Statistical Journal of the IAOS 31 (2015) 613–627
DOI 10.3233/SJI-150938
IOS Press

Different methods to complete datasets used for capture-recapture estimation: Estimating the number of usual residents in the Netherlands

Susanna C. Gerritse[a,*], Bart F. M. Bakker[b] and Peter G. M. van der Heijden[c]
[a] *Utrecht University, Utrecht, The Netherlands*
[b] *Statistics Netherlands/VU University Amsterdam, Den Haag, The Netherlands*
[c] *Utrecht University/University of Southampton, Utrecht, The Netherlands*

Abstract. We are interested in an estimate of the usual residents in the Netherlands. Capture-recapture estimation with three registers enables us to estimate the size of the total population, of which the usual residents are a part. However, usual residence cannot be used as a covariate because it is not available in one of the registers. We approach this as a missing data problem. There are different methods available to handle missing data. In this manuscript we use Expectation Maximization (EM) algorithm and Predictive Mean Matching (PMM). The EM algorithm is often used in categorical data analysis, but PMM has the advantage of flexibility in the choice for a specific part of the observed data used for the imputation of the missing data. Four scenarios have been identified where the missing data are completed via either the EM algorithm or PMM imputation, resulting in different population size estimates for usual residence. It was found that the different scenarios lead to different population size estimates. Even small changes in the completed data lead to different population size estimates. In this study PMM imputation performs best according flexibility and it is theoretically better motivated.

Keywords: Predictive mean matching, EM-algorithm, capture-recapture, usual residents, census

1. Introduction

In this manuscript we are interested in the estimation of the population size of the so-called usual residents in the Netherlands. According to the European Union, Regulation (EU) No 1260/2013 of the European Parliament, usual residence is defined as "The place where a person normally spends the daily period of rest, regardless of temporary absences for purposes of recreation, holidays, visits to friends and relatives, business, medical treatment or religious pilgrimage". An individual is considered a usual resident when they have lived in the Netherlands for a continuous period of 12 months before the reference time, or if they arrived in the 12 months before the reference time and intend to stay for at least a year. When these circumstances can not be established, "usual residence" means the place of registered residence. The registers used in this manuscript however may register a form of length of stay based on registration date, but not an intent to stay. Hence, only a length of stay may be used to assess usual residence.

The Netherlands has the advantage of having a population register (PR), wherein all registered individuals are documented. Even though for a large part the PR entails the usual residents, for obvious reasons part of the usual residents will be missed by the PR. This incompleteness of the PR has more than one reason.

[*] Corresponding author: Susanna C. Gerritse, Utrecht University, Padualaan 14, 3584 CH Utrecht, The Netherlands. E-mail: sc.gerritse@gmail.com.

First, within the European Union there is free movement and employment for individuals with a European Union nationality. Usual residents with a European Union nationality do have to register themselves in the PR. Specific rights and services can be provided only to individuals officially registered in the PR. However, unregistered individuals might have forgotten to register, do not know they need to register, or do not want to. Second, the PR is also incomplete due to immigrants, coming from outside the European Union without a working or residence permit. Additionally, as the actual population of the Dutch Census is restricted to those who are registered in the PR regardless their residence duration, the estimated true number of usual residents minus the registered population is an indicator of its under count. The registered population will also contain an over count. An over count may occur when registered individuals no longer reside in the Netherlands because of emigration, because they passed away or in the case of administrative delay. In the Netherlands however this is not as big a problem as the under count. Bakker [2] estimated an over count of 31 thousand individuals, which is only 0.2 percent of the Dutch population.

Because the PR alone is not sufficient to determine the number of usual residents, we linked the PR to an Employment Register (ER) and a Crime Suspects Register (CSR). This enables us to use capture-recapture methodology to assess the part of the population missed by all three registers [4,7,10,16]. However, because we are interested only in the usual residents of this population, the statistical problem is more complicated. For the PR and the ER residence duration can be derived. However, in the CSR there is no information on residence duration at all. Part of this lack of information in the CSR is solved because this register is linked to the PR and the ER. For the CSR individuals not linked to the PR and/or the ER the information on residence duration is missing.

Therefore we are dealing not only with estimating a population size, but also with handling missing data, since the covariate usual residence is partially missing. Partially missing covariates are usually ignored in capture-recapture problems because they lead to missing data in one or more registers. However, because the covariate usual residence is central in our research question, we cannot ignore it, and we have to solve this missing data problem before we can estimate the population size using capture-recapture methodology.

In estimating usual residence we are interested in what method handles our missing data problem best. When missingness is Missing At Random (MAR) the Expectation Maximization (EM) algorithm can be used to handle the missing data problem [9]. In previous research on capture-recapture problems with missing covariates the EM algorithm has been used [12,25,28,29]. In these contributions the data are coded into a contingency table format for which the missing data problem is solved by methods developed by Little and Rubin [18] and Schafer [24]. See also Fienberg and Manrique-Vallier [11] for a discussion on the EM algorithm in capture-recapture methods.

However, only part of the information in the observed data seems relevant for solving the missing data problem, as we have great reservations of using the information from individuals that are observed in the PR. The reason is that it is unlikely that individuals that are registered in the PR are relevant for the CSR registered individuals that do not link to the PR or ER. Large part of the individuals in the CSR that are not registered in the PR and do not work in the Netherlands, are assumed to stay in the Netherlands for only a short period; on the other hand, in the PR almost all individuals reside longer than a year in the Netherlands. From the foreign individuals not in the PR but registered in the ER, only 30 percent are usual residents. If only 30 percent of the individuals registered in the ER, but not the PR, are usual residents, compared to nearly 100 percent of the PR registered individuals, it seems implausible to assume that the CSR registered individuals that are not in the PR and ER resemble PR registered individuals. Then the ER registered individuals provide better information about the missing observations in the CSR on the variable usual residence. We note, though, that the EM algorithm is not flexible enough to use only a subpopulation of the data for solving the missing data problem. However, given the missing data is unlikely to resemble the observed data, but only a part of it, the EM algorithm may give biased results.

Another method to handle missing data is multiple imputation. In the context of in capture-recapture analyses we know of one application of multiple imputation used before to impute missing values [29], as well as more general imputation methods [20]. For categorical data Kropko et al. [17] found that conditional multiple imputation, such as PMM, gave more accurate results in a simulation study on categorical missing data compared to multiple imputation from a joint distribution. In a comparison of imputation methods for binary data (not including EM), PMM outperformed the other imputation methods [21]. In this paper we use

PMM [8,26] as a method to be compared to the EM algorithm. PMM is a sequential multiple imputation method. When data are missing PMM enables the researcher to search the data for a unit that has the same characteristics as the unit that is to be imputed [8]. The advantage of PMM is that it provides the researcher with the possibility to use only part of the observed data set as donor to impute the missing usual residence. Thus, where EM has the drawback that it has to use the complete data to impute the missing information, PMM is able to solve the missing data problem in a more appropriate way when only part of the data is relevant for the missing data.

This paper contributes by comparing the EM algorithm and PMM to handle partially missing covariates in capture-recapture analyses. Four scenarios were identified. First, the EM algorithm with a so-called maximal loglinear model will be used to complete usual residence in the CSR. In this scenario, the maximal model is used for both EM algorithm and capture-recapture estimation. Scenario 2 and 3 are used to explore different models for the EM algorithm and the capture-recapture analysis. In scenario 2 the data are completed via the EM algorithm under the maximal loglinear model, comparable to scenario 1. However, capture-recapture is now used on this completed dataset to select the best fitting model. Scenario 3 was used to select the best fitting loglinear model for both the EM algorithm and capture-recapture. Whereas in scenario 2 the loglinear model for EM and capture-recapture will be different, here the restrictive model was kept constant for both completion via EM algorithm and estimation via capture-recapture. The fourth scenario will use the PMM imputation to impute the missing residence duration and will use only ER registered individuals that are not in the PR as donors. Then capture-recapture analysis was carried out on the completed data set.

We continue as follows. In Section 2 the data sources used in this manuscript and the linkage process will be explained. In Section 3 we will present the results from previous research on the size of usual residents. In Section 4 we describe the methods to complete the data and the estimation of the population size in more detail. In Section 5 we will present the results, and discuss which scenario gives the best estimate of the usual residents missed by the population register. In Section 6 we will conclude this manuscript.

2. Data sources and their linkage

Our capture-recapture analysis makes use of three linked registers. The PR is the official Dutch Population Register, in which individuals actively have to register themselves. The ER is a register not documenting individuals but documenting jobs. For the purpose of our analyses the job-register of 2010 has been transformed into a register on individuals. Jobs were attributed to the individuals holding those jobs. Moreover, if a job started in 2010 or was ended in 2010, the jobs are registered with a starting and/or an ending date. The CSR is a register in which suspects of crimes of which the police makes a report are recorded. This register is event based: the units are the reports of the police in which one or more crimes are recorded. There is little information in the CSR and the ER on individuals with ages under 15 and over 65, individuals under 12 can not be registered in the CSR and the ER only registers between 15 and 65. Thus the population specified in this paper consists of the population aged 15 to 65.

The data have been linked for the most part deterministically using a personal identification code that is widely used in Dutch registers. Probabilistic linkage was used to improve this linkage. During linkage it was found that 38% of the units that were registered only in the CSR had missing information in the linkage variables and therefore were difficult to link to the PR and ER. There is a chance that this group consists mostly of individuals that were either tourists or criminals entering the Netherlands for a short period. This would mean that these individuals are erroneous captures because they do not belong to the population of Dutch residents. It is important to assess whether these individuals substantially affect the population size estimate. However, this research topic is not in the scope of this article and is discussed elsewhere [3]. In this manuscript we have eliminated a sample of 30 percent of the individuals in the CSR population that did not link to the PR or the ER, assuming that these individuals are erroneous captures. Of these 30 percent 80 percent were individuals that had missing values in the linkage variables and 20 percent were individuals that did not have missing values in the linkage variables. This distribution was chosen assuming that individuals with missing values in the linkage variables had a higher chance of not belonging to the population.

Neither of the three registers has a covariate directly measuring residence duration. However, for two of the three registers we can derive residence duration from

Table 1
Observed values for the three registers

PR	ER	CSR		Total
		Yes	No	
Yes	Yes	2,115	259,804	261,919
	No	4,862	350,551	355,413
No	Yes	355	112,529	112,884
	No	3,561	0	3,561
	Total	12,419	722,884	733,777

information available in those registers. In the PR data are available on the date of registration. In the ER there are data available on joblength. For more details on how the ER residence duration was derived, see [3]. For those individuals in the CSR that link to either the PR, the ER or to both we use the residence duration from the PR or the ER. When residence duration is available from both the PR and the ER the longer residence duration is assumed superior over the residence duration of one of them. In the CSR only there are no variables available to derive residence duration from.

The three linked registers were analysed with loglinear models, as is the standard approach in capture-recapture of human populations [16, compare]. Four covariates were used in the loglinear models: nationality group, age, sex and usual residence. Initially nationality group has 8 categories: (1) EU15 (excl. Netherlands) (2) Polish (3) Other EU (4) Other western (5) Turkish, Moroccan, Surinam (6) Iraqi, Iranian, Afghan, asylum seeker countries Africa (7) Other Balkan, former Soviet Union, other Asian, Latin American, and (8) Other nationalities, not mentioned elsewhere. The countries are clustered according to likely migration motives, migration legislation, regulations of the PR and size. However, in the analysis the last nationality group gave numerically unstable results, and therefore the last two nationality groups were taken together, resulting in 7 nationality groups. For age, we use four levels: (1) 15–24 (2) 25–34 (3) 35–49 and (4) 50–64 years of age.

Table 1 shows the counts for the individuals in the three linked registers ignoring the distribution over the four covariates. The zero count is a structural zero as it represents the number of individuals that belong to the population but are not registered in any of the three registers. One of the aims of the analysis is to find an estimate for this cell and for this purpose a capture-recapture analysis will be executed. Table 1 shows that for the CSR there are more individuals not present than present. In particular the number of individuals in the ER and CSR, but not in the PR is small (355), much smaller than, for example, the number of individuals that are only in the PR (355,413). Given that these 355 individuals are distributed over four covariates, there are many small cell counts in our data, including observed zeros.

3. Previous findings

There is previous research on the estimation of population sizes (most notably [2,14,27]) that overlaps with the population of usual residents that we study in this paper, and these estimates are able to place the estimates found in our scenarios in perspective. However, this previous research shows a wide variety of estimated population sizes depending on the definitions of the population and the methods used, and therefore these studies cannot be used as a simple benchmark for judging the outcomes of our scenarios. Table 2 shows their estimates on individuals not registered in the PR.

Hoogteijling [14] collected different estimates from earlier research in the nineties. In order to achieve an estimate of the size of the population not registered in the PR and living four months or longer in the Netherlands in 2000, she combined the available information from different sources. Neglecting some very small categories, the population can be estimated by adding illegal immigrants, adding the balance of wrongfully not registered residents and wrongfully registered non-residents, and recently arrived asylum seekers who have not registered because they are not allowed to do so yet. This results in an estimate of 73 to 149 thousand missed residents, with a mean of 111 thousand, being less than 1% of the registered population (Table 2).

Bakker [2] also used information from different sources to get an estimate of the under and over coverage of the PR in 2006, having the same definition of usual residence as Hoogteijling [14], so those who stay longer than four month in the Netherlands are supposed to be usual residents. He distinguishes the different categories of which it is known that they are missed or are over counted in the PR and he estimates their numbers with different sources. He estimates the total under count as 205 thousand usual residents. However, there is a large uncertainty because some of the estimates are quite arbitrary. The largest contribution is from illegal immigrants whose size is estimated between 74 and 184 thousand. The total number of missed persons is 236 thousand, where 31 thousand persons are still in the population register while they have left the country or have died.

Van der Heijden et al. [27] used capture-recapture methodology to estimate the missed portion of the pop-

Table 2
Overview of previous research to individuals residing in the Netherlands

		min.	max.	Total/mean	of which usual resident	
		x1000			%	x1000
Hoogteijling (2002), estimates for the year 2000						
	Registered population			15987		
plus	illegal immigrants	46	116	81	80	65
balance	wrongfully not registered residents and wrongfully registered non-residents	−15	−16	−15.5	80	−12
plus	asylum seekers not yet registered as residents	42	49	45.5	80	36
	missed population of residents ($\geqslant$ 4 month)	73	149	111		89
	Total population of residents ($\geqslant$ 4 month)	16,060	16,136	16,098		16,076
Bakker (2009), estimates for the year 2006						
	Registered population		16,334			
plus	illegal immigrants	74	184	129	80	103
plus	foreign labour force			64	30	19
plus	foreign students			25	30	8
plus	diplomats and NATO military			6	80	5
plus	asylum seekers not yet registered as residents			6	80	5
balance	administrative delay			5	80	4
minus	non-residents working abroad temporarily			-29	30	-9
	missed population of residents ($\geqslant$ 4 month)			205		135
	Total population of residents ($\geqslant$ 4 month)			16,509		16,469

ulation in 2009 from Poland, Bulgaria, Romania and other nationality groups in middle and eastern Europe new in the EU (i.e. Hungary, Czech republic, Slovakia, Slovenia, Latvia, Lithuania and Estonia). Therefore, their outcomes can only be compared to two nationality groups used in this paper. They used the PR and the CSR as sources and applied capture-recapture methods to estimate usual residents in the same definition as we do. A difference between Van der Heijden et al. [27] and this manuscript is that in this manuscript assumed erroneous captures have been excluded from the analysis, whereas in Van der Heijden et al. [27] these may still have impact on the estimate. Additionally, the number of usual residents is given for the total population of individuals with a middle and eastern European nationality from new EU countries residing in the Netherlands, including those registered in the PR. There are 200 thousand usual residents with a middle and eastern European nationality not registered in the PR.

It is difficult to describe the expected value of the size of the population of usual residents in 2010, because some estimations are outdated and some do not use the same definition, or both. However, by harmonizing the results for the definitional differences and looking at the developments of the number of new asylum seekers and the number of foreign workers, we can provide a range of expected outcomes. These expected outcomes could help in providing a perspective where the current estimate of usual residents may be compared to.

In the under count of 111 thousand found by Hoogteijling [14] the majority is former or present asylum seeker. Because the procedures for seeking asylum had a long duration, certainly with a mean longer than a year, we assume that most residents who were not registered as such stayed for longer than a year in the Netherlands. Therefore we assume 80 percent of the 111 thousand not registered to be usual residents, which comes down to 89 thousand usual residents in 2000.

Bakker [2] estimated an over count of 205 thousand and this estimate is difficult to harmonize with the definition of usual residence in this manuscript, because we do not have empirical information on the residence duration of the different categories that are over counted in the PR. However, if we assume (i) that 80% of the illegal immigrants are a usual resident because they still are in majority former asylum seekers and (ii) that the same percentage is true for smaller categories like asylum seekers, diplomats and NATO military and administrative delay of new born and immigration, and (iii) that 30% of the foreign work force and foreign students is a usual resident, the same percentage as we found for the foreign work force in 2010 [3], then the estimated number of usual residents not registered in the PR is 135 thousand in 2006.

The estimate of [27] for the usual residents in 2009 from Eastern European countries uses the same definition and does not have to be harmonized. However, as they did not adjust their estimation for erroneous captures, the estimate of 200 thousand is expected to be too high. [3] show a decrease of approximately 37% if they correct for erroneous captures, we expect that the estimation of the size of the usual residents would be 126 thousand only from Eastern Europe.

Two significant developments have to be mentioned to explain changes in the number of not registered usual residents between 2000 and 2010. The first is the decline of the number of asylum seekers between 2000 and 2010 (Fig. 1). The numbers dropped from almost 45 thousand in 2000 to 10 thousand in 2004, among else due to changed regulations. After 2004 there is a more or less constant number of asylum requests between 7 and 15 thousand. The other one is the sharp rise of the foreign workforce from the year 2006, in particular from Eastern Europe, who did not register themselves in the population register. This was in 2006 121 thousand and increased to 182 thousand in 2010 (CBS, StatLine, 2015). This development was possible because the civilians of these countries could enter the Netherlands without a residence permit and after 2007 for the most part could also work without a working permit.

We arrive at the following conclusion, cautiously indicating that it is always dangerous to extrapolate earlier estimates to later periods. We expect that the number of usual residents not registered in the PR has been increased since the year 2000 to 175 to 225 thousand. The total number is certainly much higher than the 135 thousand in 2006 because of the inflow of migrant workers from Eastern Europe since then. On the other hand, the number of asylum seekers has been constant since 2006 and will not cause important developments. If the estimation of the number of not registered usual residents from Eastern Europe in 2009 is correct, then it is reasonable to assume that the upper bound is approximately 225 thousand, because the 100 thousand not registered usual residents from other countries will not have disappeared.

As can be seen from Table 1 there are 116,445 registered individuals not in the PR but in the CSR and/or the ER. Of these 116 thousand individuals, we found that 33 thousand are usual residents, and thus are part of the known under coverage, these individuals have to be added to the estimate from the scenarios [3].

4. Methods

We will use different scenarios for handling the missing data and estimating the part of the population missed by all three registers. Here, we describe the scenarios and evaluate them in the context of our problem. We will first give a short introduction to capture-recapture analysis using loglinear modelling and then we will discuss the EM algorithm and multiple imputation in this context.

4.1. *Capture-recapture methodology using loglinear modelling*

For estimating the size of human populations loglinear modelling seems to be the most popular method. It was discussed in depth in the standard work by Bishop et al. [4] and since then it has been reviewed regularly, for example by Cormack [7], the International Working Group for Disease Monitoring and Forecasting [16] and Chao et al. [6].

The simplest loglinear model for estimating the size of a population is based on two linked registers, A and B. Let the levels of A be indexed by i ($i = 0$,1) where $i = 0$ stands for "not included in register A", and $i = 1$, stands for "included in register A". Similarly, let the levels of B be indexed by j ($j = 0, 1$). Expected values are denoted by m_{ij}. Observed values are denoted by n_{ij} with $n_{00} = 0$, because there are no observations for the cases that belong to the population but were not present in either of the registers.

After linkage there is an observed number of individuals both in A as well as in B, n_{11}, an observed number of individuals only in A but not in B, n_{10} and an observed number of individuals only in B and not in A, n_{01}. Individuals being neither in A nor B are missed and capture-recapture can estimate this missing number, where we denote this estimate by $\hat{m}_{00}$. Assuming statistical independence of being in A and being in B, the odds ratio between being in A and being in B is 1, i.e. $n_{11}\hat{m}_{00}/n_{10}n_{01} = 1$. It follows that $\hat{m}_{00} = n_{10}n_{01}/n_{11}$. Then the population size N is estimated as $\hat{N} = n + \hat{m}_{00}$, where n is the observed number of individuals, i.e. $n = n_{11} + n_{10} + n_{01}$. The link between these equations and loglinear modelling is that loglinear parameters are functions of odds ratios. In the loglinear model for two variables, assuming that the odds ratio is 1 comes to the same as assuming that the interaction parameter between A and B is absent. The independence model just described is denoted in loglinear model notation as [A][B], showing that being in register A is unrelated to being in register B.

By including a third register C the assumption of statistical independence is replaced by the assumption that there is no three factor interaction. This model is denoted by [AB][AC][BC]. In other words, there

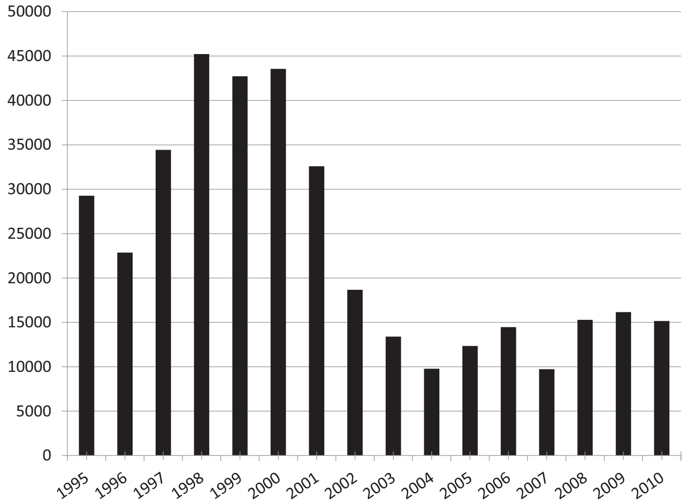

Fig. 1. Number of asylum requests in the Netherlands 1995–2010 [5].

may be interaction between A and B, but this interaction is identical in the sub-tables for individuals included in C and not in C. This also suggests a way to find an estimate $\hat{m}_{000}$: as the odds ratios are identical, $n_{111}n_{001}/n_{101}n_{011} = n_{110}\hat{m}_{000}/n_{100}n_{010}$, and this can be solved for $\hat{m}_{000}$.

Categorical covariates that are available in each of the registers, such as age and sex, can be easily added. If we collect them in a stacked variable X, the model becomes *[XAB][XAC][XBC]*. This model is saturated as the number of parameters is identical to the number of observed counts, and fitted values are equal to observed counts.

It is a crucial part of the capture-recapture procedure to search for a model that is more parsimonious than *[XAB][XAC][XBC]*, yet fits the data well. Parsimonious models have the advantage that the resulting estimate of N is more stable (i.e. has a smaller confidence interval) than saturated models [1]. The fit is usually evaluated using the deviance, that follows a chi-squared distribution with the number of counts minus the number of independent parameters when the model is true. Here the number of counts is (2*2*2 - 1)*(number of levels of the stacked covariates), where the ‘1’ refers to the cell that has a count of zero by design, hence this cell is called a structural zero. The number of possible models is large and there often is lack of theory that points out which models are of particular interest. Therefore exploratory model searches are usually employed, such as forward selection, backward elimination and stepwise procedures, that we also know from linear multiple regression. Using the difference deviance test, the AIC or the BIC a final model is chosen, where a model with either the lowest AIC or the lowest BIC is preferred. In contrast to the difference deviance test the AIC and the BIC can also be used to choose between non-nested models. Both the AIC as well as the BIC are a function of the deviance. The BIC leads to more parsimonious models, because the AIC has a penalty of $2k$ but the BIC a penalty of k log n, where k is the number of free parameters and n the observed count. We use the BIC, because in capture-recapture problems with a large observed count n, we prefer the BIC to prevent over fitting.

4.2. Missing covariate

The capture-recapture problem becomes more complicated when there are covariates involved that are not observed in every register. In the data that we study in this paper this holds for usual residence, which is not observed in the police register CSR. We consider this to be a missing data problem: the variable usual residence is missing for those individuals who are only in the register CSR. This type of missing data problem has been worked out in detail for two registers, see [12,28,29]. Here we have a capture-recapture problem with three registers instead of two, so this case deserves careful attention. Table 3 illustrates.

Table 3 shows the cross-classified counts for the three registers PR, ER and CSR, split out by UR and non-UR, for Polish individuals. Notice that there are 16 cells in total, where two are structurally zero (they refer to the individuals that are missed by all three registers), and for two cells we only know the sum, namely

Table 3
The Polish individuals by the three registers and usual residence. The two missing cells add up to 1,043

UR	PR	ER	CSR	
			Yes	No
No	Yes	Yes	32	3,523
		No	34	3,225
	No	Yes	149	60,190
		No	missing	0
Yes	Yes	Yes	183	21,309
		No	195	14,052
	No	Yes	81	20,216
		No	missing	0

for not in PR, not in ER, but in CSR. Only the sum is known because for these individuals usual residence is missing, so we cannot split them up over the cells yet. This holds for 1,043 individuals, as indicated in the header of the table.

To simplify the discussion of loglinear models, we now use the variables P for PR, E for ER and C for CSR, and U for usual residence. If usual residence would not be missing, the saturated model with as many parameters as counts would be *[PEU][PCU][ECU]*. The saturated model has 14 parameters, and it is saturated with the property that the fitted counts are identical to the observed counts. However, due to the missingness of the variable usual residence in the CSR there is one count less (the two missing counts add up to 1,043) and the maximal model for these data only has 13 parameters. The distinction between a so-called maximal model and the standard saturated model is that in the maximal model certain parameters are zero by design. The parameter for the interaction between P, C and U can be estimated from the counts 32, 3,523, 149, 60,190, 183, 21,309, 81 and 20216; the parameter for the interaction between P, C and U can be estimated from 32, 3,523, 34, 3,225, 183, 21,309, 195 and 14,052; but the parameter for the interaction between P, E and U is not identified because for the 1,043 individuals their level on U is not identified. Therefore the maximal model becomes *[PE][PCU][ECU]*. Interestingly, because of the identification problem the model has no parameter for the interaction between P, E and U. When individuals are not in P and E but only in C, the data for usual residence are missing by design. As a result the information on U in the maximal loglinear model is only available in the margins of the variables P, C and U, and the margin of E, C and U, but not in the margin of P, E and U. For more results on maximal models, see Zwane and Van der Heijden [29].

For evaluating scenarios it is important to note that model *[PE][PCU][ECU]* makes two assumptions. First the three-factor interaction between P, E and C is zero. In other words, the relation between P and E is identical for those who are in C and those who are not in C (and, similar statements can be made for the relation between P and C and between E and C). Secondly, the interaction between P, E and U is zero, meaning that the interaction between P and E is identical for those individuals who have a usual residence shorter than a year and those who have a usual residence longer than a year. However, this last assumption is not very plausible. It is known that those who stay longer in the country, in particular from Eastern Europe, assimilate fast in society, find permanent work and a partner [13]. Therefore they will register themselves more frequently than those who only live in the Netherlands for a short period. In other words, it is plausible that for those who reside in the Netherlands for longer than a year the odds ratio between P and E will be larger than for those who reside in the Netherlands for shorter than a year. Yet the maximal model cannot accommodate this because the interaction between P, E and U cannot be estimated.

These preliminaries bring us to the definition of the scenarios. Our statistical problem has two aspects:

(i) there are missing data on the variable usual residence, and

(ii) using capture-recapture methodology we are going to fit a loglinear model under which the part of the population that is missed by all three registers is estimated.

Our scenarios differ in the way that the two steps are taken. We make use of two procedures for handling the missing data problem: the Expectation-Maximization method (EM) and multiple imputation using Predictive Mean Matching (PMM).

4.3. Scenarios using the EM algorithm

The EM algorithm is a general iterative algorithm for maximum likelihood estimation when data are incomplete [18]. The EM algorithm consists of an Expectation (E) step and Maximization (M) step. In general in the E step the algorithm replaces missing values by values that are expected under a given model. Then under the M step the algorithm estimates parameters that are maximized on the expected values of the E step. Then in the next E step expectations are calculated for the missing values using the current best parameter estimates found in the last M step, after which a new M step maximizes the parameters using the data completed in the E-step. This is repeated until convergence occurs, where the joint distribution of the register and covariate variables are preserved, under the loglinear model specified. Hence MAR is assumed under the joint distribution given loglinear modeling. For the maximal model, convergence is after only one iteration.

After completion of the EM algorithm, the completed data can be used for capture recapture estimation. When the loglinear model for capture-recapture estimation is identical to the loglinear model the parameter estimates from the EM algorithm can be used to estimate the population size. Then the EM algorithm alone suffices to estimate the missed portion of the population. However, when after EM completion another loglinear model is preferred for capture-recapture analysis, the EM completed data can be used as input for the capture recapture analysis.

We distinguish the following scenarios. For scenario 1 we use the EM algorithm where the loglinear model chosen is the maximal model. For capture recapture analysis the maximal model is also used, and thus the parameter estimates from the EM algorithm can be used to estimate the missed portion of the population. In scenario 2 we also use the maximal model for the EM algorithm to complete the data. However, this completed data are then used as input for capture-recapture analysis where the function STEP in R is used to look for the best fitting, parsimonious loglinear model to the completed data. Then the EM algorithm and capture-recapture are done in 2 steps. In scenario 3, just as in scenario 1, we use the parameters from the EM algorithm to estimate the missed portion of the population. However, unlike scenario 1, scenario 3 will use more restrictive models for the EM algorithm.

4.4. *Scenario for multiple imputation using predictive mean matching*

A fourth scenario is multiple imputation using Predictive Mean Matching (PMM). When an individual has missing data PMM enables the researcher to search in the data for individuals that have the same characteristics as the individual that has values that need to be imputed, and use their observed values to impute the missing value [8]. MAR is assumed, in that units with the same background characteristics will have similar values on the missing variable, if this variable would have been observed.

Predictive mean matching is an example of a hot deck, nearest neighbour multiple imputation method. Missing values are imputed using values from the complete cases matched with respect to some metric. All individuals with the same background variables as the missing value are candidate donors for imputing. From these donors, one random donor is sampled from the candidates and the value on usual residence from this donor candidate is taken as a value for the missing unit [18,26]. By selecting individuals from the same background values, the joint distribution based on the background variables is preserved.

The PMM procedure has been repeated ten times, to account for the uncertainty of the individual imputations. To estimate the number of usual residents, the capture-recapture method has been applied to all imputed datasets. To estimate the number of usual residents, the mean of the ten estimates has been computed.

PMM has the advantage that it allows to select a specific subpopulation for which it can be assumed that it resembles the subpopulation that has to be imputed best. In this case, we have to find a donor population for the CSR-records that do not link to the PR and ER. In this donor population it is presumable that the residence duration is relatively short, because there is a positive association between residence duration and registration. Therefore, for the donors we choose individuals in the ER that do not link to the PR, i.e. individuals who are working but did not register as a resident of the Netherlands, thus we assume MAR in a conditional distribution. They also have a relatively short residence duration because of this aforementioned association.

4.5. *Concluding remarks*

We summarize the scenarios here. Both the EM algorithm and multiple imputation using PMM are established methods with a solidly grounded base in literature. Both methods assume Missing At Random (MAR). There are also two important differences.

First, the EM algorithm cannot make use of models that are more complicated than the maximal model. In the maximal model the interaction between P, E and U cannot be estimated. For the EM algorithm the missing data are completed from a joint distribution of the observed data under a given loglinear model. It has been argued above that this is a drawback for our missing data, which is assumed to resemble only a subpopulation of the observed data. On the other hand, PMM is applied using as a donor population the subpopulation of individuals that are in ER but not in PR. For this subpopulation the relation with usual residence is used.

Second, both methods handle missing data differently. EM algorithm completes the incomplete data according the loglinear model specified. Note that differences in loglinear models may result in different estimates [18,28]. Thus the choice of the loglinear model is important. Predictive Mean Matching (PMM) is a sequential multiple imputation method. When data are missing PMM enables the researcher to search the data for a unit that has the same characteristics as the unit that is to be imputed [8]. The advantage of PMM is that a missing unit will be given the same value on the missing variable of an observed unit. It is assumed that units with the same background characteristics will have similar values on the missing variable, if this variable would have been observed. Then PMM has the advantage of assuming MAR between the missing data, and the observed data that resembles the missing data best. Multiple imputation using PMM is flexible in the sense that it is possible to use only that part of the table that seems most appropriate to use for the problem at hand. So in an evaluation of both differences multiple imputation by PMM seems better suited to handle the problem that we study.

Throughout the paper, the software R has been used for all computations. For the EM algorithm the package CAT [19,22,23] was used. For multiple imputations using PMM the package MICE was used [26]. After completion, the R-function GLM was used to estimate the missing part of the population. We used parametric bootstrap confidence intervals to estimate a 95% confidence interval for the point estimate of usual residents. 10,000 bootstrap samples are used.

5. Results

5.1. Scenario 1: Maximal model for EM estimation and capture-recapture analysis

In our first scenario the missing data are completed under the maximal loglinear model, and the missing part of the population is also estimated under this model. This approach is carried out for the nationality groups separately.

For almost all of the nationality groups the population size estimates tend to infinity. To examine why we get these results, we have to take in mind that we employ a two-step process. First the incomplete data is completed via EM algorithm and then the capture-recapture analysis is carried out to get an estimate of the missed portion of the population, and problems can occur in both completion of the data or estimation under capture-recapture analysis.

Table 4
Estimated Polish usual residents that are missed by all three registers, by Age and Sex

Age	Men	Women
15–24	137	26
25–34	138	10
35–49	53	3
50–64	7	2

As an illustration of what goes wrong we go back to Table 3, which is a marginal table of individuals with a Polish nationality where we added up over the covariates sex and age. There are two structurally zero cells, representing the part of the population that is not observed, and two cells that are zero for which usual residence has to be estimated, but where the sum should be 1,043.

First the missing values for usual residence is estimated under the maximal model. This yields 655 individuals categorized as residing shorter than a year, and 376 as longer than a year. The resulting table is the input data for the capture-recapture analysis. Estimates are derived again under the maximal model. As in the maximal model fitted counts are equal to observed counts, it is important to realize that the observed counts in Table 3 are further split up over age and sex. Hook and Regal [15] discussed that some models, especially the saturated model (in our case, the maximal model), is sensitive to small or zero cells. As an example, Table 4 shows the 376 out of 1,043 individuals in the CSR only, classified as usual residents. There are only 2 women with an age between 50 and 64, whereas there are 137 young men. Because our data consists of both large and small cell numbers, and in this scenario we used the maximal model for capture-recapture analysis, it follows that the resulting estimate is implausibly high. Capture-recapture analysis thus is sensitive under the maximal model in contingency tables with small and zero cells, which we have. Thus the maximal model cannot be reliably used for capture-recapture analysis under the current data.

5.2. Scenario 2: Maximal model for EM estimation, restrictive models for capture-recapture analysis

In this scenario, after the EM algorithm was used for completing the missing data under the maximal model, we used the function STEP in R to choose the best fitting loglinear model via the BIC for the capture-recapture analysis. Table 5 shows the results. We find

Table 5
Estimates for scenario 2

Nationality	Total missed	< 1 year	⩾ 1 year	Confidence interval
	x1000	x1000	x1000	x1000
EU15	146	104	42	30–47
Polish	265	211	53	49–69
Other EU	155	132	23	13–30
Other West	16	12	4	3–5
Turkey, etc.	3	1	1	0.8–2
Iraq, etc.	9	8	2	1–2
Balkan, etc. Other.	65	51	14	10–28
Total	659	520	139	120–176

Estimates of the missed portion of the population per nationality for scenario 2, where the maximal model is used for EM algorithm. Capture-recapture analysis was done with the best fitting restrictive loglinear model.

Table 6
Estimates for scenario 3

Nationality	Total missed	< 1 year	⩾ 1 year	Confidence interval
	x1000	x1000	x1000	x1000
EU15	147	104	42	31–59
Polish	243	195	48	38–62
Other EU	142	123	20	13–35
Other West	18	13	4	3–7
Turkey	3	2	2	1–2
Iraq	10	8	2	2–3
Balkan, etc., Other.	45	34	11	9–13
Total	608	480	129	111–170

Estimates of the missed portion of the population per nationality for scenario 3, where the same restrictive loglinear model is used for EM algorithm and capture-recapture analysis.

a total of 659 thousand individuals missed by all three registers, 139 thousand of those individuals are usual residents (confidendence interval 120–176). The usual residents are 21 percent of the total missed portion of the population. Scenario 2 resulted in parsimonious and more stable models where the outcome seemed more plausible.

The loglinear models for this scenario can be found in the Appendix. The models for scenario 2 are more parsimonious and restrictive than the maximal model from scenario 1. For example, take the model of the individuals with a Polish nationality. The last term in this model [PUSA] is comparable to the first term in the model for scenario 1, which was [PCUSA], but in scenario 2 the data are collapsed over the CSR. In deleting CSR from the interaction term the distribution of individuals over the contingency table becomes more balanced, and this leads to estimates from the capture-recapture analysis that are numerically more stable.

5.3. *Scenario 3: Restrictive models for both EM estimation and capture-recapture analysis*

In scenario 3 we use more restrictive models for the EM algorithm, and keep models for EM and capture-recapture analysis equal. This is the standard approach of using the EM algorithm. Results can be found in Table 6. There are 129 thousand usual residents missed (CI is 111–170), which again is 20 percent on 608 thousand total individuals missed by the three registers. This total estimate is similar to the estimate for scenario 2, however, the estimates per country are somewhat different. Interestingly, as can be seen in the Appendix, the best fitting models for capture-recapture analysis on the completed datasets are quite different between these two scenarios.

Table 7 shows the data after completion via EM algorithm under a more restrictive model. Note that the data differ compared to Table 3. Under more restrictive models the content of the table can change to maximize the fit of the margins under the loglinear model. Thus the completed table differs from the observed Table 3.

5.4. *Scenario 4: Multiple imputation using PMM*

Table 8 shows that when we use multiple imputation using PMM and conduct a capture-recapture analysis on the data, 610 thousand individuals are missed by all three registers, of which 179 thousand are usual residents(with a lerger CI of 121–237). Now there are 29

Table 7
The data for the Polish individuals after completion with EM algorithm via restrictive loglinear models

UR	PR	ER	CSR	
			Yes	No
No	Yes	Yes	23	3,530
		No	33	3,226
	No	Yes	158	60,180
		No	781	0
Yes	Yes	Yes	193	21,300
		No	196	14,050
	No	Yes	71	20,225
		No	261	0

percent usual residents on the total number of individuals missed by all three registers, a slight increase compared to scenario 2 and 3. The CI for usual residents is higher than the estimates resulting from the EM algorithm when more restrictive models are used, such as in scenario 2 and 3.

After PMM imputation the data are similar to Table 3 in that the observed part of the data remains unchanged. However, the 1,043 individuals in the CSR are distributed differently per imputation, where generally less usual residents are imputed than for scenarios 2 and 3. As can be seen from the Appendix the loglinear model best fitting this completed data is different from the other three scenarios and a different estimate results.

5.5. *Synthesis of the results*

Four scenarios were defined to assess the effect of different methods of completing missing data on the population size estimate from capture-recapture. Usual residence has been completed in three different ways. For scenario 1 and 2 the maximal loglinear model was used via EM algorithm. The third scenario also employed the EM algorithm but used more restrictive loglinear models. Scenario 4 used multiple imputation by means of PMM. For all four scenarios different loglinear models were used. For the first scenario a highly unrealistic estimate is achieved. Scenario 2 and 3 had more plausible results than scenario 1, and these estimate were similar to each other. Scenario 4 had a higher but still plausible estimate.

For the scenario's with EM algorithm, we found that the estimates for scenario 2 and 3 were more plausible than the estimates for scenario 1. One possible explanation is that in scenario 2 and 3 compared to scenario 1, the loglinear models have less interaction terms with the CSR (Compare Tables 9 and 10). However, the CSR is not completely removed from the best fitting loglinear models and therefore we can not completely remove the CSR from the analyses. We shortly discuss a capture-recapture estimation without using the CSR. When a capture-recapture analysis is conducted on PR and ER alone only 27 thousand individuals are estimated as the missed portion of the population. This estimate is implausibly low, given that half of that number alone is an asylum seeker (see Fig. 1). In deleting the CSR from all interaction terms we assume being in PR and being in ER are statistically independent, which is not realistic. Moreover, capture-recapture analysis for only PR and ER will result in a very specific population, that does not include illegal immigrants. Hence CSR is important for our capture-recapture analysis, but may be deleted from some interaction terms.

Not surprisingly, different data sets lead to different loglinear models that fit the data best. Hence, changes in the loglinear models over the four scenarios is one part of the explanation for the different estimates for the missed part of the population. We note that the impact of minor changes in the estimates may lead to relatively large changes in the population size estimates in the presence of small or zero observed counts.

The confidence interval estimates show a great overlap between the scenarios, especially for scenario 2 and 3. Scenario 4 is more distinct and larger, and includes the point estimates of both scenario 2 and 3 in its confidence interval. Thus based on the confidence interval, the estimate of usual residents between the scenarios do not differ much. Thus the population size estimates can not be distinguished from one another based on the confidence intervals.

We indicated that PMM probably is better in dealing with the missing values on usual residence in the individuals that only appear in the CSR, which is a very specific subpopulation that will not resemble the whole dataset. Then PMM imputation is preferred since only a subgroup, best resembling the missing data, is chosen as a donor to impute the missing usual residence variable. We also bring the estimates in perspective of previous research on this subject. Following the trend laid out by previous research we expect the estimate to be between 175 thousand and 225 thousand. Given that approximately 33 thousand individuals in the ER and CSR that did not link to the PR are usual residents, we have to add this number to the estimate of usual residents missed by all three registers for the scenarios 2, 3 and 4. Then we get 172 thousand for scenario 2, 152 thousand for scenario 3 and 212 thousand for scenario 4. This means that only the estimate of scenario

Table 8
Estimates for scenario 4

Nationality	Total missed	$<$ 1 year	$\geqslant$ 1 year	Confidence interval
	x1000	x1000	x1000	x1000
EU15	156	112	46	24–68
Polish	240	178	61	28–95
Other EU	138	95	42	16–68
Other West	15	11	5	1–9
Turkey, etc.,	4	2	2	0.5–4
Iraq, etc.,	9	3	6	5–7
Balkan, etc., Other.	48	31	17	12–22
Total	610	431	179	121–237

Estimates of the missed portion of the population per nationality for scenario 4, after PMM imputation and restrictive loglinear modelling capture-recapture analysis.

4, for the PMM imputed data, lies within the prespecified range.

However, the prespecified range is based on earlier research that overlap with the research described in this manuscript, but the earlier research differs from the research discussed here in terms of data sources and models used. Given that there is no good benchmark we have to rely on other measures. Only scenario 4 lies in the prespecified range based on earlier research. However, this range may be incorrect given it is based on assumptions that may not hold. So we can not take this range alone as a benchmark, especially given the confidence intervals overlap between the scenario's and thus indicates no distinct estimates. Fomr a methodology perspective, with PMM it is possible to use only a subgroup of the population that resembles the missing units best, for which the EM algorithm is not as flexible. Additionally, in using multiple imputations, uncertanties between imputations are taken into consideration. Then the estimate from scenario 4 are based on better methodology and we assume this scenario to outperform the others. This taken together with the cautious indication for scenario 4 by the prespecified range, the estimate of usual residents not registered in the PR is 218 thousand. However, this conclusion has to be taken lightly given that the indications towards scenario 4 are small.

6. Conclusion

Often covariates are used in capture-recapture estimation for estimating hard-to-reach populations. When covariates are important for answering a research question, missing data in these covariates can pose a problem. Since such a covariate cannot be left out of the analysis, a solution has to be sought to handle the missing data problem. In this paper the covariate that poses a problem is usual residence. Since we are interested in estimating usual residents we cannot exclude the covariate.

There are multiple ways of handling the missing data for categorical data sets. In this paper we have chosen to compare the EM algorithm and multiple imputation using PMM. For EM algorithm three variants were chosen, (i) a maximal model for EM completing and capture recapture estimation, (ii), a maximal model for EM algorithm and a more restrictive model for capture-recapture analysis, and (iii) a restrictive loglinear model that is identical for EM completion and capture-recapture analysis. After multiple imputation using PMM the best fitting loglinear model is chosen for capture-recapture analysis scenario 1 gave unrealistic estimates, scenario 2 and 3 gave low estimates which are similar to one another, and scenario 4 gave a higher estimate than scenario 2 and 3. The estimate after PMM imputation is slightly preferable over the estimates after EM completion. However, scenario 2 and 3 come close to the pre-specified range, and even though the range is based on expert knowledge from previous research, it may still be incorrect. Thus, the estimates of scenario 2 and 3 are not necessarily incorrect.

Methodologically PMM is assumed to better handle missing data in our manuscript because PMM is more flexible in dealing with missing data when a specific subpopulation is missing. Additionally, the estimate after PMM imputation of 217 thousand usual residents not registered in the PR was the only one that lies in the prespecified. This range alone can not decide which scenario is best, however combined with the methodological advantage PMM imputation has over the EM algorithm, we cautiously give our preference for the estimate of scenario 4.

At the beginning of this manuscript we stated that the registers do not register intent to stay, but only a length of stay. However, we assume the individuals in the PR do have an intent to stay for a longer period. Individuals register themselves in the PR, which by law has to be done when an individual is in the Netherlands for four months or longer, or intents to stay for four months or longer. Given the assumption that a person registers in the PR with the intent to stay for a longer period, we assume all PR registered individuals to intent to stay for longer than a year. Thus all PR registered individuals will be considered usual residents. Given that there are 16,638 thousand individuals registered in the PR, and assuming these individuals have registered to reside for a longer period in the Netherlands and thus are usual residents, the total number of usual residents in the Netherlands is 16,856 thousand, of which 1.3% are not registered in the PR.

We have seen that capture-recapture analysis is sensitive to changes in loglinear models, and also that changes in data with small and zero cell counts lead to different best fitting loglinear models. These findings are in particular relevant when there are many covariates, some data sources have few units or combinations of both. Therefore we advise researchers to be cautious in interpreting the results from capture-recapture analysis in these situations.

Acknowledgements

The authors would like to thank Peter-Paul de Wolf for his continuing help. Additionally, the authors thank Jeroen Pannekoek, Sander Scholtus, Stef van Buuren and Gerko Vink for their advise.

References

[1] A. Agresti, Categorical data analysis. Wiley-Interscience, 2013.

[2] B.F.M. Bakker, Trek alle registers open! (Open all registers!). Vrije Universiteit Amsterdam, 2009.

[3] B.F.M. Bakker, S.C. Gerritse, P.G.M. van der Heijden, D.J. van der Laan, H.N. van der Vliet and M.J.L.F. Cruyff, Estimation of non-registered usual residents in the netherlands, ultimo september 2010. Conference of European Statistics Stakeholders, Rome, Italy, 2014.

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

[5] Central Bureau of Statistics. StatLine, 15 April 2015.

[6] 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** (2001), 3123–3157.

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

[8] T. De Waal, J. Pannekoek and S. Scholtus, Handbook of statistical data editing and imputation. Wiley, 2011.

[9] A.P. Dempster, N.M. Laird and D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society* **39** (1977), 1–38.

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

[11] S.E. Fienberg and D. Manrique-Vallier, Integrated methodology for multiple system estimation and record linkage using a missing data formulation, *Advanced Statistical Analysis* **93** (2008), 49–60.

[12] 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 Offcial Statistics* (2015).

[13] M. Gijsberts and M. Lubbers, Langer in Nederland, ontwikkelingen in de leefsituatie van migranten uit Polen en Bulgarije in de eerste jaren na migratie (Longer in the Netherlands, Developments on living conditions of migrants from Poland and Bulgaria in the first years after migration.). Sociaal Cultureel Planbureau, The Hague, 2015.

[14] E.J.M. Hoogteijling, Raming van het aantal niet in de gba geregistreerden (Estimates of the number of nongba registered.). Centraal Bureau voor de Statistiek, 2002.

[15] E.B. Hook and R.R. Regal, Accuracy of alternative approaches to capture-recapture estimates of disease frequency: Internal validity analysis of data from five sources, *American Journal of Epidemiology* **152** (2000), 771–779.

[16] 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** (1995), 1047–1058.

[17] J. Kropko, B. Goodrich, A. Gelman and J. Hill, Multiple imputation for continuous and categorical data: Comparing joint and conditional approaches, *Political Analysis* **22** (2014), 497–519.

[18] R.J. Little and D.B. Rubin, Statistical analysis with missing data, John Wiley and Sons, 2002.

[19] X.L. Meng and D.B. Rubin, Ipf for contingency tables with missing data via the ecm algorithm, in proceedings of the statistical computing section of the american statistical association, *American Statistical Association* Washington, DC., 1991, page 244–247.

[20] ONS. 2011 census item edit and imputation process: 2011 census: Methods and quality report. Office for National Statistics, 2012.

[21] M. Peeters, M. Zondervan-Zwijnenburg, G. Vink and A.G.J. Van der Schoot, How to handle missing data: A comparison of different approaches, *European Journal of Developmental Psychology* (2015).

[22] J.L. Schafer, Analysis of incomplete multivariate data. monographs on statistics and applied probability. Chapman and Hall, London, 1997.

[23] J.L. Schafer, Imputation of missing covariates under a general linear mixed model. Department of Statistics, Pennsylvania State University, 1997.

[24] J.L. Schafer, Analysis of incomplete multivariate data. Chapman and Hall, 1997.

[25] J.M. Sutherland, C.J. Schwarz and L.-P. Rivest, Multilist population estimation with incomplete and partial stratification, *Biometrics* **63** (2007), 910–916.

[26] S. van Buuren, Flexible imputation of missing data. CRC Press, 2012.

[27] P.G.M. Van der Heijden, M.J.L.F. Cruyff and G. Van Gils, Aantallen geregistreerde en niet-geregistreerde burgers uit MOE-landen die in Nederland verblijven, Rapportage schattingen 2008 en 2009. (The number of registered and non-registered citizens from MOE-countries residing in the Netherlands. Reporting estimations 2008 and 2009.). The Hague, Ministry of Social Affairs and Employment, 2011.

[28] 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** (2012), 831–852.

[29] 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** (2007), 1069–1089.

Appendix

Maximal model for scenario 1. *[PCUSA][ECUSA] [EPSA].*

Table 9
Loglinear models per nationality for scenario 2

Nationality	Model
EU15	[CS] [CE] [SE] [PE] [PCU] [PUS] [PCA] [PUA] [CUA] [PSA] [USA] [USE] [UAE] [SAE]
Polish	[PC] [CE] [PE] [PUS] [CUS] [PUA] [CUA] [PSA] [USA] [USE] [SAE] [PUSA]
OthE EU	[CE] [PE] [PCU] [PCS] [PUS] [CUS] [PCA] [PUA] [CUA] [PSA] [USA] [USE] [UAE] [SAE] [PCUS] [PUSA] [USAE]
OthE West	[PU] [PS] [CS] [PA] [PE] [USA] [USE] [UAE] [SAE]
TUkey etc.	[PE] [PCU] [PCS] [PUA] [CUA] [CSE] [CAE] [USAE]
Iraq, etc.	[PU] [PS] [CS] [US] [PA] [CA] [UA] [CE] [UE] [PE] [SAE]
Balkan, etc. Other.	[[CE][PE][PCU][PCS][PUS][CUS][PCA][PUA][USA][USE][UAE][SAE][PCUS]

Table 10
Loglinear models per nationality for scenario 3

Nationality	Model
EU15	[PC] [PE] [PU] [PS] [PA] [CE] [CS] [EU] [ES] [EA] [US] [UA] [SA]
Polish	[PC] [PE] [PU] [PS] [PA] [CE] [CU] [CS] [CA] [EU] [ES] [EA] [US] [UA] [SA]
Other EU	[PC] [PE] [PU] [PS] [PA] [CE] [CS] [CA] [EU] [US] [UA] [SA]
Other West	[PU] [PS] [PA] [CE] [CU] [CS] [EU] [ES] [EA] [US] [UA] [SA]
Turkey, etc.	[PC] [PE] [PU] [PS] [PA] [CE] [CU] [CS] [CA] [EU] [ES] [EA] [US] [UA] [SA]
Iraq, etc,	[PE] [PU] [PS] [PA] [CE] [CU] [CS] [CA] [EU] [ES] [EA] [US] [UA] [SA]
Balkan, etc. Other.	[PE] [PU] [PS] [PA] [CE] [CU] [CS] [CA] [EU] [ES] [EA] [US] [UA]

Table 11
Loglinear models per nationality for scenario 4

Nationality	Models
EU15	[PE] [CS] [CA] [PC] [EAU] [SAU] [ESA] [PAU] [ESU] [PSU] [PSA] [ECU]
Polish	[EC] [CS] [EA] [EU] [CU] [PC] [PE] [CA] [PSA] [PSU] [SAU] [PAU] [PSAU]
Other EU	[EC] [PC] [PSA] [PSU] [PEU] [PEA] [SAU] [PCS]
Other West	[PS] [CS] [PA] [ESA] [EAU] [PEU]
Turkey, etc.	[PC] [ESA] [SAU] [PEU] [ECA] [EAU] [ECS] [PES] [CAU]
Iraq, etc.	[CS] [PE] [CA] [SU] [PS] [PA] [EC] [PEU] [ESA] [EAU]
Balkan, etc. Other	[CS][PS][EC][PA][CA][CU][PEU][EAU][ESA][ESU][SAU][PAU][CAU]