsample that can both be perfectly and probabilistically linked. They show how the comparison between both the perfect and probabilistically linked samples can be used to correct the PSE. However, the D&F method has a drawback, because in their formulation it can only be applied in case the CRC model is based on only two samples and no covariates. This is a problem, because two samples is too limited when samples are not independent, which is often the case. Furthermore, the D&F method does not explicitly consider covariates, which are required when capture probabilities differ between groups. This is especially relevant with respect to model selection, because although the D&F method can be independently applied on subgroups, there is no clear method to test whether adding a covariate is a significant improvement of the model. Therefore, in this paper we extend the D&F linkage error correction method to a more general linkage error correction method that can deal with multiple samples and covariates. We refer to this model as the weighted capture-recapture (WCRC) model.

The WCRC model is derived from the work of Di Consiglio and Tuoto (2015) (DC&T) and De Wolf et al. (2018) (DW). DC&T show that the D&F method only allows one-way linkage errors (i.e. an individual from sample 1 that is incorrectly mislinked to an individual in sample 2) and they extend the model such that two-way errors are allowed. DW show that D&F and DC&T implicitly assume that both samples have the same size. Therefore DW extend the model further by allowing for different sample sizes. However, the most important contribution of this work is that DW show that the D&F method can be written much more comprehensively by:

$\hat{N}\_{D\&F}=\frac{n\_{1+}n\_{+1}}{E\left[n\_{11}|S,S^{\*}\right]}$ (1),

where $\hat{N}\_{D\&F}$ is the D&F method population size estimate (PSE),$ S=\left[S\_{1},S\_{2}\right]$ is a set of two population samples that are probabilistically linked, $ S^{\*}=\left[S\_{1}^{\*},S\_{2}^{\*}\right]$ is a set of two representative subsamples from $S\_{1}$ and $S\_{2}$ (i.e. $S^{\*}\in S$) that are perfectly linked, $n\_{1+}$ and $n\_{+1}$ are the sample sizes of $S\_{1}$ and $S\_{2}$ and $n\_{11}$ is the number of links between $S\_{1}$ and $S\_{2}$. When there are no linkage errors $n\_{11}=E\left[n\_{11}|S,S^{\*}\right]$, but in case there are linkage errors this equality does not hold. Equation (1) is not only a useful simplification of the D&F method, it also makes the method much easier to understand because it shows that all the D&F method does is replacing the number of links in $S$ by its expected number of links in $S$, which can be estimated with $S^{\*}$. We will use this last insight to extend the D&F method into a method that can handle multiple samples and covariates. We refer to this extended model as the weighted capture-recapture (WCRC) model for reasons that will become clear in Section 2.

The outline of this paper is as follows. In Section 2 we derive the WCRC model, in section 3 we investigate its properties in a simulation study and in section 4 we discuss the results.

**2. Derivation of the Weighted Capture-Recapture model**

A CRC model can also be written as a log-linear Poisson regression model in which observation counts are disaggregated to different samples and covariates. The input for the standard most basic CRC model is presented in Table 1. When Table 1 is used in a log-linear Poisson regression model where the Observed frequencies are ‘explained’ by $S\_{1}$ and $S\_{2}$ and this model is used to predict the frequency for $S\_{1}=0$ and $S\_{2}=0$ (i.e. the unobserved part of the population) we get the same PSE as with the standard Petersen estimate (i.e. $\hat{N}\_{Petersen}=\frac{n\_{1+}n\_{+1}}{n\_{11}}$). However, in order to mimic Equation (1) by a log-linear Poisson regression model, the last column of Linkage error corrected (LEC) frequencies are required.

**Table 1: Frequency table of two samples**

|  |  |  |  |
| --- | --- | --- | --- |
| $$S\_{1}$$ | $$S\_{2}$$ | **Observed frequencies** | **LEC frequencies** |
| 1 | 1 | $$n\_{11}$$ | $$E\left[n\_{11}|S,S^{\*}\right]$$ |
| 1 | 0 | $$n\_{10}$$ | $$n\_{1+}-E\left[n\_{11}|S,S^{\*}\right]$$ |
| 0 | 1 | $$n\_{01}$$ | $$n\_{+1}-E\left[n\_{11}|S,S^{\*}\right]$$ |

The last column holds because $n\_{1+}=n\_{11}+n\_{10}$ and $n\_{+1}=n\_{11}+n\_{01}$. In this paper we will not discuss how to obtain $E\left[n\_{11}|S,S^{\*}\right]$ extensively. A simple approach we will also use in the simulation study in the next section is to simply multiply $n\_{11}$ by $\frac{n\_{11}^{\*}}{n\_{11}\in S^{\*}}$, where $n\_{11}^{\*}$ is the number of links in $S^{\*}$ and $n\_{11}\in S^{\*}$ is the number of links in $S$ between the records that are also in $S^{\*}$.

*2.1. Adding covariates*

Table 1 can not only be used to obtain the D&F estimate from a log-linear Poisson regression model from Equation (1), it can also be easily extended with covariates. For instance, let’s say we think sex might affect capture probabilities, so we would like to add sex to Table 1, we then get:

**Table 2: Frequency table of two samples and a binary covariate**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| $$S\_{1}$$ | $$S\_{2}$$ | **Sex** | **Observed frequencies** | **LEC ‘frequencies’** |
| 1 | 1 | 1 | $$n\_{111}$$ | $$E\left[n\_{111}|S,S^{\*}\right]$$ |
| 1 | 0 | 1 | $$n\_{101}$$ | $$n\_{1+1}-E\left[n\_{111}|S,S^{\*}\right]$$ |
| 0 | 1 | 1 | $$n\_{011}$$ | $$n\_{+11}-E\left[n\_{111}|S,S^{\*}\right]$$ |
| 1 | 1 | 0 | $$n\_{110}$$ | $$E\left[n\_{110}|S,S^{\*}\right]$$ |
| 1 | 0 | 0 | $$n\_{100}$$ | $$n\_{1+0}-E\left[n\_{110}|S,S^{\*}\right]$$ |
| 0 | 1 | 0 | $$n\_{010}$$ | $$n\_{+10}-E\left[n\_{110}|S,S^{\*}\right]$$ |

Here we should note that $n\_{1+1}$, $n\_{+11}$, $n\_{1+0}$ and $n\_{+10}$ are all known, because they are simply the number of man and woman in sample 1 and 2. Furthermore, when we have $E\left[n\_{11}|S,S^{\*}\right] $and the sample $S^{\*} $contains gender as covariate as a representative sample should, we can easily obtain $E\left[n\_{111}|S,S^{\*}\right]$ and $E\left[n\_{110}|S,S^{\*}\right]$, because $E\left[n\_{11}|S,S^{\*}\right]=E\left[n\_{111}|S,S^{\*}\right]+E\left[n\_{110}|S,S^{\*}\right]$. Of course, this expansion with gender is not limited to one binary variable, any number of categorical covariates may be used to extend the ‘frequency’ table, as long as they are both available in $S$ and $S^{\*}$. We say ‘frequency’ table, because the LEC frequencies are no longer traditional counts and $E\left[n\_{11}|S,S^{\*}\right]$ is not even necessarily an integer. Therefore, instead of a log-linear Poisson regression model one should apply a log-linear Gamma regression model, which is grossly equivalent to the Poisson version but can handle a continuous dependent variable.

*2.2. Adding additional samples*

Now that we have reformulated the D&F method into a log-linear regression framework with covariates, the next step is to include additional samples. However, adding additional samples is less straightforward than adding covariates. In fact, in order to add additional samples we need to introduce the concept of individual weights $w=c(w\_{1},…,w\_{i},…,w\_{n\_{++}})$ where $n\_{++}$ is the number of records in the probabilistically linked samples $S=\left[S\_{1},S\_{2}\right]$. In the standard CRC setting one could say that each individual has a weight of 1. Adding up these weights of 1 per category gives us the Observed frequencies from Table 2. Now the question is what should $w$ be in order to obtain the LEC ‘frequencies’. The answer is simple, devide the LEC ‘frequencies’ by the Observed frequencies. For instance, the individual weight $w\_{i}$ of individual $i$ with sex = 1 and linked between $S\_{1}$ and $S\_{2}$ becomes $\frac{E\left[n\_{111}|S,S^{\*}\right]}{n\_{111}}$. Then, summing up $w$ per category gives the LEC ‘frequencies’ required to obtain the LEC PSE. In fact, when we ignore covariates we can rewrite Equation (1) as:

$\hat{N}\_{D\&F}=\frac{n\_{1+}n\_{+1}}{E\left[n\_{11}|S,S^{\*}\right]}=\frac{w\_{1+}w\_{+1}}{w\_{11}}$ (2).

Note that $w\_{1+}=n\_{1+}$ and $w\_{+1}=n\_{+1}$.

This application of $w$ is in itself useful, but even more useful is that in this way the linkage errors between $S\_{1}$ and $S\_{2}$ and its correction are ‘stored’ for later use. In fact, one can use $w$ to add additional samples, simply by repeating the linking procedure above where the linked sample from step 1 can be considered $S\_{1}$ while the new sample $S\_{3}$ can be considered $S\_{2}$. In other words, one should replace $S$ by $S=\left[\left[S\_{1}|S\_{2}\right],S\_{3}\right]$ and $S^{\*}=\left[\left[S\_{1}^{\*}| S\_{2}^{\*}\right],S\_{3}^{\*}\right]$ and instead of $n$ use the resulting $w$ from the previous step. Ignoring covariates, the ‘frequency’ table than looks like:

**Table 3: Frequency table of two samples**

|  |  |  |  |
| --- | --- | --- | --- |
| $$S\_{1}|S\_{2}$$ | $$S\_{3}$$ | **Sum of weights step 1** | **Sum of weights step 2** |
| 1 | 1 | $$w\_{11}$$ | $$E\left[w\_{11}|S,S^{\*}\right]$$ |
| 1 | 0 | $$w\_{10}$$ | $$w\_{1+}-E\left[w\_{11}|S,S^{\*}\right]$$ |
| 0 | 1 | $$w\_{01}$$ | $$w\_{+1}-E\left[w\_{11}|S,S^{\*}\right]$$ |

In Table 3 covariates can be added in the same way as was done in Table 2. The same method holds for re-splitting $S\_{1}$ and $S\_{2}$, which were considered one sample in the second linking procedure. However, in the log-linear model $S\_{1}$ and $S\_{2}$ should be separate explanatory variables again, which can simply be done by distributing the sums of weights after step 2 between $S\_{1}$ and $S\_{2}$ by their proportional sums of weights over these samples. As a result of these 2 linkage error correction steps, the weight of an individual with, for instance, sex = 0 and linked between $S\_{1}$, $S\_{2}$ and $S\_{3}$ then becomes:

 $ w\_{i}= \frac{E\left[n\_{110}|S\_{1},S\_{2},S\_{1}^{\*},S\_{2}^{\*},X\right]}{n\_{110}}\*\frac{E\left[w\_{110}|S\_{1}|S\_{2},S\_{3},S\_{1}^{\*}|S\_{2}^{\*},S\_{3}^{\*},X\right]}{w\_{110}}$ . These steps in which $w$ is updated after each new sample is linked can be repeated for any number of samples. The frequency table with these sums of weights as dependent variable can be used to estimate the log-linear Gamma model that provides a LEC PSE based on more than two samples and covariates. This model we refer to as the WCRC model.

**3. Simulation study**

In this section the WCRC model presented in section 2 is tested in a simulation study. This is useful because it allows us to check whether the derivations in the previous sections hold in practice, especially when they are applied in an environment in which multiple underlying model assumptions are violated simultaneously. In particular, how does the WCRC model behave when it is applied in a setting in which linkage errors occur simultaneously with capture probabilities that are not independent of covariates and other samples? In section 3.1 we describe the setup of this simulation study and in section 3.2 we discuss the results.

*3.1. Simulation setup*

The data used in the simulation study is a publicly available fictitious population of 26.625 persons on the UK population census[[5]](#footnote-5). In this simulation study only the file ‘person\_list.csv’ is used. This person\_list contains both a direct identifier and indirect identifiers that can be used in both probabilistically and perfectly link samples. This simulation study uses a set of three linkage keys[[6]](#footnote-6) and in order to have a certain degree of linkage errors, in each linkage key, in 3% of the records, the value is replaced by a random value from the population (i.e. a random name, day or month). Furthermore, the list contains several covariates, of which ‘SEX’ is used as covariate. For each replication three random samples are generated from the population, where individual capture probabilities are allowed to differ, depending on a person’s covariate SEX and whether he/she is in other samples or not. This gives us four scenarios that differ with respect to the covariate and sample dependence, as can be seen in Table 4.

**Table 4. Simulation study scenarios**

|  |  |  |  |
| --- | --- | --- | --- |
| **Scenario** | **Linkage errors** | **Covariate dependence** | **Sample dependence** |
| 1 | Yes | No | No |
| 2 | Yes | Yes | No |
| 3 | Yes | No | Yes |
| 4 | Yes | Yes | Yes |

Here covariate dependence implies that in each sample the covariates sex affects the probability to be in that sample. Sample dependence implies that in each sample the two-sample interactions $S\_{1}S\_{2}$, $S\_{1}S\_{3}$ and $S\_{2}S\_{3}$ affect the probability to be in that sample, but the three-sample interaction $S\_{1}S\_{2}S\_{3}$ does not. In each replication a random population of 10.000 is selected from the list that serves as our ‘true population’. From this true population three samples are generated of approximately 8000, 5000 and 2000 records. Under the first scenario these samples are truly random while under scenario 2 to 4 the capture probabilities depend on covariates, other samples or both. Finally, the sample $S^{\*}$ is generated that contains approximately 30% of records that are in any of the three samples. This part of the population can be both probabilistically and perfectly linked. For each scenario we replicate the procedure 600 times.

In order to compare the WCRC estimates with the alternatives, we consider six PSEs that are in Table 5 that can all be obtained from the simulation ingredients described above.

**Table 5: Six different PSEs compared in the simulation study**

|  |  |
| --- | --- |
| **PSE** | **Description** |
| Weighted PSE3 | Weighted CRC PSE based on 3 samples linked with linkage errors |
| Perfect PSE3 | Standard CRC PSE based on 3 samples linked without linkage errors |
| Weighted PSE2 | Weighted CRC PSE based on 2 samples linked with linkage errors |
| Perfect PSE2 | Standard CRC PSE based on 2 samples linked without linkage errors |
| Naïve PSE3 | Standard CRC PSE based on 3 samples linked with linkage errors |
| Naïve PSE2 | Standard CRC PSE based on 2 samples linked with linkage errors |

The estimators in Table 5 differ in three aspects. First, they differ in the number of samples that are involved in the estimation. Second, they differ in their linkage process (i.e. the PSEs that start with ‘Perfect’ are based on perfect linkage while the others are not). Third, they differ with respect to linkage error correction (i.e. the PSE that starts with ‘Weighted’ are corrected for linkage errors and with ‘Naïve’ are not).

The perfect PSEs serve as benchmarks that cannot be obtained in practice when a perfect identifier is not available. The Naïve PSEs serve as benchmarks that can be obtained in practice and show what happens if linkage errors are not taken into account. In each replication al six estimates are calculated for the same true population and the same two (i.e. the two that are largest) or three samples.

*3.2. Simulation results*

In figure 1 below the simulation results of the four scenarios are presented as density plots.

**PSE density plots**



In each of these four figures we see the densities of the six PSEs from Table 5. Ideally the density revolves around the true population size of 10.000. This is clearly not the case for all PSEs. In the first figure we see the results for scenario 1, in which there is no covariate or sample dependence. Here we see that due to linkage errors (remember, only 3% in each of the three linkage keys), both Naïve PSEs are clearly biased. The same holds for scenario 3, under which there is only covariate dependence. In that case also the WCRC model with 2 samples provides PSEs that revolve around 10.000, which implies that the extension of the D&F method with covariates performes well. In contrast, the same does not hold for scenario 2 and 4 in which there is sample dependence, logically leading to unbiased PSEs based on 2 samples. The only two PSEs that hold under all four scenarios are the weighted and perfect CRC PSE, which makes the WCRC model a viable alternative when perfect linkage over the full sample is not an option.

**4. Discussion**

In this paper we derived and tested the WCRC model, a model that is equivalent to the D&F method for linkage errors but is more general in the sense that it incorporates covariates and multiple samples, which might be necessary to correct for other sources of bias. Our simulation study indicates that in a controlled setting the method works well, which is good news. And although in practise the WCRC model might not yet solve all the linkage error problems, some justified criticism that concerns the D&F method still holds for the WCRC method. The first criticism that comes to mind in applying the WCRC model successfully in practise is how to obtain a good $S^{\*}$. Ideally $S^{\*}$ is a truly random sample of $S$ or at least it should be representative for $S$, both with respect to covariates and the quality of linkage keys. However, when $S^{\*}$ is based on a rematch study, as with D&F, individuals with good quality linkage keys might be more easy to trace and accurately rematch due to the quality of their linkage keys or they might be more responsive in a rematch study. When this is the case, the proportion of linkage errors in $S^{\*}$ may be different than the real proportion in $S$. Such a difference would lead to a biased value of $E\left[n\_{11}|S,S^{\*}\right]$ and therefore a biased D&F and WCRC PSE. Another issue is the size of $S^{\*}$, especially when the population contains small specific groups with low capture probabilities, it won’t be easy to have them well represented in $S^{\*}$. How large the impact of such issues is, requires further research. The good news is now that the linkage error correction method is incorporated in the more general family of log-linear regression models, it can also be more easily studied within other studies in this field.

**5. References**

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. 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.

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.

Gerritse, S.C., Bakker, B.F.M., Van der Heijden, P.G.M. (2017). The impact of linkage errors and erroneous captures on the population size estimator duet o implied coverage. Discussion paper 2017-16, Statistics Netherlands, The Hague/Heerlen

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>.

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.

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

1. Corresponding author [↑](#footnote-ref-1)
2. Cooperating author [↑](#footnote-ref-2)
3. Research supervisor and cooperating author [↑](#footnote-ref-3)
4. Research supervisor [↑](#footnote-ref-4)
5. Created for the European Statistical System Data Integration project (ESSnet DI) (McLeod, Heasman and Forbes, 2011), a European project on data integration (Record Linkage, Statistical Matching, Micro integration Processing) run from 2009 to 2011. [↑](#footnote-ref-5)
6. ‘PERNAME2’, ‘DOB\_DAY’ and ‘DOB\_MON’ served as linkage variables, which corresponds to the ‘bronze scenario’ in DC&T. [↑](#footnote-ref-6)