Discussion Paper

The impact of linkage errors and erroneous captures on the population size estimator due to implied coverage

The views expressed in this paper are those of the author and do not necessarily reflect the policies of Statistics Netherlands.

2017 | 16

Susanna C. Gerritse
Bart F. M. Bakker
Peter G. M. van der Heijden

Contents

Capture-recapture estimation is an often used methodology for hard-to-reach populations and relies heavily on a couple of assumptions. In this paper for two of these assumptions, sensitivity analyses are conducted: we investigate the effect of linkage error and erroneous captures on the population size estimate. These sensitivity analyses are conducted for the two and three register case, and for two different nationality groups. It was found that for the nationality group with a high implied coverage of one register over the other, the population size estimator was robust to linkage error and erroneous captures. For the nationality group with a low implied coverage, the population size estimator was not robust. It is concluded that researchers should consider the implied coverage of their registers for capture-recapture estimation, since this plays a crucial role in the robustness of the population size estimation. [1)]

1 Introduction

Capture-recapture methods are commonly used to estimate the size of hard-to-reach populations [9, 1, 6, 11]. When linking the individuals from two or more registers we can use these methods to estimate the part of the population that was missed by all registers. Capture-recapture estimation relies on five assumptions: 1) For the two register case, the registers are assumed to be independent in the sense that the inclusion probability of one register is independent of the inclusion probability of the other register. For three registers this assumption is relaxed and it is only assumed that the three factor interaction is zero, such that dependence between pairs of two registers may occur; 2) The registers are perfectly linked: when one unit is captured in two (or more) registers, perfect linkage assumes that we correctly identify all of these units as recaptures; 3) The population is closed: registers with continuous recording such as a population register are closed when one point in time is chosen. For incidence registers it is recommended to take a small sampling period to limit a possible violation of the closed population assumption; 4) All individuals in the registers belong to the population, i.e. there are no erroneous captures; 5) Assumptions related to homogeneity of inclusion probabilities [13]. Heterogeneity occurs when one register has heterogeneous inclusion probabilities, for example, when the probability to include men is higher than the probability to include women. If there is one source of heterogeneity, the estimate is unbiased when at least for one of the two registers the inclusion probabilities are homogeneous [4, 14]. If there is a source of heterogeneity in each of two registers, the estimates are unbiased if the inclusion probabilities of the two sources of heterogeneity are statistically independent [10, 12, p. 86].

In capture-recapture methodology it is not possible to verify from the data whether the assumptions are met, let alone to verify from the data to what extent possible violations occur. Some research has been conducted to investigate the effect of violations of the assumptions of capture-recapture analysis on the population size estimate for two registers. For the independence assumption, it has been shown that violating the independence assumption in the two register case can lead to biased results, but the bias is not necessarily large. This result is explained via the implied coverage of the register. Assume that we have two registers, register 1 and register 2. Implied coverage is the number of units included in the first register, that are also included in the second register, relative to the size of the second register. We propose that register 1 is always the register with the highest coverage of the population, such that the

[1)] This paper was reviewed by Jeroen Pannekoek

addition of register 2 leads to an increased coverage of the population, relative to the size of register 2. A high implied coverage results from a high number of units being included in the second register that are also included in the first register. When the data has a high implied coverage the capture-recapture estimation is relatively robust to a violation of independence between the registers [3, 2, 10].

Boden [2] used sensitivity analyses to test the effect of violating the assumptions of independence, perfect linkage and heterogeneity. The violation of independence between the sources had a substantial effect on the population size estimate. However, the results of violating perfect linkage and heterogeneity showed only a small bias. The small effect of linkage error on the capture-recapture estimation may be explained by the small number of linkage errors in the paper, although the author concludes that in his case even a moderate amount of linkage errors would have a minor impact on the population size estimate.

In this paper we investigate the effect of linkage error and erroneous captures on the population size estimation due to implied coverage, assuming all other assumptions have been met. We expect, similar to the impact of the violation of the independence assumption, that if the implied coverage is high the impact of violation of these assumptions is low and vice versa. For this purpose, we use administrative data from Statistics Netherlands: the Population Register (PR), an Employment Register (ER) and a Crime Suspects Register (CSR), all from 2010. The three registers used have been linked for the most part via deterministic linkage, made possible by the use of a unique Personal Identification Number (PIN) that all PR registered individuals in the Netherlands have. When there are no administrative errors, this PIN can identify every individual correctly. However, these registers may contain administrative errors such that deterministic linkage may not identify all links. To be able to identify these links we also used probabilistic linkage. Probabilistic linkage estimates for each possible pair of individuals in two registers their probability of a correct link [8]. Those individuals in the ER and CSR that did not already link to the PR deterministically, were linked probabilistically. The advantage of this method is that it allows for small errors in the identifying variables.

After linkage it was found that 37 percent of the individuals in the CSR that did not link to the PR and ER had missing information on one of the linkage variables, and could not be linked. It is possible that these individuals do not belong to the population but it is impossible to know whether they should have been linked or not. Thus there is the possibility of two assumptions being violated: the one on perfect linkage and the one on no erroneous captures. There is also a possibility that part of the data of the CSR that have missing data are duplicate units, thus one person that is registered twice but can not be identified as the same person due to missing information. We assume in this paper that duplicate units do not occur.

Since the number of captured and recaptured units are used to estimate the number of units missed by the registers for population size estimation, the process of record linkage is an important aspect of capture-recapture methodology when using administrative data. Errors in record linkage will result in violation of the perfect linkage assumption. There are two types of linkage errors. For simplicity sake we exemplify the possible errors in linkage for two registers only, register 1 and register 2, and two unique units, X and Y. One error in linkage occurs when unit X in register 1 is falsely linked to unit Y in register 2. This type of error is also known as a mislink or a false positive. A second error in linkage occurs when unit X in register 1 is falsely not linked with unit X in register 2. This type of error is known as a missed link or a false negative.

If there are no covariates involved, we are in a relatively simple situation where one false positive can be compensated by one false negative, and thus there will be no effect on the population size estimation. Thus linkage errors are the number of false positives minus the number of false negatives, or in other words the number of mislinks minus the number of missed links. Then linkage errors are seen as a balance of two possible errors. Throughout this paper we will most commonly refer to errors in record linkage as linkage error. Erroneous captures occur when units that do not belong to the population are in the data.

The three registers from Statistics Netherlands together contain data on individuals residing in the Netherlands. In this paper we study two nationality groups from the three registers. The first nationality group contains data from all three registers on one nationality only: Polish. In the EU individuals with EU nationalities are free to move and work within the EU. Hence, Polish individuals are free to live in the Netherlands and a high number of the labor migrants in the Netherlands in 2010 were from Poland. The second nationality group contains data from all three registers on individuals with an Iraqi, Iranian and Afghan nationality. Individuals from these three countries need a visum and working permit to enter the Netherlands and are undocumented immigrants when residing in the Netherlands without either of these. These nationality groups differ substantially in their implied coverage: the Polish have a low implied coverage, whereas the Iraqi, Iranian and Afghan have a high implied coverage.

We continue as follows. In section 2, a sensitivity analysis is conducted on these two nationality groups for the simplest form of capture-recapture: using only two registers. Section 2.1 will discuss the effect of linkage error on the population size estimate and section 2.2 will discuss the effect of erroneous captures on the population size estimate. In section 3 we extend the sensitivity analysis to the multiple register case, where three registers are used to exemplify one form of a multiple register case. In section 3.1 a sensitivity analysis will be conducted on the effect of linkage errors on the population size estimate. In section 3.2 the effect of erroneous captures on the population size estimate is established via sensitivity analysis. Section 4 will give a discussion of the results and we conclude this paper in section 5.

2 Capture-recapture for two registers

The simplest population size estimation model makes use of two registers, 1 and 2. Let variables A and B respectively denote inclusion in registers 1 and 2. Let the levels of A be indexed by i (i = 0 (No), 1 (Yes)) where $i = 0$ stands for "not included in register 1", and $i = 1$, stands for "included in register 1". Similarly, let the levels of B be indexed by j ($j = 0(No), 1(Yes)$). Expected values are denoted by m_{ij} and fitted values are denoted by $\hat{m}_{ij}$. Observed values are denoted by n_{ij}, with $n_{00} = 0$ by design.

One of the assumptions in population size estimation is that the probability of being in the first register is independent of the probability of being in the second register. Under independence the saturated loglinear model for the counts n_{01}, n_{10} and n_{11} is:

$$\log m_{ij} = \lambda + \lambda_i^A + \lambda_j^B. \tag{1}$$

where we used the identifying restrictions $\lambda_0^A = \lambda_0^B = 0$. Two ways to derive the maximum likelihood estimate of the missed part of the population, are first, by using Poisson loglinear modeling such that $\hat{m}_{00} = \exp(\hat{\lambda})$ and second, by using the property that the odds ratio under independence is 1, i.e., $m_{00}m_{11}/m_{10}m_{01} = 1$ so that:

$$\hat{m}_{00} = \frac{\hat{m}_{10}\hat{m}_{01}}{\hat{m}_{11}} = \frac{n_{10}n_{01}}{n_{11}}. \tag{2}$$

Poisson loglinear modeling is more flexible in more complicated loglinear models using covariates. The odds ratio provides a simple trick to explore capture-recapture methodology, but becomes increasingly difficult to use as the number of registers and covariates increases [10].

2.1 Linkage errors in two registers

An important assumption of capture-recapture methodology is perfect linkage. The assumption is met when all units in register 1 that are also in register 2 are correctly linked to their counterparts in register 2 and when all units in register 2 that are also in register 1 are correctly linked to their counterparts in register 1. We are interested in what happens if the two registers have not been linked perfectly, while capture-recapture analysis is done assuming the assumptions to be met.

When linking two registers the contingency table with the expected values m_{ij} are shown in the left hand side of Table 2.1, where m_{00} will be estimated by the maximum likelihood estimate (2). The population size estimate under assumed perfect linkage is
$\hat{N} = \hat{m}_{11} + \hat{m}_{01} + \hat{m}_{10} + \hat{m}_{00} = n_{11} + n_{01} + n_{10} + \hat{m}_{00}$.

Assume we have linkage errors of size b, where b is the number of false positive links minus the number of false negative links. Then b is negative when the number of false negative links outbalances the number of false positive links, and b is positive when the number of false positive links outbalances the number of false negative links.

Table 2.1 Expected values of being present in register 1 and register 2 under perfect linkage on the left side of the table and the expected values of being present in register 1 and register 2 under linkage error on the right side.

	A		
B	1 (yes)	0 (no)	Total
1 (yes)	m_{11}	m_{10}	m_{1+}
0 (no)	m_{01}	m_{00}	m_{0+}
Total	m_{+1}	m_{+0}	m_{++}

	A		
B	1 (yes)	0 (no)	Total
1 (yes)	$\tilde{m}_{11}$	$\tilde{m}_{10}$	$\tilde{m}_{1+}$
0 (no)	$\tilde{m}_{01}$	$\tilde{m}_{00}$	$\tilde{m}_{0+}$
Total	$\tilde{m}_{+1}$	$\tilde{m}_{+0}$	$\tilde{m}_{++}$

Under perfect linkage we have $b = 0$ for expected values m_{ij} and observed values n_{ij}. Under linkage error $b \neq 0$ we denote expected values by $\tilde{m}_{ij}$, and observed values are denoted $\tilde{n}_{ij}$. Then $\tilde{m}_{11} = m_{11} + b$, $\tilde{m}_{10} = m_{10} - b$ and $\tilde{m}_{01} = m_{01} - b$, and $\tilde{n}_{11} = n_{11} + b$, $\tilde{n}_{10} = n_{10} - b$ and $\tilde{n}_{01} = n_{01} - b$. The contingency table with the expected values $\tilde{m}_{ij}$ is shown in the right hand side of Table 2.1.

Unfortunately, we can not verify from the data to what extent perfect linkage has been violated. We can however use chosen values of b to investigate the effect of linkage error on the population size estimate in a sensitivity analysis. To choose values of b we define linkage error

rate β for m_{01} and thus for observed values n_{01}. For this specific example we have chosen linkage error rate β on m_{01} because m_{01} is the number of added cases of register 2 relative to register 1, such that β is specified based on the implied coverage of register 1 given register 2. Then,

$$\beta = \frac{\tilde{m}_{01}}{m_{01}} = \frac{\tilde{n}_{01}}{n_{01}} \tag{3}$$

where $\beta = 1$ denotes perfect linkage. Linkage error rate β enables us to simulate linkage error, where $\tilde{n}_{01} = n_{01}{}^{*}\beta$. In creating such a linkage error rate β we can denote linkage error in percentages, and by defining linkage error in percentages we can better compare the effect of β on the population size estimate between the two nationality groups. Then,

$$\frac{\hat{\tilde{m}}_{10}\hat{\tilde{m}}_{01}}{\hat{\tilde{m}}_{11}} = \frac{\tilde{n}_{10}\tilde{n}_{01}}{\tilde{n}_{11}} = \frac{(n_{10}-b)(n_{01}-b)}{n_{11}+b} = \hat{m}_{00(\beta)}, \tag{4}$$

where $\hat{m}_{00(\beta)}$ is the size of the individuals missed by the two registers. The population size estimate under linkage error is $\hat{N}_{\beta} = \hat{m}_{00(\beta)} + (n_{11}+b) + (n_{10}-b) + (n_{01}-b)$.

It can be shown that if b is positive, and thus the number of false positive links outbalances the number of false negative links, $\hat{m}_{00(\beta)}$ will be smaller than $\hat{m}_{00}$ and $\hat{m}_{00}$ is an overestimation of $\hat{m}_{00(\beta)}$. If b is negative, and thus the number of false negative links outbalance the number of false positive links, $\hat{m}_{00(\beta)}$ will be larger than $\hat{m}_{00}$ and $\hat{m}_{00}$ is an under estimation of $\hat{m}_{00(\beta)}$. Then, when we estimate $\hat{m}_{00}$ while it has linkage error of size β, estimate $\hat{m}_{00}$ will be biased.

Note that n_{1+} and n_{+1} are fixed values, because these are the total number of individuals in register 1 and register 2. Under linkage error the number of individuals for register 1 and register 2 does not change. However, the sizes of the expected values $\tilde{m}_{ij}$ do change.

2.2 Implied coverage

Under register 1 and 2, the maximum likelihood estimate of the missed portion of the population can be estimated by equation(2). Using (2) we can estimate the conditional probabilities:

$$\hat{p}(0|1) = \frac{n_{01}}{n_{+1}} \text{ and, } \hat{p}(1|1) = \frac{n_{11}}{n_{+1}}, \tag{5}$$

where $\hat{p}(0|1)$ is the estimated probability of n_{01}, given n_{+1}: the conditional, estimated probability of only being registered in register 2, given all registered cases in register 2, including the overlap with register 1. Similarly, $\hat{p}(1|1)$ is the estimated probability of n_{11}, given n_{+1}: the conditional, estimated probability of being in the overlap between register 1 and register 2, given all registered cases in register 2. Thus $\hat{p}(0|1)$ is the estimated probability of new cases from register 2, compared to those cases already registered in register 1, and $\hat{p}(1|1)$ is the estimated probability of already known cases from register 1, compared to all the cases from register 2.

These two probabilities together add up to 1. Given these probabilities we can rewrite equation (2) as

$$\hat{m}_{00} = \frac{n_{10} * \hat{p}(0|1)}{\hat{p}(1|1)}, \tag{6}$$

the number of observations uniquely in register 1, multiplied with the estimated odds of a new observation found in register 2. The larger the estimated odds, the larger $\hat{m}_{00}$ will be. It can be seen from equation (6) that the estimated number of individuals missed by the two registers is a result from the estimated probability of new cases added by register 2, compared to the number of cases already known in register 1.

When the estimated probability of new cases $\hat{p}(0|1)$ is relatively small compared to the estimated probability of already known cases $\hat{p}(1|1)$, the effect of the added new cases of register 2 on $\hat{m}_{00}$ will be small and the population size estimator is robust. Then the coverage of the population by register 1, implied by register 2 is high, which means that register 1 already captures a high number of the individuals in the population compared to register 2.

However, when the estimated probability of new cases $\hat{p}(0|1)$ is relatively large compared to the estimated probability of known cases $\hat{p}(1|1)$, the effect of the added new cases of register 2 on $\hat{m}_{00}$ will be large as well and the population size estimator is not robust. Then the coverage of the population by register 1, implied by register 2 is low, such that register 2 captures a relatively larger number of unique cases compared to register 1. The coverage of register 1 is implied by register 2 because these two registers are our reference point, and the true coverage is unknown. In this paper we will denote the coverage of register 1 implied by register 2 as implied coverage.

2.3 Linkage error for two registers

To illustrate, we use the data on individuals with an Afghan, Iraqi and Iranian nationality residing in the Netherlands, and data of individuals with a Polish nationality residing in the Netherlands. For this section only two of the three registers introduced in the introduction will be used, the Dutch Population Register (PR) and a Crime Suspect Register (CSR). The data are shown in Table 2.2.

Table 2.2 Left are the values of the Afghan, Iraqi and Iranian individuals and on the right are values of the individuals of Polish nationality, estimated values are in italics.

	CSR		
PR	Yes	No	Total
Yes	1,356	58,891	60,247
No	320	*13,898*	*14,218*
Total	1,675	*72,789*	*74,465*

	CSR		
PR	Yes	No	Total
Yes	444	42,109	42,553
No	1,659	*157,340*	*158,999*
Total	2,103	*199,449*	*201,552*

We illustrate the use of β and b by using data of individuals with an Afghan, Iraqi and Iranian nationality. Assuming perfect linkage under $\beta = 1$, we get the maximum likelihood estimate of the missed part of the population by $58,891 * 320/1,356 = 13,898$. Then a total of $\hat{N} = 1,356 + 58,891 + 320 + 13,898 = 74,465$ individuals with an Afghan, Iraqi and Iranian nationality are residing in the Netherlands.

However, when we introduce linkage error of size $\beta = 0.9$, then $\tilde{n}_{01} = n_{01} * 0.9 = 320 * 0.9 = 288$, and $b = 320 - 288 = 32$. Then

$\tilde{n}_{10} = 58,891 - 32 = 58,859$ and $\tilde{n}_{11} = 1,356 + 32 = 1,388$, which can be seen on the right hand side of Table 2.3. We estimate that 12,213 individuals are missed by the population. Then we have a total of $\hat{N}_{\beta} = 1,388 + 58,859 + 288 + 12,213 = 72,748$ individuals with an Afghan, Iraqi and Iranian individuals residing in the Netherlands. Thus the estimate of 74,465 is a result of capture-recapture analysis when assuming perfect linkage, and will be an overestimation if we have linkage error and have not adjusted for this.

Table 2.3 Left are the values of the Afghan, Iraqi and Iranian individuals as observed in the registers, and on the right are the values after linkage error adjustment, estimated values are in italics.

	CSR		
PR	Yes	No	Total
Yes	1,356	58,891	60,247
No	320	*13,898*	*14,218*
Total	1,675	*72,789*	*74,464*

	CSR		
PR	Yes	No	Total
Yes	1,388	58,859	60,247
No	288	*12,213*	*12,501*
Total	1,676	*71,072*	*72,748*

Parameter β enables us to conduct a sensitivity analysis on linkage errors. The results can be seen in Figure 2.1 and details can be found in the Appendix. When a linkage error range of $\beta = 0.5$ to $\beta = 1.5$ is introduced on the data of the Afghan, Iraqi and Iranian individuals, the estimate of the missed portion of the population does not differ greatly from the estimate under perfect linkage. If we introduce linkage error of size $\beta = 0.5$ (and also for $\beta = 1.5$) and estimate assuming perfect linkage without adjusting for the linkage error, we have a bias of only 12 percent. Thus there is only a 12 percent difference between the actual population size estimate $\hat{N}_{\beta}$, and the population size estimate under the observed values $\hat{N}$ where perfect linkage is assumed. Thus for the Afghan, Iraqi and Iranian individuals the population size estimator is relatively robust.

To compare we also conducted a sensitivity analysis on the effect of linkage error on the individuals with a Polish nationality. We introduced linkage error range of $\beta = 0.5$ to $\beta = 1.5$. As can be seen from the figure, the population size estimator is not robust to linkage error. Introducing linkage error of $\beta = 0.5$, we overestimate the population size estimate with 187 percent, compared to when we would estimate under $\beta = 1$. Note that the upper range of β is lower for the individuals with a Polish nationality because of the low cell count of 444 in cell (1,1). For the individuals with an Afghan, Iraqi and Iranian nationality the population size estimate under linkage errors is quite stable, for those with a Polish nationality this is not the case.

Thus for the Afghan, Iraqi and Iranian individuals the population size estimator is relatively robust to linkage errors, but for the Polish individuals the population size estimator is not robust. The reason is that for the former, there is a larger implied coverage of the PR given the CSR than for the latter. For the individuals with an Afghan, Iraqi and Iranian individuals $\hat{p}(0|1) = 320/1,675 = 0.19$ and $\hat{p}(1|1) = 1,356/1676 = 0.81$, which means that the CSR does not add many new cases to the PR and the estimated conditional coverage implied by the PR given the CSR is high. For the Polish individuals $\hat{p}(0|1) = 1,659/2,103 = 0.79$ and $\hat{p}(1|1) = 444/2,103 = 0.21$, which means that the CSR actually adds a lot of new cases to the PR and the estimated conditional coverage implied by the PR given the CSR is low. Then for the individuals with an Afghan, Iraqi and Iranian, due to the high implied coverage the population size estimator is robust. Whereas for the Polish individuals, due to the low implied coverage the population size estimator is not robust at all.

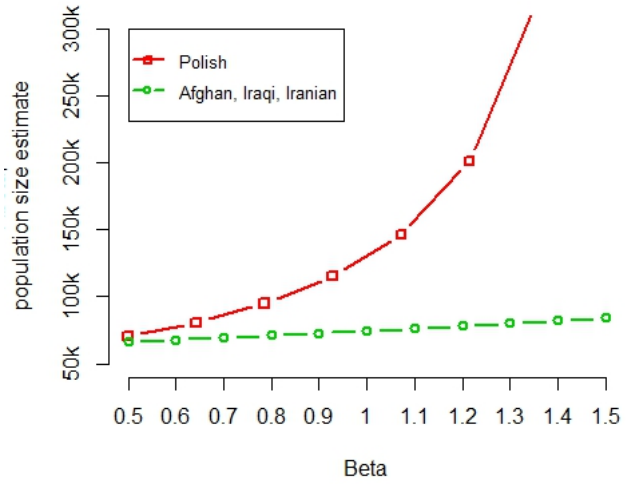

Figure 2.1 Population size estimate for the two register capture recapture sensitivity analysis for both nationalities under a β ranging from 0.5 to 1.5. $\beta = 1$ defines perfect linkage.

2.4 Erroneous captures, theoretical example

Another important assumption of capture-recapture methodology is that all units in the observed data belong to the population, and thus the data contain no erroneous captures in either the first or the second capture. It is not always possible to assess from the data which units actually belong to the population and which do not. Therefore we aim to investigate the effect on the population size estimate when erroneous captures are present.

Again m_{ij} are the expected values of the observed values n_{ij}. For m_{ij} and n_{ij} we assume no erroneous captures. If erroneous captures are introduced expected values are denoted by $\bar{m}_{ij}$, and observed values under linkage error are $\bar{n}_{ij}$. We can define an erroneous capture rate γ for n_{01}, where $\gamma = \bar{n}_{01}/n_{01}$, such that $\bar{n}_{01} = n_{01}{}^{*}\gamma$. Erroneous captures are units in the data that should not have been observed, and therefore $\bar{n}_{01}$ will always be smaller than n_{01}, and $0 \leq \gamma \leq 1$. Erroneous capture rate γ has been defined on n_{01} because that is the number of added cases by register 2, relative to register 1. We find

$$\hat{m}_{00(\gamma)} = \frac{n_{10}(n_{01}{}^{*}\gamma)}{n_{11}} = \frac{n_{10}\bar{n}_{01}}{n_{11}} = \gamma\hat{m}_{00}, \tag{7}$$

where $\hat{m}_{00}$ is the estimate when there are no erroneous captures defined in (2). The effect of erroneous captures is thus:

$$N = n_{10} + n_{01} + n_{11} + \tilde{m}_{00} = a + \gamma\hat{m}_{00}, \tag{8}$$

which is a straight line with intercept $a = n_{10} + n_{01} + n_{11}$ and slope γ, if a is large compared to $\hat{m}_{00}$ the line is more horizontal and the estimate is more robust.

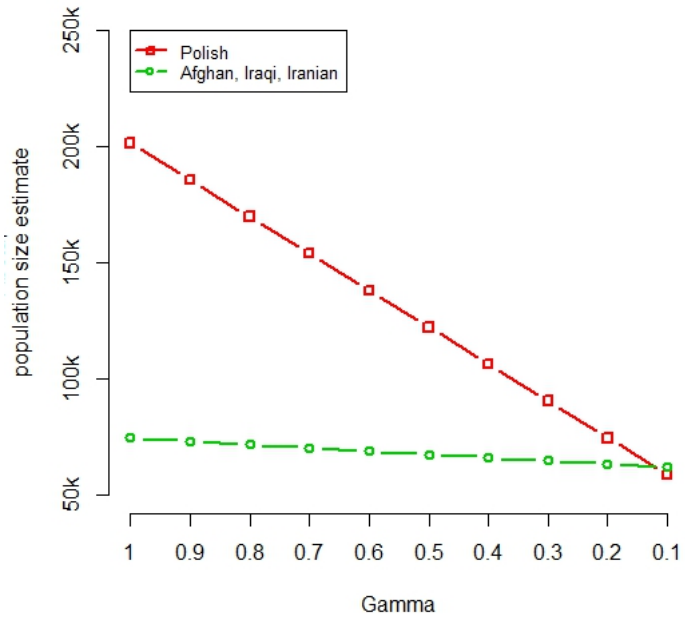

Figure 2.2 Population size estimate for the two register capture-recapture sensitivity analysis for both nationalities under a Gamma ranging from 1 to 0.1. Under $\gamma = 1$ we assume no erroneous captures, and when $\gamma < 1$ the data contain erroneous captures.

2.5 Erroneous captures

For the sensitivity analyses to study the effect of erroneous captures on the population size estimator, we use again the data from Table 2.2. We chose the range $\gamma = 0.9$ to $\gamma = 0.1$. This is a rather extreme range, such extreme ranges may not be very realistic but it does give us a complete insight into the effect of γ. Table 2.4 shows the expected values when there are 10 percent erroneous captures in n_{01} and hence $\gamma = 0.9$. Then
$\tilde{n}_{01} = n_{01}{}^*\gamma = n_{01}{}^*0.9 = 320{}^*0.9 = 288.$

Table 2.4 The values for the Afghan, Iraqi and Iranian people residing in the Netherlands in 2010, adjusted for $\gamma = 0.9$ erroneous capture in the CSR. Values in italics are estimated values

	CSR		
PR	Yes	No	Total
Yes	1,356	58,891	60,247
No	288	*12,508*	*12,796*
Total	1,644	*71,399*	*73,043*

It follows that, $\tilde{m}_{00(\gamma)} = 12,508$, which is 1,390 cases smaller than $\hat{m}_{00} = 13,898$.
Figure **??**figuur2) shows the sensitivity analysis for both nationalities. For the individuals with an Afghan, Iraqi and Iranian nationality erroneous captures have only a small effect on the population size estimator and the estimator is relatively robust. Figure 2 also shows the sensitivity analysis for the Polish individuals. For the Polish individuals erroneous captures have a large effect on the population size estimator and the estimator is not robust.

Here again the population size estimator is more robust against violation of an assumption for the data of the Afghan, Iraqi and Iranian individuals than for the Polish individuals. The effect of

erroneous captures is smaller than the effect of linkage error, because it only affects one expected value, whereas linkage errors affect all three cells, compare (4) and (7).

Here again the relative size of the implied coverage of the PR given the ER is the explanation. Given erroneous captures are deleted from n_{01}, it will influence the implied coverage. For the Afghan, Iraqi and Iranian individuals $\hat{p}(0|1) = 320/1,675 = 0.19$, but when $\gamma = 0.9$, $n_{01} = 288$ instead of 320 and $\hat{p}(0|1) = 288/1,675 = 0.17$. Thus when eliminating the erroneous captures from n_{01}, the implied coverage of the PR given the CSR increases.

3 Capture-recapture for three registers.

In section 2 we discussed the impact of linkage errors and erroneous captures in the simplest form of capture-recapture for 2 registers. However, often multiple registers are available. When linkage errors or erroneous captures are present in three registers the effect of these violations may become more complex to explain. We now investigate for three registers what the effect is of violations of perfect linkage and no erroneous captures on the population size estimator.

Assume we have three registers, 1, 2 and 3. Let variables A, B and C respectively denote inclusion in registers 1, 2 and 3. Let the levels of A be indexed by i (i = 0,1) where $i = 0$ stands for "not included in register 1", and $i = 1$, stands for "included in register 1". Similarly, let the levels of B be indexed by j ($j = 0, 1$), and let the levels of C be indexed by k ($k = 0, 1$). Table 3.1 shows the expected values denoted by m_{ijk}. Observed values are denoted by n_{ijk} with $n_{000} = 0$.

Table 3.1 The table of expected counts for three registers

		C	
A	B	1 (Yes)	0 (No)
1 (Yes)	1 (Yes)	m_{111}	m_{110}
	0 (No)	m_{101}	m_{100}
0(No)	1 (Yes)	m_{011}	m_{010}
	0 (No)	m_{001}	m_{000}

For three variables the saturated loglinear model is denoted by:

$$\log m_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC}, \tag{9}$$

with identifying restrictions that a parameter equals zero when i, j or $k = 0$. The model assumption is that the three factor interaction parameter $\lambda_{ijk}^{ABC} = 0$. Model $[AB][BC][AC]$ is the saturated model, as the number of observed parameters equals the number of parameters to be estimated. This model assumes that the odds ratio between A and B is the same for $k = 0$ and $k = 1$, i.e.,

$$\frac{m_{110}m_{000}}{m_{100}m_{010}} = \frac{m_{111}m_{001}}{m_{101}m_{011}}. \tag{10}$$

An estimate for $\hat{m}_{000}$ is easily derived from (10).

$$\frac{\hat{m}_{010}\hat{m}_{001}\hat{m}_{100}\hat{m}_{111}}{\hat{m}_{011}\hat{m}_{110}\hat{m}_{101}} = \frac{n_{010}n_{001}n_{100}n_{111}}{n_{011}n_{110}n_{101}} = \hat{m}_{000}. \tag{11}$$

When linkage error is present, the same rules apply as for two registers. Here again we can define a linkage rate β, where we investigate linkage error only in one register. As an example, linkage error rate β is only investigated in register 3:

$$\beta = \frac{\tilde{n}_{001}}{n_{001}}. \tag{12}$$

Under these conditions, we can again conduct a sensitivity analysis on data where three registers are linked.

3.1 Linkage error in three registers

For our sensitivity analysis we focus on the CSR, because this is the smallest register and more prone to errors. We again use data on the Afghan, Iraqi and Iranian individuals residing in the Netherlands which is shown on the left side of Table 3.2, however now by the three registers. Additionally, we again use the data on the Polish individuals residing in the Netherlands, shown on the right side of Table 3.2, also by the three registers.

Table 3.2 The observed values for the Afghan, Iraqi and Iranian individuals by three registers on the left side of the table and the observed values for the Polish individuals on the right side of the Table.

		CSR	
PR	ER	Yes	No
Yes	Yes	309	13,862
	No	1,647	45,029
No	Yes	7	266
	No	313	0

		CSR	
PR	ER	Yes	No
Yes	Yes	215	24,832
	No	229	17,277
No	Yes	230	80,406
	No	1,429	0

We introduce linkage error rate $\beta < 1$, where individuals in the CSR are to be linked to either the ER or the PR. A part of the CSR individuals that did not link to the PR will in part also have to be linked to the intersection of PR and ER. Then for the individuals in the CSR that will be linked to the PR, a part will also be linked to the intersection of PR and ER. The proportion of linkage errors in the CSR to be linked to either the PR or the intersection of the PR and the ER will be the same as the distribution of the CSR individuals that are in the PR. Then, for the individuals with an Afghan, Iraqi and Iranian nationality 309 individuals are in the intersection of all three registers, compared to 1,647 in the PR and CSR, which is only 16 percent of all the individuals in both the PR and the CSR. The percentage linkage error of size β related to 313 that should have linked to the PR will link for 16 percent to the intersection of the PR and the ER and for 84 percent to the PR alone.

As can be seen from Figure 3.1, again for the Afghan, Iraqi and Iranian individuals linkage errors have only a small effect on the population size estimation. However, for the Polish individuals linkage errors have a large effect on the population size estimation. Table I.6 in the Appendix

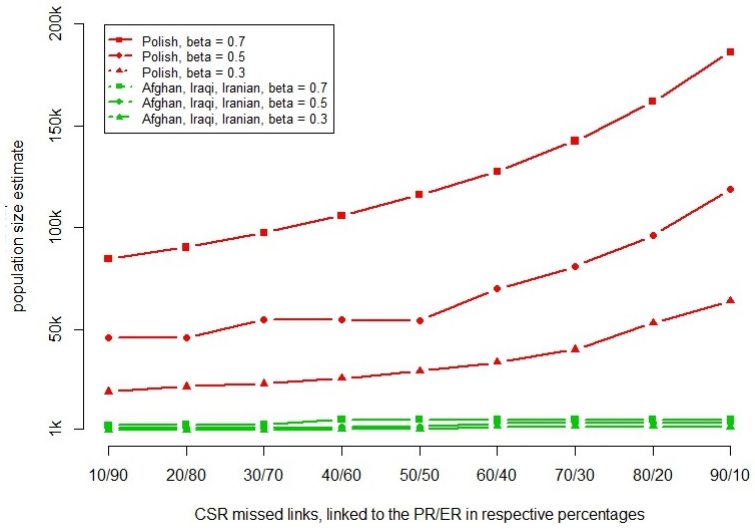

Figure 3.1 Population size estimate for both nationalities with linkage error rate $\beta = 0.7, 0.5$ or 0.3. The respective b has been distributed over the PR and ER according the percentages on the X axis. For example the first tick on the X axis, $10/90$ means 10 percent of b were linked to the PR and 90 percent of b to the ER. Note that part of b linked to the PR, a beforehand specified part is distributed to the intersection of PR and ER.

shows the estimates for the full sensitivity analysis. However, compared to the Polish individuals, the capture-recapture analysis for the Afghan, Iraqi and Iranian individuals is quite robust under three registers.

The difference between the effect linkage errors have on the population size estimate between the nationalities again are the result of the implied coverage. In the sensitivity analysis we link individuals from the CSR to the PR and the ER. However, for the Afghan, Iraqi and Iranian individuals there are only 313 individuals in the CSR that did not link to the PR and ER, which is 1,429 cases for the Polish individuals. The coverage of the PR and ER is smaller for the Polish individuals compared to the Afghan, Iraqi and Iranian individuals, given 1,429 cases are new cases added by the CSR compared to only 313. Thus for the Polish individuals the population size estimate is less robust than for the Afghan, Iraqi and Iranian individuals.

3.2 Three registers - erroneous captures

When three registers are used and one register contains erroneous captures, a sensitivity analysis with γ can be carried out. We present here one example. We define an erroneous capture rate $\gamma = \bar{n}_{001}/n_{001}$. Erroneous capture rate γ was introduced for $\bar{n}_{001}$ because the CSR has the most administrative error, thus the added cases of the CSR relative to the PR and ER may have the highest probability of erroneous captures. For the sensitivity analysis a range from $\gamma = 0.9$ to $\gamma = 0.1$ of erroneous captures are introduced to the observed values.

The bias resulting from erroneous captures in the capture-recapture analysis for both nationalities can be found in Figure 3.2. For erroneous captures in three registers, the population size estimate for the individuals with a Polish nationality is not robust to violating this

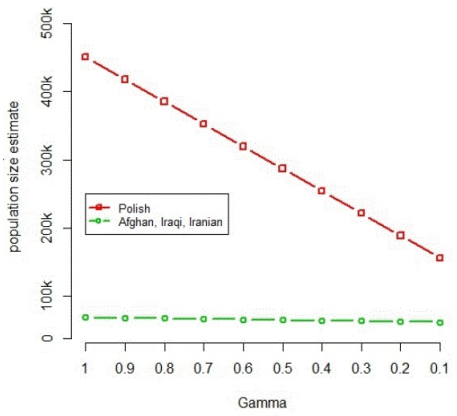

Figure 3.2 Population size estimate for both nationalities under a Gamma ranging from 1 to 0.1. Under $\gamma < 1$ no erroneous captures are assumed, and for $\gamma \neq 1$ erroneous captures are assumed

assumption. When assuming that all the individuals belong to the population, the total number of Polish individuals in the Netherlands is 450,945. When however erroneous captures of $\gamma = 0.5$ is introduced, the population size is 286,953. Thus we overestimate the population size estimate by 65 percent when erroneous captures are present.

Again we find that when violating erroneous captures the population size estimate for the individuals with an Afghan, Iraqi and Iranian nationality is quite robust to erroneous captures. Under no erroneous captures the total number of Afghan, Iraqi and Iranian individuals in the Netherlands is 68,682. When, however, there are $\gamma = 0.5$, thus 50 percent, erroneous captures in the CSR the population size estimate is 64,889, and we overestimate by only six percent.

As was stated in the previous section, the implied coverage of the PR and ER is higher for the Afghan, Iraqi and Iranian individuals than for the Polish individuals. As such the population size estimate under erroneous captures are more robust for the Afghan, Iraqi and Iranian individuals than for the Polish individuals.

4 Conclusion

In this paper we have compared two rather different nationality groups: data of Afghan, Iraqi and Iranian individuals residing in the Netherlands are compared to data of Polish individuals residing in the Netherlands to assess the effect of linkage errors and erroneous captures on the population size estimate. These two different nationality groups have been chosen because they have different legal requirements to reside in the Netherlands. This results in two rather different contingency tables, as can be seen from Table 2.2 and 3.2. Both nationality groups have a high number of individuals registered in the PR only. However, for the individuals registered in the CSR there is a large difference between the nationality groups on the implied coverage of the PR over the CSR.

Because individuals with a Polish nationality are free to move and work in the EU, less Polish individuals registered in the CSR are also in the intersection with PR. Thus the implied coverage

of the PR relative to the CSR is low. For the individuals with an Afghan, Iraqi and Iranian nationality we see the opposite. Because these individuals need a working or residence permit to enter the Netherlands more individuals that are registered in the CSR are also in the intersection with the PR. This is probably due to the fact that those CSR registered individuals that are not registered in the PR are illegally residing in the Netherlands, whereas legally the Polish individuals only registered in the CSR are not illegally residing in the Netherlands.

For the two register case, because of the different conditional probabilities of the PR, given the CSR between the two nationality groups, we see a difference in the effect that violations have on the population size estimator. Because the implied coverage of the PR over the CSR is different between the two nationality groups, we see that the effect of linkage errors and erroneous captures is more dramatic for the Polish data than for the Afghan, Iraqi and Iranian data. Because the implied coverage of the PR given the CSR for the Afghan, Iraqi and Iranian individuals is already relatively high, given the CSR add relatively few new cases, the population size estimate is more robust to violation of the assumptions. However, for the Polish data the implied coverage of the PR given the CSR is rather small, such that the population size estimator is less robust to violations of the same percentage as for the Afghan, Iraqi and Iranian individuals.

Given the implied coverage has a substantial impact on the population size estimation when assumptions are violated, it is important that all units are linked correctly. Currently there are some developments in the theory and practice of capture-recapture methods that aims at linkage error-unbiased estimates. One of which is the research from [5], based on [7] for probabilistic linkage, where they propose to use the estimated number of false positives and false negatives. For our current data we were unable to use their method, because of the 37% of the individuals in the CSR that are without background information. These cases create unrealistic false positives and false negatives probabilities.

We have assumed in this paper that when assessing the effect of either linkage error or erroneous captures on the population size estimate, all other assumptions are met. This is not very realistic. However, this did allow us to assess the effect of violating the assumptions on the population size estimator. To introduce more than one violation per data set, would make the presentation more complex.

In this paper we have chosen relatively random linkage error rates and erroneous capture rates. These were chosen to assess the effect of minor deviations in perfect linkage and no erroneous captures to a more extreme deviation. Given that we can not assess the extent of deviation of the assumptions in observed data, we use sensitivity analyses to assess the effect of relative to extreme deviations.

In this paper, implied coverage is important. When the implied coverage of the PR given the CSR is large, the population size estimate is relatively robust to violations of assumptions. However when the implied coverage of the PR given the CSR is small, the population size estimate is not robust to violations of the assumptions. For the individuals with an Afghan, Iraqi en Iranian nationality we can conclude that deviations will have an effect, albeit a small one. For the individuals with a Polish nationality it can be concluded that even the smallest deviation will give a large effect on the population size estimate. In conclusion, it is important to assess the implied coverage of the registers, because this will have an effect on the population size estimate.

References

[1] Y. M. M. Bishop, S. E. Fienberg, and P. W. Holland. *Discrete multivariate analysis*. MIT press, Cambridge, MA., 1975.

[2] I. Boden, L. Capture–recapture estimates of the undercount of workplace injuries and illnesses: Sensitivity analysis. *American Journal Of Industrial Medicine*, 57:1090 – 1099, 2014.

[3] J. J. Brown, I. D. Diamond, R. L. Chambers, L. J. Buckner, and A. D. Teague. A methodological strategy for a one-number Census in the UK. *Journal of the Royal Statistical Society. Series A*, 162:247 – 267, 1999.

[4] A. Chao, P. K. Tsay, S-H. Lin, W-Y. Shau, and D-Y. Chao. Tutorial in biostatistics. the application of capture-recapture models of epidemiological data. *Statistics in Medicine*, 20:3123 – 3157, 2001.

[5] L. D. Consiglio and T. Tuoto. Coverage evaluation on probabilistically linked data. *Journal of Official Statistics*, 31:415 – 429, 2015.

[6] R. M. Cormack. Log-linear models for capture-recapture. *Biometrics*, 45:395 – 413, 1989.

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

[8] I. P. Fellegi and A. B. Sunter. A theory of record linkage. *Journal of the American Statistical Association*, 64:1183 – 1210, 1969.

[9] S. E. Fienberg. The multiple recapture census for closed populations and incomplete 2k contingency tables. *Biometrika*, 59:409 – 439, 1972.

[10] S. C. Gerritse, P. G. M. Van der Heijden, and B. F. M. Bakker. Sensitivity of population size estimation for violating parametric assumptions in loglinear models. *Journal of Official Statistics*, 31:357 – 379, 2015.

[11] International Working Group for Disease Monitoring and Forecasting. Capture-recapture and multiple-record systems estimation I: History and theoretical development. *American Journal of Epidemiology*, 142:1047 – 1058, 1995.

[12] G. A. F. Seber. *The Estimation of Animal Abundance and Related Parameters*. Edward Arnold: London, 1982.

[13] P. G. M. Van der Heijden, J. Whittaker, M. J. L. F Cruyff, B. F. M. Bakker, and H. N. Van der Vliet. People born in the Middle East but residing in the Netherlands: Invariant population size estimates and the role of active and passive covariates. *The Annals of Applied Statistics.*, 6:831 – 852, 2012.

[14] E. N. Zwane and P. G. M. Van der Heijden. Analysing capture-recapture data when some variables of heterogeneous catchability are not collected or asked in all registries. *Statistics in Medicine*, 26:1069 – 1089, 2007.

Appendix

I Tables

I.1 Tables for section 2.1

Results for the sensitivity analysis for linkage error under error rate β in two registers for the Afghan, Iraqi and Iranian individuals can be found in Table I.1. The first row shows the estimate of the missed portion of the population under a β from 0.5 to 1.5. The second row gives the population size estimate $\hat{N} = m_{11} + m_{01} + m_{10} + \hat{m}_{00}$. The third row gives a relative bias, the bias of the estimate under assumed perfect linkage $\hat{N}$ to the estimate adjusted for linkage error $\hat{N}(\beta)$, where $\hat{N}_{(\beta)} = \hat{m}_{00(\beta)} + m_{11+b} + m_{10-b} + m_{01-b}$.

Table I.1 Robustness analysis of the population size estimate for the people residing in the Netherlands in 2010 with an Afghan, Iraqi and Iranian nationality.

β	0.5	0.6	0.7	0.8	0.9	1	1.1	1.2	1.3	1.4	1.5
$\hat{m}_{00(\beta)}$	6,199	7,603	9,070	10,605	12,213	13,898	15,665	17,522	19,475	21,531	23,699
$\hat{N}_{(\beta)}$	66,606	68,042	69,541	71,208	72,748	74,465	76,264	78,153	80,138	82,226	84,426
$\hat{N}/\hat{N}_{(\beta)}$	1.12	1.09	1.07	1.05	1.02	1	0.98	0.95	0.93	0.90	0.88

Results for the sensitivity analysis for linkage error under error rate β in two registers for the Polish individuals can be found in Table I.2. The first row shows the estimate of the missed portion of the population under a β from 0.5 to 1.2, given that it was impossible to take more linkage error. The second row gives the population size estimate $\hat{N}$. The third row gives a relative bias, the bias of the estimate under assumed perfect linkage $\hat{N}$ to the estimate adjusted for linkage error $\hat{N}(\beta)$.

Table I.2 Robustness analysis of the population size estimate for the people residing in the Netherlands in 2010 with a Polish nationality.

β	0.5	0.6	0.7	0.8	0.9	1	1.1	1.2
$\hat{m}_{00(\beta)}$	26,894	37,218	51,285	71,441	102,657	157,340	277,525	754,465
$\hat{N}_{(\beta)}$	70,278	80,766	94,999	115,321	146,703	201,552	321,903	799,009
$\hat{N}/\hat{N}_{(\beta)}$	2.87	2.50	2.12	1.74	1.37	1	0.63	0.25

I.2 Tables for section 2.2

Table I.3 shows the robustness analysis for the Afghan, Iraqi and Iranian people considering erroneous captures of size γ in the CSR register. Row $\hat{m}_{00(\gamma)}$ shows the maximum likelihood estimate under erroneous captures of size γ. The second row is the total population size estimate $\hat{N}_{(\gamma)} = \hat{m}_{00(\gamma)} + m_{10} + m_{01\gamma} + m_{11} = \hat{N} + (\gamma - 1)(\hat{m}_{00} + m_{01})$. The third row is the relative bias of $\hat{N}_{(\gamma)}$ to $\hat{N}$.

Table I.3 Robustness analysis of the population size estimate for the people residing in the Netherlands in 2010 with an Afghan, Iraqi and Iranian nationality when adjusting for γ erroneous captures.

γ	1	0.9	0.8	0.7	0.6	0.5	0.4	0.4	0.2	0.1
$\hat{m}_{00(\gamma)}$	13,898	12,508	11,118	9,728	8,339	6,949	5,559	4,169	2,780	1,390
$\hat{N}_{(\gamma)}$	74,465	73,043	71,621	70,199	68,778	67,356	65,934	64,512	63,091	61,669
$\hat{N}/\hat{N}_{(\gamma)}$	1	1.02	1.04	1.06	1.08	1.11	1.13	1.15	1.18	1.21

Table I.3 shows the robustness analysis for the Polish people considering erroneous captures of size γ in the CSR register. Row $\hat{m}_{00(\gamma)}$ shows the maximum likelihood estimate under erroneous captures of size γ. The second row is the total population size estimate $\hat{N}_{(\gamma)}$. The third row is the relative bias of $\hat{N}_{(\gamma)}$ to $\hat{N}$.

Table I.4 Robustness analysis of the population size estimate for the people residing in the Netherlands in 2010 with a Polish nationality,when adjusting for γ erroneous captures.

γ	1	0.9	0.8	0.7	0.6	0.5	0.4	0.4	0.2	0.1
$\hat{m}_{00(\gamma)}$	157,340	141,596	125,853	110,109	94,366	78,717	62,974	47,230	31,487	15,743
$\hat{N}_{(\gamma)}$	201,552	185,642	169,733	153,824	137,914	122,100	106,190	90,281	74,372	58,462
$\hat{N}/\hat{N}_{(\gamma)}$	1	1.09	1.19	1.31	1.46	1.65	1.90	2.23	2.71	3.45

I.3 Tables for section 3.2

Table I.5 shows the sensitivity analysis for the individuals with a Polish nationality by the three registers. The rows display the percentage that has been taken from the CSR. The columns show how much of the percentage taken from the CSR has been linked to either the PR or the ER.

Table I.6 shows the sensitivity analysis for the individuals with an Afghan, Iraqi and Iranian nationality by the three registers. The rows display the percentage that has been taken from the CSR. The columns show how much of the percentage taken from the CSR has been linked to either the PR or the ER.

Table I.5 Resulting population size estimates when percentages linkage errors are taken from the CSR (rows) and they are divided differently over the PR and ER (columns) for the Polish individuals.

Percentage in PR→	10	20	30	40	50	60	70	80	90
Percentage in ER →	90	80	70	60	50	40	30	20	10
Percentage from CSR									
10	188,054	195,671	204,215	211,794	222,621	232,873	245,087	258,598	273,813
20	122,628	129,366	137,900	148,190	159,065	171,643	186,595	205,168	226,379
30	84,768	90,377	97,674	105,885	116,075	127,696	142,554	161,835	186,264
40	55,395	64,623	70,021	76,776	84,787	95,314	108,542	125,916	150,558
50	46,185	45,966	54,906	54,723	54,544	70,060	81,068	96,087	118,808
60	29,404	30,769	35,100	39,108	43,869	54,047	53,828	70,987	89,733
70	19,542	22,000	23,439	26,110	29,520	34,019	40,220	53,220	63,912
80	11,684	12,730	14,074	15,668	17,805	20,644	24,625	30,536	40,534
90	5,280	5,754	6,366	7,132	8,146	9,477	11,372	14,290	19,259

I.4 Tables for section 3.3

Table I.7 shows the robustness analysis for the Afghan, Iraqi and Iranian people considering erroneous captures of size γ in the CSR register. Row $\hat{m}_{00(\gamma)}$ shows the maximum likelihood estimate under erroneous captures of size γ. The second row is the total population size estimate $\hat{N}_{(\gamma)}$. The third row is the relative bias of $\hat{N}_{(\gamma)}$ to $\hat{N}$.

Table I.6 Resulting population size estimates when percentages linkage errors are taken from the CSR (rows) and they are divided differently over the PR and ER (columns) for the Afghan, Iraqi and Iranian individuals.

Percentage in PR →	10	20	30	40	50	60	70	80	90
Percentage in ER →	90	80	70	60	50	40	30	20	10
Percentage from CSR									
10	7,695	7,686	7,671	7,657	7,647	7,632	7,623	7,609	7,594
20	6,813	6,788	6,766	6,745	6,724	6,699	6,678	6,658	6,637
30	3,024	3,185	3,340	5,868	5,843	5,814	5,785	5,757	5,729
40	2,190	2,377	2,547	2,733	4,973	4,942	4,909	4,880	4,847
50	1,488	1,675	1,856	2,042	2,235	4,061	4,030	3,998	3,967
60	923	1,104	1,285	1,455	1,635	3,223	3,194	3,164	3,134
70	495	651	810	965	1,117	2,402	2,375	2,350	2,324
80	133	214	295	376	457	1,595	1,575	1,556	1,537
90	39	133	224	314	405	495	585	693	769

Table I.7 Erroneous captures for three registers for the Afghan, Iraqi and Iranian individuals

γ	1.00	0.90	0.80	0.70	0.60	0.50	0.40	0.30	0.20	0.10
$\hat{m}_{00(\gamma)}$	7,249	6,531	5,790	5,072	4,354	3,613	2,895	2,177	1,459	718
$\hat{N}_{(\gamma)}$	68,682	67,932	67,160	66,411	65,662	64,889	64,140	63,391	62,642	61,869
$\hat{N}/\hat{N}_{(\gamma)}$	1.00	1.01	1.02	1.03	1.05	1.06	1.07	1.08	1.10	1.11

Table I.8 Erroneous capture for three registers for the Polish individuals

γ	1.00	0.90	0.80	0.70	0.60	0.50	0.40	0.30	0.20	0.10
$\hat{m}_{00(\gamma)}$	326,327	293,671	261,016	228,360	195,705	163,049	130,622	97,966	65,311	32,655
$\hat{N}_{(\gamma)}$	450,945	418,146	385,348	352,549	319,751	286,953	254,383	221,584	188,786	155,987
$\hat{N}/\hat{N}_{(\gamma)}$	1.00	1.08	1.17	1.28	1.41	1.57	1.77	2.04	2.39	2.89

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

Prepress: Statistics Netherlands, Grafimedia
Design: Edenspiekermann

Information
Telephone +31 88 570 70 70, fax +31 70 337 59 94
Via contact form: www.cbs.nl/information

Where to order
verkoop@cbs.nl
Fax +31 45 570 62 68
ISSN 1572-0314