Correction for Linkage Error in Population Size Estimation FIRST DRAFT

B.F.M. Bakker[12], L. Di Consiglio[3], D.J. van der Laan[1], T. Tuoto[4], P.-P. de Wolf[1], D. Zult[1]

[1] *Statistics Netherlands, The Hague, Netherlands; bfm.bakker@cbs.nl*
[2] *VU University, Amsterdam, Netherlands;*
[3]*Eurostat, Luxembourg;*
[4] *Istat, Rome, Italy;*

Abstract
In capture-recapture models perfect linkage between the used sources is assumed. Violation of this assumption leads to biased estimates and this happens most of the time. In this paper, two methods are presented. The first one is to distinguish different scenarios for different levels of linkage error. The second one is developed by Ding and Fienberg (1994) and extended by Di Consiglio and Tuoto (2015), based on the idea that if the sources are linked probabilistically, the capture – recapture outcomes should be corrected by the probabilities of a pair being a correct match, as well as by the probability of missing a true match. It is applied to data from the Netherlands and the second method seems to lead to implausible results when accurate evaluation of linkage errors are not available

Keywords: Capture – recapture, linkage quality, probability linked data

1. Introduction

An important aspect of the quality of data sources is the coverage of the target population. There could be over coverage, i.e. elements that do not belong to the target population are present in the source, and under coverage, i.e. elements are in the target population that are not represented in the source (Bakker and Daas, 2012). If two or more lists are available, it is possible to estimate the under coverage of each of these list by applying capture-recapture models (CRC), also called multi-capture when more than two lists are used (Fienberg, 1972; Bishop et al., 1975; Van der Heijden *et al.* 2012; Baffour, Brown & Smith, 2013). To get accurate outcomes from these models, several assumptions have to be met. The main assumptions are independence of inclusion probabilities, closed population, no erroneous captures and perfect record linkage. In practice these assumptions cannot always be met, particularly when data are originally not collected for statistical purposes.

Gerritse et al. (2016) have estimated the usual resident population not registered in the Population Register in the Netherlands at ultimo September 2010 aged 15-65 between 88 and 185 thousand persons. The sensitivity of the results to violation of the independence, perfect linkage and no erroneous captures assumptions is quite large, in particular if the overlap between one and the other sources is relatively small. For the independence assumption and the assumption that there are no erroneous captures they found reasonable solutions, but for the linkage error they did not (Gerritse et al. 2015a, 2015b, 2016). To assess the impact of linkage error on the estimate, they considered eight different scenarios. In these scenarios they assumed several combinations of proportions of missed links, that should have been linked in a certain way. This paper has two objectives. The first one is to replicate the estimation of the population size of the usual residents in the Netherlands for ultimo March 2014 and calculate the outcomes for the same eight scenarios. The second one is to correct the population size estimate for linkage error. We propose to apply the method described by Di Consiglio and Tuoto (2015) which in fact is an extension of the work of Ding and Fienberg (1994).

2. Method

Often, data sources are linked probabilistically. The linkage is based on the idea that for two files *I* and *J*, all possible pairs of records in these files can be divided into two disjoint sets *M* (Matched) and *U*(Unmatched). A pair of records (*i,j*) is a member of *M* if the two records are truly related to the same element. Otherwise it is a member of *U*. In reality, the members of *M* and *U* are unknown. The record linkage process aims to classify each record pair as belonging to either *M* or *U*, by observing whether the actual values on corresponding linkage variables within each pair agree (Fellegi and Sunter, 1969). In probabilistic linkage, weights are determined for each possible pair of records on the basis of the estimation of belonging to M or U. The parameters of this linkage process can be used to correct the CRC-estimation.

Let N be the population unknown total, N_1 and N_2 be the population size counts in the first and second list, respectively. Let x_{11} be the number of units recorded in both lists, $x_{12} = N_1 - x_{11}$ the number of units reported only in List 1 and $x_{21} = N_2 - x_{11}$ the number of units reported

only in List 2. The counts can be organized in a 2 x 2 contingency table, with x_{22} , the unknown number of units missed by both lists.

Table 1. *Contingency table of the counts in the two lists*

		List 2	
		Present	*Absent*
List 1	*Present*	x_{11}	x_{12}
	Absent	x_{21}	x_{22}

As the linkage process is subject to errors, we might have differences in the realized links, let $x_{11}^*, x_{12}^*, x_{21}^*$ be the corresponding observed quantities resulting from the linkage procedure. Finally, let $N_{1\cup 2}$ be the observed number of records in list 1 or list 2.

Let α be the probability that a true match is linked (also known as "precision") and therefore 1-α the probability that a true match is not linked. Let β be the probability that a record without a match in the other source is linked. Under the assumption that false links in records of pairs belonging to *M* are negligible as that occur when at least two errors are made (that is, records are incorrectly linked and the correct link is missed), the adjusted estimator is given by Di Consiglio and Tuoto (2015):

$$\tilde{N}_{MDF} = \frac{N_{1\cup 2}}{\hat{\tau}_{1,MDF} + \hat{\tau}_{2,MDF} - (\alpha\,\hat{\tau}_{1,MDF}\,\hat{\tau}_{2,MDF} + \beta\,(\hat{\tau}_{1,MDF} + \hat{\tau}_{2,MDF} - 2\hat{\tau}_{1,MDF}\,\hat{\tau}_{2,MDF}))} \tag{1}$$

where

$$\hat{\tau}_{1,MDF} = \frac{2\beta x_{11}^* + \beta x_{12}^* + \beta x_{21}^* - x_{11}^*}{(2\beta - \alpha)(x_{11}^* + x_{21}^*)} \tag{2}$$

$$\hat{\tau}_{2,MDF} = \frac{2\beta x^*_{11} + \beta x^*_{12} + \beta x^*_{21} - x^*_{11}}{(2\beta - \alpha)(x^*_{11} + x^*_{12})}. \quad (3)$$

are the ML estimators of the probabilities of being recorded in lists 1 and 2, respectively.

In the previous formulas, the probabilities of the cell counts in the contingency table involve also the linkage probabilities. If the linkage procedure is error free, i.e. $\alpha = 1$ and $\beta = 0$, the adjusted estimators turn into the ones of the classical CRC approach. The proposed estimators are based on the assumption that linkage errors are constant. If this assumption holds at least in sub-groups the estimators can be applied within strata in which linkage error probabilities (and capture probabilities) can be assumed more homogeneous than in the whole population.

Compared to Di Consiglio and Tuoto (2015) we have to make two extensions. Because we had to link very large datasets, we applied probabilistic linkage with several blocking variables. Because of the use of different blocking variables, the α - and β -parameters from the two different blockings have different meanings and cannot be used to estimate parameters. We have not found a solution to this problem yet, but we might look for subgroups in which these estimated probabilities are more homogeneous, taking the blocking variables into account. The second extension is to add a third source. For the moment we assume that two sources are perfectly linked, and handle the combined file as if it is one source. The third source is linked to this combined file.

3. Data and data linkage

3.1. Data

For the application of the method, three administrative data sources are used. This relaxes the assumption of independency between the registers[1]. The first register is the Dutch Population Register (PR). It includes the total registered population of the Netherlands and is the population of the register based census (Schulte Nordholt et al. 2014). The second one contains the employees in registered Dutch companies and is called the Employment Register (ER). The third one contains the crime suspects that are registered as such by the police

[1] Because we assume that two of the three registers have been linked perfectly, this is not entirely true.

(CSR). For the PR and ER the assumption that the population is closed is easily met because ultimo March 2014 is chosen as the reference date. This cannot be applied to the CSR because this register is event-based: crime suspects for which the police makes a report are registered. The number of events on one specific day is not enough to apply the capture – capture method. In order to satisfy the assumption as well as possible, we restrict the period of the CSR to the first half year of 2014. Not all elements of the population have a positive probability of being registered. The ER is restricted to the population of 15-65 years of age, while the CSR is restricted to the persons of 12 years and older. Because of these restrictions, we are not able to estimate the total population, but only the population of 15-65 years of age.

To prevent erroneous captures as much as possible, we removed records from the ER and CSR of persons who do not belong to the population: (a) the few persons with the Dutch nationality not registered in the PR because we expect them to be expats working in another country and therefore not belonging to the usual resident population; (b) persons with an address in Belgium or Germany, the neighbouring countries of The Netherlands, because it is likely that they live in Belgium or Germany and are only temporary in The Netherlands to work, to go to school, to shop or to have a short holiday; (c) persons who are reported for a crime by the border police at the airport or elsewhere because they did not enter the country at all.

Nationality group has 7 categories: (1) EU15 (excl. Netherlands) (2) Polish (3) Other EU (4) Other western (5) Turkish, Moroccan, Antillean, Surinam (6) Iraqi, Iranian, Afghan, asylum seeker countries Africa (7) Other Balkan, former Soviet Union, other Asian, Latin American, and not mentioned elsewhere. The countries are clustered according to main migration motives, migration legislation, regulations of the PR and size. Nationality is derived in the following order (the first available one is selected): nationality from the PR, nationality from the ER, nationality from the CSR, and country of birth from the PR. For age, we use four categories: (1) 15-24 (2) 25-34 (3) 35-49 and (4) 50-64 years of age. Sex has the categories (1) male and (2) female. Usual residence has the categories (0) shorter than 1 year and (1) longer than 1 year. None of the registers contain information on usual residence. For the PR is assumed that all registered persons are usual residents. For the ER, those who are not

registered in the PR are defined as a usual resident if they are employed longer than a year without a gap of more than a month. Usual residence in the CSR is imputed with Predictive Mean Matching, using a log-linear regression model (Gerritse et al., 2015b).

3.2.Data linkage.

The three registers were linked in two steps each consisting of a combination of deterministic and probabilistic linkage. The selections on nationality group, country of residence, age, and residence duration were applied after record linkage as these variables could only be determined accurately after linkage by combining information from all three registers. Therefore, linkage was performed on the complete registers.

First, the PR was linked to the ER. To allow for changes in addresses records were selected from 2013-06-01 to 2014-12-31. Most links (99.99%) between the registers were established using deterministic linkage using the Civil Service Number (CSN). The remaining links were established using probabilistic linkage on postcode, house number, date of birth and sex. We assume that the PR and ER are perfectly linked. The linked PR and ER were combined into one register (P-ER) which was linked to the CSR.

The CSR consists of two parts. First, records that were linked to the PR by the police. For these records only the CSN was available. Second, records for which the CSN was not available, date of birth, sex, postcode (for Dutch addresses), street, house number, city, and country were available for probabilistic linkage. Three separate blocking variables were used: postcode, date of birth, and a combination of city and date of birth. For the last blocking variable city and date of birth were combined as follows: city, year and month of birth were used for large cities; city and year of birth for medium sized cities; and city for small cities. After each of the probabilistic linkages 1-to-n linkage was enforced (one person from the P-ER can be linked to multiple records from the CSR; one record from the CSR can be linked to only one record from the P-ER). The pairs from the three probabilistic record linkage steps were then combined. When there are duplicate pairs, the pair with the highest posterior matching probability was selected. After linking the three files, the selections mentioned in section 3.1. were applied.

Table 1: Overview of linkage results

identified in			Missing linkage keys in CSR			
PR	ER	CSR	DOB	Address 1[a]	Address 2[b]	N
			%			*abs.*
No	No	Yes	0	53	79	6672
No	Yes	No		-	-	30118
No	Yes	Yes	0	0	3	36
Yes	No	No		-	-	415646
Yes	No	Yes	0	1	2	7552
Yes	Yes	No		-	-	323399
Yes	Yes	Yes	0	1	2	2327

[a] *Missing value in either city, house number and street.*

[b] *Missing value in either postcode and house number.*

[c] *CSR records for which it is not certain that crime was committed in first half of 2014.*

Table 1 shows the results of the linkage process. In the records that were not linked to the other two registers, 50- 80% has missing values in the address variables. This makes linkage of these records unlikely.

4. Results

Gerritse et al. (2016) presented a series of estimates of non-registered usual residents aged 25-65 in the Netherlands ultimo September 2010. They had several estimates, because it was unclear how one should deal with captures in the CSR that have an incomplete linkage key, as they are potentially new captures, recaptures (i.e. missed link) or erroneous captures. Each estimate was therefore the result of a different scenario in which a share of the unlinked captures in the CSR with incomplete linkage key is redefined as erroneous and missed links. Furthermore, they considered a small share of unlinked captures in the CSR with complete linkage keys as potential erroneous captures, because the linkage variables could contain errors. Table 2 describes the same scenarios as in Gerritse et al. (2016) and complement this with outcomes for ultimo March 2014 obtained by an identical estimation method. An important difference between the March 2014 and September 2010 data is the share of unlinked captures in the CSR with incomplete linkage keys, which was 37% in September

2010 but has increased to 56% in March 2014. The baseline estimation is described because it shows the results if no correction for linkage error and erroneous captures are made.

Table 2: Overview of the scenarios and the maximum likelihood estimates of the missed portion of the population

	incomplete linkage key		complete linkage key		estimation	
	erroneous captures	linkage error	erroneous captures	linkage error	September 2010	March 2014
scenario	*%*				*x1000*	
Baseline	0	0	0	0	249	470
1	75	25	0	0	66	146
2	75	25	5	0	66	146
3	75	25	0	5	54	134
4	75	25	5	5	56	134
5	100	0	0	0	151	143
6	100	0	5	0	151	143
7	100	0	0	5	91	132
8	100	0	5	5	92	133

Although the baseline results differ substantially, the estimates for the different scenarios are closer. The substantial increase of the baseline estimate is due to the higher share of incomplete linkage keys in the CRS. However, this higher share also implies that in all the scenarios a higher fraction is redefined as erroneous capture or missed link, which in turn substantially suppresses the estimates, bringing them down to levels similar to that of September 2010. In that year the estimations varied from 87 to 185 thousand. Adding the 37 thousand registered persons in the ER and CSR who are not linked to the PR, the results vary from 169 to 183 thousand usual residents not registered in the PR in 2014.

In the next step we have applied the correction method for linkage error. However, the estimation of the α - and β -parameters was not straightforward. Because we used different blockings, the m- and u-parameters differ in their meaning and therefore, the α - and β - parameters deduced from the m- and u-parameters from the different blockings also differ in their meaning. Naïve application of formulas (1) – (3) to obtain the population size estimation,

simply ignoring the problem of multiple blockings, leads to implausible results that we do not report here. The effectiveness of the adjustment proposed by Di Consiglio and Tuoto (2015) strongly depends on the accuracy of the estimated linkage error probabilities. We have not found a solution for this problem yet.

5. Conclusions and discussion

This paper had two objectives. The first one was to replicate the estimation of the population size of the Netherlands for the year 2014. The estimated number of usual residents not registered in the Population Register ultimo September 2010 varied between 88 and 185 thousand persons. We estimated the population of usual residents ultimo March 2014 between 169 and 183 thousand persons. The difference between both groups of estimates is the degree of variation between them: for the 2014 the variation is much lower than for 2010.

The second aim of this paper was to correct the estimation for linkage error. However, because of multiple blocking of the probability linkage, we could not yet compute the necessary parameters and straightforwardly apply the adjustments.

References

Baffour, B., Brown, J.J., and Smith, P.W.F. (2013), An investigation of triple system estimators in censuses, Statistical Journal of the International Association for Official Statistics, 29, pp. 53-68.

Bakker, B.F.M. and Daas, P. (2012), Some Methodological Issues of Register Based Research, Statistica Neerlandica, 66, pp. 2-7.

Bishop, Y.M.M., Fienberg, S.E. and Holland P.W. (1975), Discrete multivariate analysis, MIT press, Cambridge, MA.

Di Consiglio, L. and Tuoto, T. (2015), Coverage Evaluation on Probabilistically Linked Data, Journal of Official Statistics, 31, pp. 415–429.

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

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

Fienberg, S.E. (1972), The multiple recapture census for closed populations and incomplete 2k contingency tables, Biometrika, 59, pp.409-439.

Gerritse, S.C. , Van der Heijden, P.G.M. and Bakker, B.F.M. (2015a), Sensitivity of Population Size Estimation for Violating Parametric Assumptions in Log-linear Models, Journal of Official Statistics, 31, pp. 357–379.

Gerritse, S.C. , Bakker, B.F.M. and Van der Heijden P.G.M. (2015b), Different methods to complete datasets used for capture-recapture estimation, Statistical Journal of the IAOS, 31, pp. 613-627.

Gerritse, S.C., Bakker, B.F.M., De Wolf, P.-P. and Van der Heijden, P.G.M., 2016, Under coverage of the population register in the Netherlands, 2010. Discussion paper 2016-02, Statistics Netherlands, The Hague / Heerlen.

Schulte Nordholt, E. et al. (2014), Dutch Census 2011. Analyses and methodology, Statistics Netherlands, The Hague / Heerlen.

Van der Heijden, P.G.M., Whittaker, J., Cruyff, M., Bakker, B., and Van der Vliet, H. (2012), People born in the Middle East but residing in the Netherlands: invariant population size estimates and the role of active and passive covariates, Annals of Applied Statistics, 6, pp. 831-852.